Procesamiento del Lenguaje Natural, Revista nº **51**, septiembre de 2013

sociedad española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT | FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA | 2012

ISSN: 1135-5948

Artículos

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013

ISSN: 1135-5948

Comité Editorial

Consejo de redacción

ISSN: 1135-5948
ISSN electrónico: 1989-7553
Depósito Legal: B:3941-91
Editado en: Universidad Complutense de Madrid
Año de edición: 2013
Editores: Alberto Díaz Esteban Universidad Complutense de Madrid albertodiaz@fdi.ucm.es
Publicado por: Sociedad Española para el Procesamiento del Lenguaje Natural
Departamento de Informática. Universidad de Jaén
Campus Las Lagunillas, EdificioA3. Despacho 127. 23071 Jaén
secretaria.sepln@ujaen.es

Consejo asesor

Felisa Verdejo	UNED
Manuel Vilares	Universidad de A Coruña
Leonel Ruiz Miyares	Centro de Lingüística Aplicada de Santiago de Cuba
Luis Villaseñor-Pineda	Instituto Nacional de Astrofísica, Óptica y Electrónica, México
Manuel Montes y Gómez	Instituto Nacional de Astrofísica, Óptica y Electrónica, México
Alexander Gelbukh	Instituto Politécnico Nacional, México

Revisores adicionales

Emmanuel Anguiano	Instituto Nacional de Astrofísica, Óptica y Electrónica, México
Miguel Ballesteros	Universidad Pompeu Fabra
Susana Bautista	Universidad Complutense de Madrid
Zoraida Callejas	Universidad de Granada
Alberto Díaz	Universidad Complutense de Madrid
Arantza Casillas	Universidad del País Vasco
Manuel C. Díaz-Galiano	Universidad de Jaén
Miguel Ángel García Cumbreras	Universidad de Jaén
Manuel García-Vega	Universidad de Jaén
Milagros Fernández Gavilanes	Universidad de A Coruña
Daniel González	Universidad de Vigo
Víctor Flores	Universidad Europea de Madrid
Virginia Francisco	Universidad Complutense de Madrid
Pablo Gervás	Universidad Complutense de Madrid
Ignacio Giráldez	Universidad Europea de Madrid
Iakes Goenaga	Universidad del País Vasco
José M. Gómez	Optenet
David Griol	Universidad Carlos III de Madrid
Raquel Hervás	Universidad Complutense de Madrid
Adrian Pastor Lopéz-Monroy	Instituto Nacional de Astrofísica, Óptica y Electrónica, México
Montserrat Marimon	Universidad de Barcelona
Fernando Martínez Santiago	Universidad de Jaén
Gonzalo Rubén Méndez Pozo	Universidad Complutense de Madrid
Simon Mille	Universidad Pompeu Fabra
Rafael Muñoz	Universidad Europea de Madrid
Eva Navas	Universidad del País Vasco
Jesús Peral	Universidad de Alicante
Guillermo Pérez	Universidad de Sevilla
Francisco José Ribadas-Pena	Universidad de Vigo
Damiano Spina	UNED
David Vilares	Universidad de A Coruña
Julio Villena	Daedalus
Arkaitz Zubiaga	Dublin Institute of Technology

Procesamiento del Lenguaje Natural, Revista nº **51**, septiembre de 2013

ISSN: 1135-5948

Preámbulo

La revista "Procesamiento del Lenguaje Natural" pretende ser un foro de publicación de artículos científico-técnicos inéditos de calidad relevante en el ámbito del Procesamiento de Lenguaje Natural (PLN) tanto para la comunidad científica nacional e internacional, como para las empresas del sector. Además, se quiere potenciar el desarrollo de las diferentes áreas relacionadas con el PLN, mejorar la divulgación de las investigaciones que se llevan a cabo, identificar las futuras directrices de la investigación básica y mostrar las posibilidades reales de aplicación en este campo. Anualmente la SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural) publica dos números de la revista, que incluyen artículos originales, presentaciones de proyectos en marcha, reseñas bibliográficas y resúmenes de tesis doctorales. Esta revista se distribuye gratuitamente a todos los socios, y con el fin de conseguir una mayor expansión y facilitar el acceso a la publicación, su contenido es libremente accesible por Internet.

Las áreas temáticas tratadas son las siguientes:

- Modelos lingüísticos, matemáticos y psicolingüísticos del lenguaje.
- Lingüística de corpus.
- Desarrollo de recursos y herramientas lingüísticas.
- Gramáticas y formalismos para el análisis morfológico y sintáctico.
- Semántica, pragmática y discurso.
- Lexicografía y terminología computacional
- Resolución de la ambigüedad léxica.
- Aprendizaje automático en PLN.
- Generación textual monolingüe y multilingüe.
- Traducción automática.
- Reconocimiento y síntesis del habla.
- Extracción y recuperación de información monolingüe, multilingüe y multimodal.
- Sistemas de búsqueda de respuestas.
- Análisis automático del contenido textual.
- Resumen automático.
- PLN para la generación de recursos educativos.
- PLN para lenguas con recursos limitados.
- Aplicaciones industriales del PLN.
- Sistemas de diálogo.
- Análisis de sentimientos y opiniones.
- Minería de texto.
- Evaluación de sistemas de PLN.
- Implicación textual y paráfrasis.

El ejemplar número 51 de la revista de la Sociedad Española para el Procesamiento del Lenguaje Natural contiene trabajos correspondientes a tres apartados diferenciados: comunicaciones científicas, resúmenes de proyectos de investigación y descripciones de herramientas. Todos ellos han sido aceptados mediante el proceso de revisión tradicional en la

revista que ha sido llevado a cabo según el calendario previsto. Queremos agradecer a los miembros del Comité asesor y a los revisores adicionales la labor que han realizado.

Se recibieron 53 trabajos para este número de los cuales 45 eran artículos científicos y 8 correspondían a resúmenes de proyectos de investigación y descripciones de herramientas. De entre los 45 artículos recibidos 22 han sido finalmente seleccionados para su publicación, lo cual fija una tasa de aceptación del 48.9%. Autores de otros 10 países han participado en los trabajos publicados en la revista. Estos países son: Alemania, Inglaterra, Escocia, Irlanda, Irán, Cuba, México, EEUU, Brasil y Ecuador.

El Comité asesor de la revista se ha hecho cargo de la revisión de los trabajos. Este proceso de revisión es de doble anonimato, se mantiene oculta la identidad de los autores que son evaluados y de los revisores que realizan las evaluaciones. En un primer paso cada artículo ha sido examinado de manera ciega o anónima por tres revisores. En un segundo paso, para aquellos artículos que tenían una divergencia mínima de tres puntos (sobre siete) en sus puntuaciones sus tres revisores han reconsiderado su evaluación en conjunto. Finalmente, la evaluación de aquellos artículos que estaban en posición muy cercana a la frontera de aceptación ha sido supervisada por más miembros del Comité.

Estimamos que la calidad de los artículos es alta. El criterio de corte adoptado ha sido la media de las tres calificaciones, siempre y cuando haya sido igual o superior a 5 sobre 7.

Septiembre de 2013
Los editores

Procesamiento del Lenguaje Natural, Revista nº **51**, septiembre de 2013

ISSN: 1135-5948

Preamble

The *Natural Language Processing* journal aims to be a forum for the publication of quality unpublished scientific and technical papers on Natural Language Processing (NLP) for both the national and international scientific community and companies. Furthermore, we want to strengthen the development of different areas related to NLP, widening the dissemination of research carried out, identifying the future directions of basic research and demonstrating the possibilities of its application in this field. Every year, the Spanish Society for Natural Language Processing (SEPLN) publishes two issues of the journal that include original articles, ongoing projects, book reviews and the summaries of doctoral theses. All issues published are freely distributed to all members, and contents are freely available online.

The subject areas addressed are the following:

- Linguistic, Mathematical and Psychological models to language
- Grammars and Formalisms for Morphological and Syntactic Analysis
- Semantics, Pragmatics and Discourse
- Computational Lexicography and Terminology
- Linguistic resources and tools
- Corpus Linguistics
- Speech Recognition and Synthesis
- Dialogue Systems
- Machine Translation
- Word Sense Disambiguation
- Machine Learning in NLP
- Monolingual and multilingual Text Generation
- Information Extraction and Information Retrieval
- Question Answering
- Automatic Text Analysis
- Automatic Summarization
- NLP Resources for Learning
- NLP for languages with limited resources
- Business Applications of NLP
- Sentiment Analysis
- Opinion Mining
- Text Mining
- Evaluation of NLP systems
- Textual Entailment and Paraphrases

The 51th issue of the *Procesamiento del Lenguaje Natural* journal contains scientific papers and investigation projects and tools descriptions summaries. All of these were accepted by the traditional peer reviewed process. We would like to thank the Advisory Committee members and additional reviewers for their work. Fifty-three papers were submitted for this issue of which forty-five were scientific papers and eight investigation projects and tools descriptions summaries. From these forty-five papers, we selected twenty-two (48.9% for publication).

Authors from other 10 countries have submitted papers to the journal. These countries are: Germany, England, Scotland, Ireland, Iran, Cuba, Mexico, USA, Brazil and Ecuador.

The Advisory Committee of the journal has reviewed the papers in a double-blind process. Under double-blind review the identity of the reviewers and the authors are hidden from each other. In the first step, each paper was reviewed blindly by three reviewers. In the second step, the three reviewers have given a second overall evaluation to those papers with a difference of three or more points out of 7 in their individual reviewer scores. Finally, the evaluation of those papers that were in a position very close to the acceptance limit were supervised by the editorial board.

We believe that the quality of the articles is high. The cut-off criteria adopted was the average of the three scores given.

September 2013
Editorial board

Procesamiento del Lenguaje Natural, Revista nº **51**, septiembre de 2013

sociedad española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT | FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA | 2012

ISSN: 1135-5948

Artículos

Artículos

Traducción Automática

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 15-22 recibido 27-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Peeking through the language barrier: the development of a free/open-source gisting system for Basque to English based on apertium.org*

Asomándose por encima de la barrera lingüística: desarollo de un sistema libre/de código abierto euskera–inglés para la asimilación basado en apertium.org

Jim O'Regan
Eolaistriu Technologies
Thurles, Tipperary (Ireland)
joregan@gmail.com

Mikel L. Forcada
Dept. Lleng. i Sist. Informàtics
Universitat d'Alacant
E-03071 Alacant (Spain)
mlf@dlsi.ua.es

Resumen: El artículo describe el desarrollo de un sistema de traducción automática del euskera al inglés pensado para la asimilación (comprensión) construido sobre la plataforma de traducción automática libre/de código fuente abierto basada en reglas Apertium, y lo evalúa preliminarmente usando un nuevo método basado en tests de clausura (*cloze tests*) en los que se pide que se rellenen huecos en una traducción de referencia. Los resultados indican que la disponibilidad de las traducciones en bruto producidas por un sistema con un diccionario de unas 10.000 entradas y unas 300 reglas de traducción incrementan significativamente la capacidad de quien lee para completar los tests con éxito.
Palabras clave: traducción automática, euskera, inglés, software libre, código abierto, asimilación, evaluación

Abstract: The article describes the development of a machine translation system from Basque to English designed for assimilation (*gisting*) built on the free/open-source rule-based machine translation platform Apertium, and evaluates it preliminarly using a new method based in Cloze tests in which readers are asked to fill out gaps in a reference translation. The results indicate that the availability of the raw translations by a system with a dictionary of about 10,000 entries and about 300 translation rules increase significantly the ability of readers to complete the tests successfully.
Keywords: Machine translation, Basque, English, Free Software, Open Source, Gisting, Assimilation, Evaluation

1 Introduction

Machine translation (MT) has become a viable technology that may help individuals in assimilation (to get the gist of a text written in a language the reader does not understand) and dissemination (to produce a draft translation to be post-edited for publication) tasks.

An unfortunate reality for the Basque people on a global scale is that they are little known, and very often for the political conflict and the (fortunately past) violent activity of armed Basque separatist groups. A Basque to English translator would allow the wider world to get a better knowledge about the Basque people and culture through translations of their media (such as the digital edition of the Basque-language daily Berria, http://www.berria.info). Although Google Translate does provide Basque translation, we feel that a free/open-source, rule-based alternative is not only justified, but desirable, for three reasons. Firstly, Google's system often favours fluency over fidelity. Even between larger languages, such as English and Spanish, Google Translate can make errors such as translating *billón* to *million* or removing words from translations (such as *not*) with severe damage to meaning. Secondly, Google's data harvesting has given them something of a 'Big Brother' reputation; to those concerned with privacy, Google

* Development was funded through a grant of the European Association for Machine Translation

ISSN 1135-5948

products are not an option — for a discussion of the risks of online MT see (Porsiel, 2008). Thirdly, having a free/open-source rule-based alternative means that linguistic knowledge (dictionaries, rules) will be explicitly encoded and will therefore be available for reuse in other technologies, thanks to the licence used.

This article describes the development of such a MT system from Basque to English, aimed primarily at assimilation ("gisting"), based on the Apertium[1] free/open-source rule-based MT platform (described in section 2.1), similar to the Basque–Spanish system existing in the project. This involves the creation of a language-pair package which includes monolingual Basque and English dictionaries, a bilingual dictionary, and a set of translation (*structural transfer*) rules. As in the case of the existing Basque to Spanish system (Ginestí-Rosell et al., 2009), the aim was to build a prototype language pair with around 6,000 lemmata in its bilingual dictionary and about 200 structural transfer rules (existing data are described in 2.2; the resulting lexicon and translation rules are described in sections 2.3–2.5. Secondary goals include the creation of the lexical resources necessary for a possible future English to Basque system, and of resources which may be used to augment future work in statistical MT, by using a subset of the rules to generate phrase candidates, to overcome the data sparseness problem (Tyers, 2009).

The evaluation of MT for assimilation purposes is an open research subject which, in our opinion, has not received the attention it deserves in view of the fact that internet-based MT is massively used for assimilation — one could indeed say that assimilation is the most frequent application of MT nowadays. The obvious choice of setting up reading comprehension tests such as multiple choice tests (Jones et al., 2007) is costly and labour-intensive; therefore, alternative ways have been thought. For instance, the evaluation of the Basque to Spanish system (Ginestí-Rosell et al., 2009) involved a two-step procedure very similar to that used in the WMT 2009[2] and WMT 2010[3] workshops on MT (Callison-Burch et al., 2009; Callison-Burch et al., 2010): first, the raw MT output (and nothing else) was shown to target-language monolingual who had to do their best to guess what the sentence meant and post-edit it into a fluent sentence with the same meaning. Then, bilingual people would compare the resulting target sentence with the source sentence and subjectively rate how good a translation it was. This was still a rather costly procedure, required the availability of bilingual experts, and was affected by subjective judgements by bilinguals. For a preliminar evaluation the resulting prototype, we have designed and implemented a novel method for the evaluation of the usefulness of the MT for assimilation or *gisting*, which is described in section 3. To the best of our knowledge, this is the first time that Cloze or gap-filling tests have been used to evaluate the informativeness of MT: note that Cloze tests have indeed been used but in a different way (Somers and Wild, 2000): readability of raw MT was measured by introducing gaps in it and having the subjects fill the gaps. The method described in this paper fills gaps in a reference translation instead. The results indicate that the availability of the raw translations by a system with a dictionary of about 10,000 entries and about 300 translation rules increase significantly the ability of readers to complete the tests successfully.

The paper ends up describing future work (section 4) and giving some concluding remarks (section 5).

2 Development

2.1 The Apertium platform

The system is based on the Apertium MT platform. Originally designed for the Romance languages of Spain, the platform has been extended to support other, more divergent language pairs, including a Basque to Spanish translator (Ginestí-Rosell et al., 2009). Apertium is licensed under the GNU General Public License,[4] as published by the Free Software Foundation, and all software, data, and related source code, for the engine, tools, and all 33 supported language pairs are available to download from the project website.

Apertium uses a shallow-transfer engine. Finite-state transducers are used for lexical processing, hidden Markov models are used for part-of-speech tagging, and finite-state based chunking is used for structural transfer. Linguistic data is encoded in standard-

[1] `http://www.apertium.org`

[2] `http://www.statmt.org/wmt09/`

[3] `http://www.statmt.org/wmt10/`

[4] `http://www.fsf.org/licensing/licenses/gpl.html`

ised, XML-based formats, which are transformed into efficient binary representations using tools which accompany the engine. A full description of the Apertium engine is available in Forcada et al. (2011); what follows is a brief description of the components of the engine.

2.1.1 The pipeline

A typical Apertium-based translation system uses a pipeline of eight components.

- The **deformatter** separates the text to be translated from the formatting markup (e.g., HTML, RTF, OpenDocument, etc.). The format information is encapsulated within "superblanks", which are treated the same as blank characters by the rest of the system.
- The **morphological analyser** tokenises the text into *surface forms*, and adds to each surface form one or more *lexical forms*, containing the *lemma*, and a sequence of tags describing the lexical category and morphological inflection. The analyser is able to analyse both contractions, and fixed-length multi-word units. Multi-word units are processed as "words with spaces", which may themselves be inflected.
- The **part-of-speech tagger** uses a pretrained hidden Markov model (HMM) to determine which of a list of ambiguous analyses is the most likely in the context of its neighbours, delivering a single lexical form.
- The **structural transfer module** uses finite-state pattern matching to detect fixed-length patterns of lexical forms that need special processing due to divergences between the languages: the insertion or deletion of words, reordering, agreement operations (number, gender, person, etc.).
- The **lexical transfer module** is called by the structural transfer module for each source language lexical form, delivering a target-language lexical form.
- The **morphological generator** produces the target-language surface form from each target-language lexical form by inflecting it according to the operations that have been carried out by the structural transfer module.
- The **post-generator** performs orthographical operations, such as contractions.
- Finally, the **reformatter** restores the format information encapsulated by the deformatter module into the translated text, extracting it from the superblanks.

In the case of more complicated language pairs, such as Basque to English, the structural transfer module is split into three components:

- a **chunker**, which performs lexical transfer and local syntactic operations, and segments the input sequence of lexical units into *chunks*; that is, sequences of lexical forms corresponding to some syntactic constituent or feature, such as a verb phrase;
- an **interchunk** module, which performs more global operations on and among chunks; and
- and a **postchunk** module, which performs further local operations on each chunk, before restoring the contents of the chunk to the text stream, to be processed by the morphological generator.

All modules of the system communicate using text streams, via Unix pipes. Individual modules can be replaced, or additional modules added, without any architectural changes to the engine. This allows individual modules to be reused for other natural language processing tasks; for example, the morphological analysis and part-of-speech tagger modules have been reused for opinion classification (Bonev, Ramírez-Sánchez, and Rojas, 2012).

2.2 Existing data

We were able to reuse monolingual data from the Apertium Basque to Spanish system (Ginestí-Rosell et al., 2009), which in turn is based on data from the Matxin (Mayor et al., 2011) Spanish to Basque system.

The Basque morphological dictionary was almost identical to that in the Basque–Spanish system, aside from the addition of extra entries, and the correction of some erroneous entries. Although we added some new inflectional paradigms (for example, for place names ending in *-m* which were

missing from the Basque—Spanish analyser), most of the new paradigms were added for number handling.

The dictionaries for the initial system were built by crossing (or "triangulating") existing Basque to Spanish and Spanish to English lexical data in the Apertium platform using `apertium-dixtools`;[5] the resulting dictionaries were then manually extended for coverage and tuned to yield the most useful lexical equivalents. The creation of the lexicon is further described in section 2.3.

Although the system described here is rule-based, and therefore not corpus-based, it is desirable for many reasons to have a bilingual corpus to aid in the development of the system: to use statistical word alignment tools, to aid in lexicon creation; as a basis for comparison while checking the output of rules; and as a source with which to build a statistical system for the purpose of comparative evaluation; to name but a few. It is also desirable that this corpus be made available under an open-content licence, to be able to distribute it along with the translator[6]

The creation of a bilingual corpus has proved more difficult than expected, as there is hardly any parallel text with a free/open-source licence. Only very recently, two corpora have been made available from OPUS (Tiedemann, 2009),[7] one containing messages from the KDE4 Desktop ("KDE4") and the other one containing film subtitles ("OpenSubtitles2011"); these corpora are not suitable for evaluation: the first one is too dirty and contains only a few complete sentences, and the second one has translations that are very far from being *literal* enough for MT uses. Therefore, two corpora have been built; one is available in the project's repository[8] and contains the Basque and English version of Kafka's *Metamorphosis* - this corpus is cleaner and was used to guide the building of dictionaries, etc., but was not used for evaluation in view of the fact that many translations were found to be problematic; the other corpus was harvested from the website of the International Contact Group for the Basque Country[9] and, after filtering, is a very small corpus (223 sentences); as of today we have not obtained permission to distribute it, but is available on request from the authors.

2.3 Lexicon

The lexicon in the Basque to Spanish translator was based on the lexicon from Matxin, which has a prototype English to Basque translator available from its source repository (Mayor and Tyers, 2009). Despite the availability of the lexicon of this system, we opted not to use it initially: as with Matxin Spanish to Basque, only a small portion of the data is available under a free/open-source licence; of that data, many of the English translations were quite obscure. We chose instead to create our initial lexicon by triangulation, using the Basque to Spanish lexicon and the lexicon from Apertium's English to Spanish translator, and `apertium-dixtools`.

`apertium-dixtools` uses tags, such as those representing lexical category, to restrict the triangulation, thus reducing the number of incorrectly generated entries; that is, the Basque–Spanish entry `izaki<n>:ser<n>` will be aligned with the Spanish–English entry `ser<n>:being<n>`, and not with `ser<v>:be<v>`. In addition, as entries are processed, the software creates a model of the patterns of combinations of tags observed in the source dictionaries, which it sorts by frequency. The crossing model it generates can then be incrementally refined, to generate dictionary entries as close to handwritten entries as possible.

We ran through 6 iterations of this refinement process until we reached the singleton entries which needed to be edited by hand (the majority of these entries were closed category words, though a few were the result of errors in one of the source dictionaries). To keep the Basque–English translator as close as possible to the Basque–Spanish translator, we made use of the greater availability of English–Spanish resources by adding missing entries to the English–Spanish dictionary, and extracted them from a further iteration.

Despite the reduction in triangulation errors achieved by using `apertium-dixtools`, there were still errors due to polysemy and gender ambiguity introduced by the pivot language (for instance, where Spanish nouns in-

[5] `http://wiki.apertium.org/wiki/Apertium-dixtools`

[6] Apertium's development host, Sourceforge, requires that all content distributed via its infrastructure be under a free/open licence.

[7] `http://opus.lingfil.uu.se/`

[8] `http://sourceforge.net/projects/apertium/files/apertium-eu-en/0.3.0/metamorphosis.tmx/download`

[9] `http://hnteuskalherria.org/`

flect like adjectives, such as *hermano* and *hermana*, dictionaries treat the feminine noun as an inflected form of the masculine, leading to an artificial ambiguity).

We used the data from Matxin to automatically resolve *true* polysemy, and to assist in manually resolving the *artificial ambiguity*. Although at the time of development Apertium lacked a lexical selection mechanism for polysemous words,[10] we took this opportunity to annotate the polysemous words we found, for future use. We similarly used Google Translate and online dictionaries to add more data from Matxin (approx. 2,000 words) and from Wikipedia (approx. 1,000 words), though we later discovered that a number of those later additions were duplicates (approx. 500).

The released version of the bilingual lexicon contains about 10,000 entries, of which approx. 350 cannot be reached, because the source-language (Basque) morphological analyser has no corresponding entry. The Basque–Spanish lexicon contains a large number of entries, the majority of which cannot be reached. We estimate it at around 7,000 words.

2.4 Part-of-speech tagger

The Basque part-of-speech tagger in `apertium-eu-es` was used without modification.

2.5 Structural transfer rules

An initial set of structural transfer rules was built by adapting existing rules in the Apertium Basque to Spanish system and then manually extended to deal with the most frequent structural mismatches between English and Basque. The transfer rules were kept as close as possible to the originals, though in most cases the output was quite different, to adapt to English syntax. Most of the new transfer rules were added to handle verb negation, where no rule was required for Spanish. A number of new rules were added to handle phrases expressing dates, and to better handle hyphenated noun–noun constructs.

The existing rules were extended to generate the subject for verbs[11] and to generate auxiliaries; unhyphenated *noun noun* constructs were extended to choose among $noun_1$ $noun_2$, $noun_1$ *'s* $noun_2$ and $noun_2$ *of* $noun_1$ as could be determined. Triple noun phrases were left as-is ($noun_3$ *of* $noun_2$ *of* $noun_1$) as it was the best option, given the data available, though we presume that the availability of further data will show that the possibilities are as varied as for double noun phrases.

[10] A feasible lexical selection module capable of using hand-written rules or rules inferred from a corpus has been only recently described and is in the process of being released (Tyers, Sánchez-Martínez, and Forcada, 2012).

[11] Both Basque and Spanish are pro-drop, so this was not needed in the Basque–Spanish translator.

3 Preliminary evaluation

Since this system was basically aimed at helping English speakers to understand Basque text (assimilation or "gisting"), evaluation has tried to measure the ability of readers to make sense of sentences extracted from documents produced by the International Contact Group cited above. Therefore we recruited (mostly through the mailing list for the Apertium project) 27 people with a good command of English that did not have any command whatsoever of Basque. Of these, 23 people responded by the deadline given.

To that end, we have used a simple method. *Holes* or *gaps* were created in the English version of each sentence by blanking out 20% of the words that were not stopwords to create *Cloze* tests (Taylor, 1953) where subjects had to try their best to guess the missing words and fill them in. In preliminary tests, the 20% percentage was shown to be safely beyond the point where monolingual guessing may be successful. For some randomly chosen sentences, the following hints were given: the original Basque sentence (a weak hint which, however, could be expected to be useful to fill proper names or cognate words), the output of the `apertium-en-es` MT system (which, in case of being maximally useful, would provide information to fill all holes), or both. Each one of the two hints were given with 50% probability separately. Each informer got 32 sentences, roughly 8 in each category.

Instead of using Apertium-eu-en version 0.3.0, released in November, a slightly improved version (subversion repository revision 36906, as of March 24, 2012). Table 2 shows an example where both hints are given.

The sentences rebuilt by the subjects were then compared to the actual English reference sentences and separate success rates were obtained for each of the 4 levels of hinting.

A synonym list (86 entries) was built by manually inspecting the mismatches between words filled in by the informants and those in

Item	Count
Number of bidirectional bilingual dictionary entries	9,565
Number of specific Basque–English bilingual dictionary entries (not to be used in future English–Basque MT)	29
Number of specific English–Basque bilingual dictionary entries (unused in Basque–English MT, ready for future development)	4,727
Number of entries used for Basque–English	9,594
Total dictionary entries	14,321
Structural transfer: chunking rules	197
Structural transfer: inter-chunk rules	55
Structural transfer: post-chunk rules	20
Total structural transfer rules	272

Table 1: Current status of bilingual dictionaries and rules (revision 36906 of `apertium-eu-en`)

Test number 3	
Basque (source language) hint:	`Bruselako Adierazpenaren sinatzaileek argi eta garbi zuzendu dute Adierazpen horrek ordezkatzen duen nazioarteko komunitatearen eskaera.`
Machine translation hint:	`the signatories of the Statement of Brussels clear and clean they have addressed this Statement he of the international community that substitutes the request.`
Problem sentence:	`[sm]@@143: The ##### of the ##### Declaration have addressed in ##### ##### the demands of the international ##### which the ##### Declaration represents.`
Reference sentence:	`The endorsers of the Brussels Declaration have addressed in unequivocal terms the demands of the international community which the Brussels Declaration represents.`

Table 2: Evaluation of `apertium-eu-en`: two types of hints, problem sentence with gaps, and reference sentence with the gaps filled.

the reference sentences. A very conservative approach was taken to select these allowed substitutions (the only substitutions allowed were those not affecting the sense of the reference: actual synonyms in place, alternative spellings, reference-side and informant misspellings, etc.). Table 3 contains some entries of the synonym list.

The number of holes that informants were able to fill out successfully without hints (the baseline) is relatively high. The context provided by other sentences and the homogeneous nature of the corpus (texts related to talks to help the resolution of the conflict in the Basque Country) may be a reason for that. A small increase in success was observed when the Basque source sentence was given as the only hint.

The increase obtained when the output of the Apertium-eu-en machine translation system was presented to the informants is reasonably significant, but lower than expected. Part of the small difference may be due to the fact that the baseline was already rather high.

The most surprising result is that the performance decreases with both the machine-translated and the source sentence are shown to the informants. The decrease is actually quite systematic. This may be due to an "information glut" effect: too much information to integrate.

Hint level	Total number of one-word holes filled	% holes succesfully filled (exactly as in reference)	% holes succesfully filled (using a short synonym list)
No hint	575	26% (sd 13%)	30% (sd 14%)
Source hint (Basque sentence)	543	29% (sd 12%)	34% (sd 14%)
Machine translation hint (Apertium translation)	597	48% (sd 13%)	54% (sd 13%)
Both source and machine translation hint	589	43% (sd 13%)	51% (sd 14%)

Table 4: The results obtained (averages and standard deviations between informants ("sd") are shown).

mesures	measures
mandate	Mandate
likewise	also
legalization	legalisation
lawful	legitimate
laid	set
kept	maintained
international	International
HNT	ICG
financial	economic
evaluation	assessment
end	cessation
encourage	facilitate
demonstrate	prove
declaration	statement
change	transformation
big	major
agents	stakeholders
affected	hit
action	practice
amnisty	amnesty
to	To
truly	sincerely
statemet	declaration
Spain's	Spanish
richest	wealthiest

Table 3: A selection of the synonyms used during the evaluation

When using the synonym list, all results improve slightly, but the trend is the same as when no synonyms are used.

4 Future work

As regards evaluation, this is the first time that MT has been evaluated by measuring its usefulness as a hint in a Cloze test — until now, Cloze tests used for machine translation evaluation asked informants to fill gaps in the MT output itself (Somers and Wild, 2000). However, the experiments reported are still preliminary, and avenues for improvement can easily be identified. For instance, the evaluation corpus was too homogeneous and therefore constituted a very powerful hint, as shown by the results. A general-purpose or multi-domain corpus with wider vocabulary sentences would definitely evaluate it better. Another task that we plan to perform is a comparison task in which Apertium-translated and Google Translate-translated hints were compared.

5 Concluding remarks

This paper has described the building of a free/open-source machine translation system from Basque to English, based on the Apertium platform. The system aims at being a useful tool for assimilation or 'gisting' purposes.

The preliminary results show that it is possible to build, in a few months, a Basque to English MT system capable of producing translations that measurably improve the level of understanding, on the part of non-Basque speakers, of the contents of Basque text.

Acknowledgements: The authors thank the European Association for Machine Translation for financial support, all 23 informants for participating in the evaluation, Francis M. Tyers for useful discussions on the idea of using Cloze tests for machine translation evaluation, and Mireia Ginestí for her help on questions regarding Basque data.

References

Bonev, Boyan, Gema Ramírez-Sánchez, and Sergio Ortiz Rojas. 2012. Opinum: statistical sentiment analysis for opinion classification. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 29–37. Association for Computational Linguistics.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden, July. Association for Computational Linguistics. Revised August 2010.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March. Association for Computational Linguistics.

Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Ginestí-Rosell, Mireia, Gema Ramírez-Sánchez, Sergio Ortiz-Rojas, Francis M Tyers, and Mikel L Forcada. 2009. Development of a free Basque to Spanish machine translation system. *Procesamiento del Lenguaje Natural*, 43:187–195.

Jones, Douglas, Martha Herzog, Hussny Ibrahim, Arvind Jairam, Wade Shen, Edward Gibson, and Michael Emonts. 2007. ILR-based MT comprehension test with multi-level questions. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 77–80. Association for Computational Linguistics.

Mayor, Aingeru, Iñaki Alegria, Arantza Díaz De Ilarraza, Gorka Labaka, Mikel Lersundi, and Kepa Sarasola. 2011. Matxin, an open-source rule-based machine translation system for Basque. *Machine Translation*, 25(1):53–82.

Mayor, Aingeru and Francis M. Tyers. 2009. Matxin: moving towards language independence. In *Proceedings of the first international workshop on free/open-source rule-based machine translation, Alacant*, pages 11–17.

Porsiel, Jörg. 2008. Machine translation at volkswagen: a case study. *Multilingual Computing & Technology*, 100.

Somers, Harold and Elizabeth Wild. 2000. Evaluating machine translation: the cloze procedure revisited. In *Translating and the Computer 22: proceedings of the Twenty-second International Conference on Translating and the Computer, 16-17 November 2000*.

Taylor, Wilson L. 1953. "Cloze procedure": a new tool for measuring readability. *Journalism Quarterly*, 30:415–433.

Tiedemann, Jörg. 2009. News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume 5, pages 237–248.

Tyers, Francis M. 2009. Rule-based augmentation of training data in breton-french statistical machine translation. In *Proceedings of the 13th Annual Conference of the European Association of Machine Translation, EAMT09*, pages 213–218.

Tyers, Francis M., Felipe Sánchez-Martínez, and Mikel L. Forcada. 2012. Flexible finite-state lexical selection for rule-based machine translation. In Mauro Cettolo, Marcello Federico, Lucia Specia, and Andy Way, editors, *EAMT 2012: Proceedings of the 16th Annual Conference of the European Association for Machine Translation, Trento, Italy, May 28-30 2012*, pages 213–220.

Extracción y Recuperación de Información

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 25-32 recibido 03-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Consultas con Errores Ortográficos en RI Multilingüe: Análisis y Tratamiento*

Misspelled Queries in Cross-Language IR: Analysis and Management

David Vilares Calvo
Depto. de Computación
Universidade da Coruña
Campus de Elviña s/n
15071 – A Coruña
david.vilares@udc.es

Adrián Blanco González
Depto. de Informática
Universidade de Vigo
Campus As Lagoas s/n
32004 – Ourense
adbgonzalez@uvigo.es

Jesús Vilares Ferro
Depto. de Computación
Universidade da Coruña
Campus de Elviña s/n
15071 – A Coruña
jvilares@udc.es

Resumen: Este artículo estudia el impacto de los errores ortográficos en las consultas sobre el rendimiento de los sistemas de recuperación de información multilingüe, proponiendo dos estrategias para su tratamiento: el empleo de técnicas de corrección ortográfica automática y la utilización de n-gramas de caracteres como términos índice y unidad de traducción, para así aprovecharnos de su robustez inherente. Los resultados demuestran la sensibilidad de estos sistemas frente a dichos errores así como la efectividad de las soluciones propuestas. Hasta donde alcanza nuestro conocimiento no existen trabajos similares en el ámbito multilingüe.
Palabras clave: Recuperación de información multilingüe; traducción automática; errores ortográficos; corrección ortográfica; n-gramas de caracteres.

Abstract: This paper studies the impact of misspelled queries on the performance of Cross-Language Information Retrieval systems and proposes two strategies for dealing with them: the use of automatic spelling correction techniques and the use of character n-grams both as index terms and translation units, thus allowing to take advantage of their inherent robustness. Our results demonstrate the sensitivity of these systems to such errors and the effectiveness of the proposed solutions. To the best of our knowledge there are no similar jobs in the cross-language field.
Keywords: Cross-language information retrieval; machine translation; misspellings; spelling correction; character n-grams.

1. Introducción

En el marco actual de globalización de la red muchas veces un documento relevante para el usuario está escrito en una lengua diferente a la suya. Como respuesta a esta problemática surge la *recuperación de información multilingüe* (RIM)[1] (Nie, 2010), un caso particular dentro de la *recuperación de información* (RI) en el que consultas y documentos están en idiomas diferentes.

Para ello la mayoría de dichos sistemas introducen algún tipo de *fase de traducción* intermedia que permita reducir el problema original a la clásica RI monolingüe con consultas y documentos en el mismo idioma. Debido a limitaciones prácticas suele optarse por traducir las consultas de su idioma original (denominado *origen*) al de los documentos (denominado *destino*) (Nie, 2010).

Fruto de esta misma globalización cada vez es más necesario disponer de sistemas capaces de operar sobre textos con *errores ortográficos*,[2] en particular en el caso de las consultas (Guo et al., 2008). Esto se debe a que los modelos formales de RI fueron diseñados para operar sobre textos sin errores, por lo que su presencia dañará substancialmente el rendimiento. Hablaremos entonces de *recuperación de información tolerante a errores*[3] (Manning, Raghavan, y Schütze, 2008, Cap. 3).

En este contexto nuestro trabajo aborda el

* Trabajo parcialmente subvencionado por el Ministerio de Economía y Competitividad y FEDER (proyectos TIN2010-18552-C03-01 y TIN2010-18552-C03-02) y por la Xunta de Galicia (ayudas CN 2012/008, CN 2012/317 y CN 2012/319).

[1] *Cross-language information retrieval* (CLIR)

[2] Tanto aquéllos fruto del desconocimiento de la ortografía como errores tipográficos o producto del ruido durante su generación (ej. OCR) (Kukich, 1992).

[3] *Tolerant Information Retrieval*.

ISSN 1135-5948

estudio del impacto de los errores ortográficos en las consultas sobre el proceso de recuperación multilingüe así como el diseño de entornos robustos capaces de operar en ese contexto.

El tratamiento de consultas mal escritas suele basarse en sustituir o modificar el algoritmo de búsqueda de correspondencias exactas original para permitir correspondencias aproximadas. Conforme al estado del arte, consideramos dos estrategias genéricas diferentes (Manning, Raghavan, y Schütze, 2008): una que opera a nivel de palabra y otra a nivel de subpalabra.

La primera de éstas opera a nivel de palabra y consiste en añadir una fase de preprocesamiento para la corrección de los errores ortográficos de la consulta empleando técnicas de *procesamiento del lenguaje natural* (PLN) basadas en diccionarios. Debemos señalar que a diferencia de otros ámbitos clásicos de aplicación de los sistemas de corrección (ej. procesadores de texto), en el caso de RI se requieren soluciones que permitan un tratamiento totalmente automático del error (Kukich, 1992) sin necesidad de la intervención del usuario. Se pueden distinguir dos enfoques: la *corrección de palabras aisladas* en la cual se intenta corregir cada palabra por separado (Savary, 2002), y el aprovechamiento de la información lingüística de su contexto para la corrección (Otero, Graña, y Vilares, 2007).

Una segunda estrategia consiste en emplear n-gramas de caracteres como unidad de procesamiento en lugar de palabras (McNamee y Mayfield, 2004a; Robertson y Willett, 1998).

En este trabajo probaremos ambas aproximaciones con errores humanos reales en un contexto de recuperación *de-español-a-inglés* (consultas en español y documentos en inglés). Hasta donde alcanza nuestro conocimiento no existen trabajos similares con este grado de detalle en el ámbito multilingüe.

La estructura del artículo es como sigue. La Sección 2 aborda nuestras propuestas basadas en corrección, mientras que la Sección 3 describe nuestra propuesta basada en el empleo de n-gramas de caracteres. La Sección 4 detalla nuestro metodología de prueba, obteniendo los resultados recogidos en la Sección 5. Finalmente, la Sección 6 presenta nuestras conclusiones y propuestas de trabajo futuro.

2. *Aproximaciones basadas en corrección ortográfica*

La primera de las estrategias contempladas pasa por preprocesar la consulta empleando técnicas de corrección automática basadas en PLN para detectar y corregir sus errores ortográficos, la cual ha sido ya aplicada con éxito en RI monolingüe (Vilares, Vilares, y Otero, 2011). En nuestro contexto actual la consulta inicial es preprocesada y, una vez corregida, es traducida aplicando técnicas de *traducción automática* (TA)[4] y luego lanzada contra el motor de recuperación. Asimismo se han considerado dos posibles técnicas corrección, aislada y contextual, que describimos a continuación.

2.1. Corrección aislada

Como punto de partida aplicaremos el algoritmo de reparación global propuesto por Savary (2002), capaz de encontrar todas las palabras cuya distancia de edición (Levenshtein, 1966) con la errónea sea mínima; esto es, el número de *operaciones de edición*[5] a aplicar para transformar una cadena en otra.

Este algoritmo tiene como núcleo un *autómata finito* (AF) que reconoce el léxico del idioma considerado. Para cada palabra a procesar el AF intenta reconocerla intentando ir desde el estado inicial a uno final a través de las transiciones etiquetadas con los caracteres de la cadena de entrada. Si el AF se detiene en un estado por no haber transiciones de salida etiquetadas con el siguiente carácter de la entrada, es que se ha detectado un error ortográfico. Pasamos entonces a aplicar sobre la configuración actual del autómata cuatro posibles *hipótesis de reparación* elemental, cada una de ellas correspondientes a una operación elemental (inserción, borrado, sustitución y transposición) y con un coste asociado para así intentar alcanzar una nueva configuración que nos permita continuar con el proceso de reconocimiento. Dichas operaciones se aplican recursivamente hasta alcanzar una configuración correcta. El algoritmo también reduce dinámicamente el espacio de búsqueda, quedándose en todo momento únicamente con las correcciones mínimas y tratando de alcanzar la primera solución tan pronto como sea posible.

[4] *Machine translation* (MT).

[5] Inserción, borrado o substitución de un carácter, o transposición de dos caracteres contiguos.

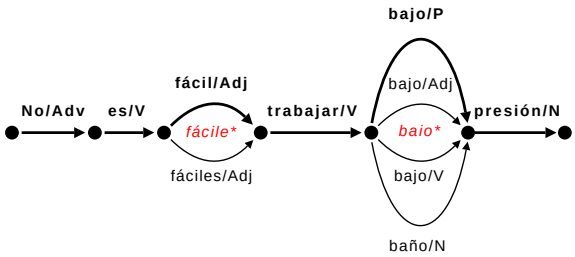

Figura 1: Alternativas de corrección ortográfica representadas en una retícula (secuencia correcta resaltada).

Desafortunadamente, este algoritmo puede devolver varias correcciones candidatas si existen varias palabras a distancia mínima de edición de la palabra original. Tomemos como ejemplo la frase *"No es fácile* trabajar baio* presión"*,[6] representada junto con sus posibles correcciones en la Figura 2.1; en este caso el algoritmo devolvería dos posibles correcciones candidatas para *"fácile*"* (*"fácil"* y *"fáciles"*) y otros dos para *"baio*"* (*"bajo"* y *"baño"*). Asimismo este algoritmo está limitado a aquellos errores correspondientes a palabras no válidas, por lo que puede fallar a la hora de detectar errores que dan lugar a otras palabras válidas del idioma, como podría ser el caso de la consulta *"materiales compuestos ligueros*"*[7] pues tomados por separado cada uno de sus términos es válido.

2.2. Corrección contextual

Existe a su vez una extensión al algoritmo anterior, que denominaremos de *corrección contextual*, que permite emplear el contexto lingüístico de la palabra a corregir para resolver las limitaciones del algoritmo original (Otero, Graña, y Vilares, 2007). Para ello se utiliza la información lingüística contextual embebida en un proceso de etiquetación con el fin de podar las correcciones candidatas de tal forma que sólo se acepten aquéllas que encajen en el contexto morfosintáctico de la palabra a corregir.

Este modelo emplea un etiquetador morfosintáctico estocástico basado en una extensión dinámica del algoritmo de Viterbi sobre *Modelos Ocultos de Markov* (Graña, Alonso, y Vilares, 2002) de segundo orden que se aplica sobre retículas en lugar de enrejados, haciéndola mucho más flexible al permitirnos representar un par *palabra/etiqueta* en cada arco, y luego calcular la secuencia más probable mediante una adaptación del algoritmo de Viterbi, como se muestra en la Figura 2.1.[8]

Así, y ya restringiéndonos al ejemplo de dicha figura, las correcciones devueltas serían únicamente las correspondientes a la secuencia de etiquetas correcta: *"fácil"* (`Adj`) para *"fácile*"* y *"bajo"* (`P`) para *"baio*"*.

3. Aproximaciones basadas en n-gramas

Un *n-grama de caracteres* es una secuencia de n caracteres dentro de una palabra. De esta forma `tomato` se descompone en los 3-gramas superpuestos: `-tom-`, `-oma-`, `-mat-` y `-ato-`. Las ventajas que aporta el tratamiento a nivel de n-grama —simplicidad, eficiencia, robustez, completitud e independencia del dominio— lo han convertido en una técnica habitual en el procesamiento de textos (Robertson y Willett, 1998; Vilares, Vilares, y Otero, 2011). En el caso concreto de RI, los sistemas clásicos suelen emplear conocimiento y recursos lingüísticos tales como listas de *stopwords*, *stemmers*, lexicones, tesauros, etiquetadores, etc. (McNamee y Mayfield, 2004a), mientras que la *tokenización* en n-gramas no emplea ninguno: el texto es meramente dividido en n-gramas superpuestos (McNamee y Mayfield, 2004b), que son procesados por el motor de recuperación como cualquier otro término. Se trata, pues, de una aproximación independiente del idioma y del dominio, donde además el empleo de correspondencias a nivel de n-grama constituye en sí mismo un mecanismo de normalización que permite trabajar con gran variedad de idiomas sin procesamiento alguno a mayores (McNamee y Mayfield, 2004b; Robertson y Willett, 1998; McNamee y Mayfield, 2004a). Es además un proceso robusto, debido a la redundancia introducida por el proceso de *tokenización* (Vilares, Vilares, y Otero, 2011).

3.1. RIM basada en n-gramas

En el caso de RIM, sin embargo, tales ventajas quedan comprometidas por el proceso de traducción, que debe hacerse a nivel de palabra o frase, pudiendo *tokenizarse* la consulta en n-gramas sólo tras traducirse siendo

[6] Por *"No es fácil trabajar bajo presión"* y donde los asteriscos señalan palabras mal escritas.

[7] Por *"materiales compuestos ligeros"*.

[8] Si bien sería posible emplear el algoritmo original basado en enrejados, su aplicación sería mucho más compleja y costosa (Graña, Barcala, y Vilares, 2002).

además el proceso de traducción muy sensible a los errores ortográficos, palabras desconocidas, falta de recursos lingüísticos apropiados, etc. De este modo, por ejemplo, una palabra mal escrita como "fácile*" no podría ser traducida correctamente, obteniendo[9] "facile*" en lugar de "easy", perdiendo así el procesamiento posterior con n-gramas su capacidad de realizar correspondencias aproximadas. Sólo si también se pudiese traducir a nivel de n-grama[10] podrían beneficiarse plenamente los sistemas de RIM de las ventajas derivadas del uso de n-gramas.

McNamee y Mayfield (2004b) fueron pioneros en este campo, empleando para ello un algoritmo de traducción de n-gramas basado en el alineamiento de corpus paralelos a nivel de n-grama mediante técnicas estadísticas. Posteriormente, Vilares, Oakes, y Vilares (2007) desarrollaron un sistema alternativo que difiere en el proceso de generación de alineamientos, preservando las bondades del sistema previo pero solventando sus principales desventajas.

Estas aproximaciones permiten extender las ventajas del empleo de n-gramas como unidad de procesamiento al proceso de traducción y, consecuentemente, también a los sistemas de RIM, pudiendo así evitar algunas de las limitaciones de las técnicas clásicas, tales como la necesidad de normalizar las palabras o la imposibilidad de traducir palabras desconocidas. Además, al no emplear ningún tipo de procesamiento particular dependiente del idioma, puede aplicarse cuando la disponibilidad de recursos lingüísticos es reducida, lo que, en contra de lo que pueda parecer, no es infrecuente incluso para los principales idiomas europeos (Rehm y Uszkoreit, 2011).

Seguidamente describiremos brevemente el sistema de Vilares, Oakes, y Vilares (2007), que será el que empleemos.

3.2. Traducción de n-gramas

El algoritmo de alineamiento de n-gramas, sobre el que se asienta el sistema de traducción, consta de dos fases. Primero se alinea a nivel de palabra un corpus paralelo de los idiomas origen y destino deseados empleando la herramienta estadística GIZA++ (Och y Ney, 2003), obteniendo las probabilidades de traducción entre las palabras de ambos idiomas. Dicho alineamiento será bidireccional (Koehn, Och, y Marcu, 2003), aceptando un alineamiento $idiomaOrigen \rightarrow idiomaDestino$ (w_s, w_t), donde w_s denota la palabra en lengua origen y w_t su traducción candidata, sólo si existe también el alineamiento inverso $idiomaDestino \rightarrow idiomaOrigen$ (w_t, w_s). Se desecharán también aquellos alineamientos con probabilidad menor de 0,15.

A continuación, en la segunda fase del algoritmo, se realiza el alineamiento a nivel de n-gramas propiamente dicho calculando medidas estadísticas de asociación (Dale, Moisi, y Somers, 2000, Cap. 24) entre los n-gramas contenidos en los pares de palabras alineadas en la fase anterior. Sin embargo, a la hora de realizar los cálculos deberá ponderarse la frecuencia de las observaciones de las coocurrencias de n-gramas en base a las probabilidades de los alineamientos de las palabras que los contienen. Esto se debe a que no se trata de coocurrencias *reales*, sino únicamente *probables*, de forma que una misma palabra origen puede estar alineada con más de una traducción candidata. Tomemos como ejemplo el caso de las palabras en español `leche` y `lechoso`, y las inglesas `milk`, `milky` y `tomato`; un posible alineamiento a nivel de palabra sería:

w_s	w_t	**prob.**
`leche`	`milk`	0,98
`lechoso`	`milky`	0,92
`leche`	`tomato`	0,15

En este caso la frecuencia de coocurrencia del par de n-gramas *(`-lech-`, `-milk-`)* no sería 2, sino 1,90, ya que si bien dicho par coocurre en dos alineamientos a nivel de palabra, *(`leche`, `milk`)* y *(`lechoso`, `milky`)*, dichas coocurrencias se ponderan en base a las probabilidades de sus alineamientos:

$$0{,}98\ (\texttt{leche, milk}) + 0{,}92\ (\texttt{lechoso, milky}) = \mathbf{1{,}90}$$

4. Metodología

4.1. Marco de evaluación

Hemos escogido un contexto de RIM *de-español-a-inglés* (consultas en español y documentos en inglés) por varias razones: (a) la conveniencia de incluir el inglés al ser la lengua dominante en la web; (b) al estar disponible en la web más información en inglés

[9] Empleando Google Translate para traducir de español a inglés.

[10] Realmente no estaríamos ante una *traducción* propiamente dicha desde un punto de vista lingüístico, sino sólo a efectos de recuperación, de ahí que debería hablarse más bien de *pseudo-traducción*.

que en otros idiomas, es lógico que actúe como *idioma destino*; (c) muchos usuarios, aún entendiendo inglés, tienen problemas para expresarse en él, por lo que emplearían su lengua materna como *idioma origen*; y (d) el empleo del español como *lengua origen* se debe a que la gran variedad de procesos morfológicos que presenta hacen de él un buen sujeto de ensayo a la hora de trabajar sobre errores ortográficos (Vilares, Otero, y Graña, 2004).

Respecto al corpus de evaluación, la colección utilizada es la *LA Times 94* (56.472 documentos, 154 MB), empleada en la *robust task* del *ad hoc track* del CLEF 2006 (CLEF, 2013), la cual reciclaba consultas de ediciones anteriores (Di Nunzio et al., 2006).[11] En cuanto a las consultas, se emplearon los 60 *topics* del *conjunto de entrenamiento* para dicha *task*,[12] que constan de tres campos: *título*, un breve título como su nombre indica; *descripción*, una somera frase de descripción; y *narrativa*, un pequeño texto especificando los criterios de relevancia. Con objeto de analizar en mayor detalle la influencia de la longitud de la consulta y la redundancia de la información que ésta contiene, se contemplaron dos series de pruebas de acuerdo a los campos del *topic* empleados para generar la consulta: (1) *consultas cortas:* empleando únicamente el campo *título* (longitud media 2,75 palabras); (2) *consultas de longitud media:* empleando *título* y *descripción* (longitud media 9,88). Por su longitud y complejidad éstas se corresponden con las consultas habituales en motores de búsqueda comerciales y otros sistemas de RI (Bendersky y Croft, 2009; Jansen, Spink, y Saracevic, 2000).

En cuanto a otros recursos empleados, los correctores requieren un diccionario del idioma, y el contextual precisa también de un corpus de entrenamiento para el etiquetador. En ambos casos se ha empleado el corpus español del MULTEXT-JOC (Véronis, 1999), con alrededor de 200.000 palabras etiquetadas y un lexicón de 15.548 términos. En el caso de las subpalabras, se empleó la versión v6 del conocido corpus paralelo EUROPARL (Koehn, 2005), con 51 millones de palabras, para la primera fase del algoritmo de alineamiento de n-gramas

[11] La otra subcolección, la *Glasgow Herald 95* no pudo ser empleada pues al haber sido introducida con posterioridad a la *LA Times 94*, no se dispone de referencias de relevancia de sus documentos (los denominados *qrels*) para gran parte de las consultas.

[12] Topics C050-C059, C070-C079, C100-C109, C120-C129, C150-159 y C180-189.

4.2. Generación de errores

Para evaluar las diversas aproximaciones introduciremos errores ortográficos en los *topics* con una *tasa de error* T creciente:

$$T \in \{0\,\%, 10\,\%, 20\,\%, \ldots, 60\,\%\}$$

donde una tasa T implica que el $T\,\%$ de las palabras incluyen errores,[13] permitiéndonos así emular incluso entornos ruidosos como aquéllos en que la entrada se obtiene de dispositivos móviles o basados en escritura a mano (ej. PDAs, bolígrafos digitales o tabletas digitalizadoras), o de interfaces por habla. Debe tenerse en cuenta que el uso de tasas tan altas no es en absoluto excesivo, pues obsérvese que en consultas cortas, como es el caso de nuestros experimentos, la tasa de error debe ser más alta para que ésta se refleje en las consultas.[14] Hemos empleado además una metodología que permite trabajar con errores humanos reales, mucho más complejos de generar y controlar, pero de mayor interés (Vilares, Vilares, y Otero, 2011).

4.3. Indexación y recuperación

Deberemos diferenciar dos casos en función de la unidad de procesamiento empleada: palabras o n-gramas de caracteres.

En el caso de usar palabras el texto es normalizado de una forma clásica empleando el *stemmer* SNOWBALL,[15] basado en Porter, y las *stopwords* de la UniNE,[16] ambos conocidos y de amplio uso, para luego ser procesado por el motor. A la hora de realizar la consulta, ésta es previamente traducida con Google Translate[17] antes de normalizarla. Consideraremos, a su vez, tres casos: (a) la consulta es traducida tal cual, con errores, siendo ésta nuestra *línea de base* (denotada *stm*); (b) la consulta es corregida previamente con el algoritmo de Savary para palabras aisladas (*Sav*), y en caso de devolver varias correcciones candidatas, se emplean todas; y (c) la consulta es corregida con el algoritmo de corrección contextual (*cont*).

[13] Donde $T{=}0\,\%$ corresponde al *topic* original.

[14] En el caso de consultas de dos palabras se precisaría una tasa del 50 % para que haya de media un error por consulta, y con tres palabras, del 33 %.

[15] `http://snowball.tartarus.org`

[16] `http://members.unine.ch/jacques.savoy/clef/index.html`

[17] `http://translate.google.es`

En el caso de los n-gramas de caracteres (denotado *4gr*), el texto es normalizado pasándolo a minúscula y *tokenizándolo* en 4-gramas (McNamee y Mayfield, 2004b) para ser luego indexado (documentos) o traducido (consultas). En este último caso se ha empleado *log-likelihood* como medida de asociación para el cálculo de alineamientos de n-gramas, para luego, durante la traducción, reemplazar cada n-grama de la consulta original por su n-grama traducción con la medida de asociación más alta (Vilares, Oakes, y Vilares, 2007).

Nótese que no se han empleado técnicas de expansión de la consulta ni de realimentación por relevancia, y así estudiar el comportamiento de las aproximaciones consideradas sin introducir distorsiones en los resultados por la integración de otras técnicas.

El motor de indexación empleado ha sido TERRIER (Ounis et al., 2007), con un modelo de ordenación DFR InL2.[18]

5. Resultados experimentales

Los resultados obtenidos se recogen en el Cuadro 1, mostrando para cada tasa de error T: la precisión media obtenida (MAP); la caída porcentual de dicha precisión respecto a la original —T=0 %— (*%loss*), resaltando en negrita aquellos casos en los que dicha caída es estadísticamente significativa;[19] y el número de consultas, respecto al original, que han dejado de devolver documentos relevantes ($^{[\Delta\varnothing]}$). La media de ambas pérdidas se muestra al final de cada serie de resultados.

En el caso del uso de palabras como unidad de procesamiento, los resultados para nuestra *línea de base* (*stm*) muestran en todos los parámetros empleados un claro impacto negativo de los errores en el comportamiento del sistema, incluso con tasas de error bajas. Al ser mayor la importancia de cada término cuanto más corta la consulta, el impacto es mucho mayor para consultas *cortas*.

El empleo de técnicas de corrección tiene un notable efecto positivo que permite reducir dicha pérdida. En el caso del algoritmo de Savary (*Sav*), éste es muy estable en lo que respecta a la pérdida media de MAP, en torno al 24 % independientemente de la longitud (frente a 34 %/26 % para *stm*), si bien dicha caída tarda más en hacerse significativa para consultas cortas (con $T\geq$40 % vs. $T\geq$30 %). La corrección contextual (*cont*), por contra, se comporta mucho mejor con consultas más largas debido a que el reducido contexto lingüístico de las consultas más cortas limita su efectividad. De este modo en el caso de consultas cortas el algoritmo de Savary se comporta mejor (caída media del MAP de 24 % significativa con $T\geq$40 % vs. 29 % con $T\geq$30 %), mientras que con consultas medias es mejor el contextual (19 % con $T\geq$40 % vs. 24 % con $T\geq$30 %).

En el caso de los n-gramas (*4gr*), los resultados confirman su robustez también en este nuevo ámbito multilingüe, al sufrir una caída de rendimiento claramente menor que en las aproximaciones basadas en palabras, particularmente tanto para consultas cortas como para tasas de error muy altas. No sólo se muestra mucho más robusto que la *línea de base* (*stm*) para palabras (caída del MAP de 11 % vs. 34 % para consultas cortas y 14 % vs. 26 % para medias, siendo dicha caída significativa y "perdiendo" consultas sólo para los T más altos), sino que también supera a las aproximaciones basadas en corrección (caída del 11 % vs. 24 %/29 % para consultas cortas y 14 % vs. 24 %/19 % para medias). Todo ello sin aplicar ningún tipo de procesamiento específico para el tratamiento de errores.

6. Conclusiones y trabajo futuro

Se han estudiado los efectos perniciosos de los errores ortográficos en las consultas en entornos de recuperación de información multilingüe, planteándose dos posibles estrategias para abordar dicha problemática como primer paso hacia el desarrollo de sistemas de información multilingüe más robustos.

En primer lugar, una estrategia clásica basada en el uso de palabras como términos de indexación y unidad de procesamiento. En este caso se ha estudiado el empleo de mecanismos de corrección ortográfica para el tratamiento de los errores en la consulta origen, presentándose dos alternativas. Por una parte, el algoritmo de Savary, que procesa de forma aislada cada término devolviendo todas sus correcciones candidatas a distancia mínima de edición, con el consiguiente riesgo de introducir ruido si devuelve varias, siendo además incapaz de detectar errores que den lugar a palabras existentes. Por otra parte, un algoritmo de corrección contextual que

[18] Frecuencia inversa de documento con normalización 2 de Laplace.

[19] Se han empleado tests-t bilaterales sobre las MAP con α=0,05.

	stm		***Sav***		***cont***		***4gr***	
T	MAP	***%loss*** $^{[\Delta\varnothing]}$	MAP	***%loss*** $^{[\Delta\varnothing]}$	MAP	***%loss*** $^{[\Delta\varnothing]}$	MAP	***%loss*** $^{[\Delta\varnothing]}$
	CONSULTAS CORTAS							
0	0,2705	– –	–	– –	–	– –	0,1637	– –
10	0,2355	-12,94 $^{[-1]}$	0,2617	-3,25 $^{[-1]}$	0,2467	-8,80 $^{[-1]}$	0,1608	-1,77 $^{[0]}$
20	0,2153	**-20,41** $^{[-3]}$	0,2487	-8,06 $^{[-1]}$	0,2362	-12,68 $^{[-1]}$	0,1554	-5,07 $^{[0]}$
30	0,2091	**-22,70** $^{[-6]}$	0,2252	-16,75 $^{[-1]}$	0,2115	**-21,81** $^{[-1]}$	0,1542	-5,80 $^{[0]}$
40	0,1765	**-34,75** $^{[-9]}$	0,2031	**-24,92** $^{[-3]}$	0,1865	**-31,05** $^{[-3]}$	0,1455	-11,12 $^{[0]}$
50	0,1473	**-45,55** $^{[-15]}$	0,1665	**-38,45** $^{[-5]}$	0,1533	**-43,33** $^{[-6]}$	0,1409	**-13,93** $^{[-5]}$
60	0,0945	**-65,06** $^{[-22]}$	0,1360	**-49,72** $^{[-9]}$	0,1249	**-53,83** $^{[-9]}$	0,1193	**-27,12** $^{[-7]}$
media	–	-33,57 $^{[-9,33]}$	–	-23,52 $^{[-3,33]}$	–	-28,58 $^{[-3,50]}$	–	-10,80 $^{[-2,00]}$
	CONSULTAS DE LONGITUD MEDIA							
0	0,3273	– –	–	– –	–	– –	0,2042	– –
10	0,3166	-3,27 $^{[0]}$	0,3128	-4,43 $^{[0]}$	0,3147	-3,85 $^{[0]}$	0,2006	-1,76 $^{[0]}$
20	0,2952	-9,81 $^{[-2]}$	0,2825	-13,69 $^{[-1]}$	0,2917	-10,88 $^{[-1]}$	0,1800	-11,85 $^{[+1]}$
30	0,2604	**-20,44** $^{[-3]}$	0,2712	**-17,14** $^{[-2]}$	0,2890	-11,70 $^{[-2]}$	0,1782	-12,73 $^{[+1]}$
40	0,2339	**-28,54** $^{[-4]}$	0,2570	**-21,48** $^{[-2]}$	0,2655	**-18,88** $^{[-2]}$	0,1782	-12,73 $^{[+1]}$
50	0,2068	**-36,82** $^{[-4]}$	0,2141	**-34,59** $^{[-2]}$	0,2338	**-28,57** $^{[-2]}$	0,1700	-16,75 $^{[+1]}$
60	0,1500	**-54,17** $^{[-11]}$	0,1633	**-50,11** $^{[-3]}$	0,1906	**-41,77** $^{[-3]}$	0,1464	**-28,31** $^{[-2]}$
media	–	-25,51 $^{[-4,00]}$	–	-23,57 $^{[-1,67]}$	–	-19,27 $^{[-1,67]}$	–	-14,02 $^{[+0,33]}$

Cuadro 1: Resultados experimentales.

resuelve dichas limitaciones filtrando las alternativas a partir de información lingüística contextual. Nuestros experimentos demuestran que esta estrategia es muy sensible a los errores en la consulta, particularmente en el caso de consultas cortas, si bien la utilización de mecanismos de corrección permite reducir notablemente sus efectos. Asimismo durante nuestras pruebas el algoritmo de Savary se mostró más apropiado para consultas cortas, mientras que la corrección contextual fue superior para consultas de mayor longitud.

La segunda estrategia propuesta plantea el empleo de n-gramas de caracteres como unidad de procesamiento tanto para indexación como para traducción. Esto nos permite beneficiarnos de la robustez propia del procesamiento a nivel de n-grama y trabajar directamente con la consulta original con errores sin realizar ningún procesamiento a mayores. Este enfoque ha mostrado una gran robustez, con una caída del rendimiento notablemente menor que en el caso de las palabras, aún aplicando mecanismos de corrección. Se podría argumentar que el rendimiento *per se* de esta aproximación basada en n-gramas es menor que el de las aproximaciones clásicas, pero al compararlo con el caso monolingüe (Vilares, Vilares, y Otero, 2011) se observa que buena parte de tal desfase se debe a que los mecanismos de traducción a nivel de subpalabra tienen actualmente un rendimiento menor por estar aún poco desarrollados (Vilares, Oakes, y Vilares, 2007). Al mismo tiempo debe tenerse también en cuenta que se trata de un enfoque *ligero* desde el punto de vista del conocimiento y recursos empleados, y que no se basa en ningún procesamiento particular dependiente del idioma, pudiéndose aplicar para una amplia variedad de idiomas, incluso cuando la disponibilidad de información y recursos lingüísticos sea reducida. Por contra, otros enfoques más clásicos de RIM precisan recursos específicos del idioma como listas de *stopwords*, diccionarios, lematizadores, etiquetadores, corpus de entrenamiento, etc., no siempre disponibles.

De cara al futuro pretendemos mejorar los procesos de indexación-recuperación y traducción de n-gramas con objeto de aumentar el rendimiento del sistema.

Bibliografía

Bendersky, M. y W.B. Croft. 2009. Analysis of long queries in a large scale search log. En *Proc. of WSCD'09*, págs. 8–14. ACM.

CLEF. 2013. `http://www.clef-initiative.eu`.

Dale, R., H. Moisi, y H. Somers, eds. 2000. *Handbook of Natural Language Processing*. Marcel Dekker, Inc.

Di Nunzio, G.M., N. Ferro, T. Mandl, y C. Peters. 2006. CLEF 2006: Ad Hoc Track Overview. En *Working Notes of the CLEF 2006 Workshop*, págs. 21–34.

Graña, J., M.A. Alonso, y M. Vilares. 2002. A common solution for tokenization and part-of-speech tagging: One-pass Viterbi algorithm vs. iterative approaches. *LNCS*, 2448:3–10.

Graña, J., F.M. Barcala, y J. Vilares. 2002. Formal methods of tokenization for part-of-speech tagging. *LNCS*, 2276:240–249.

Guo, J., G. Xu, H. Li, y X. Cheng. 2008. A unified and discriminative model for query refinement. En *Proc. of ACM SIGIR'08*, págs. 379–386. ACM.

Jansen, B.J., A. Spink, y T. Saracevic. 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management*, 36(2):207–227.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. En *Proc. of MT Summit X*, págs. 79–86. Corpus disponible en `http://www.statmt.org/europarl/`.

Koehn, P., F.J. Och, y D. Marcu. 2003. Statistical phrase-based translation. En *Proc. of NAACL'03*, págs. 48–54. ACL.

Kukich, K. 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys (CSUR)*, 24(4):377–439.

Levenshtein, V.I. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics-Doklandy*, 6:707–710.

Manning, C.D., P. Raghavan, y H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

McNamee, P. y J. Mayfield. 2004a. Character N-gram Tokenization for European Language Text Retrieval. *Information Retrieval*, 7(1-2):73–97.

McNamee, P. y J. Mayfield. 2004b. JHU/APL experiments in tokenization and non-word translation. *LNCS*, 3237:85–97.

Nie, J.-Y. 2010. *Cross-Language Information Retrieval*, vol. 8 de *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

Och, F.J. y H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51. Herramienta disponible en `http://code.google.com/p/giza-pp/`.

Otero, J., J. Graña, y M. Vilares. 2007. Contextual Spelling Correction. *LNCS*, 4739:290–296.

Ounis, I., C. Lioma, C. Macdonald, y V. Plachouras. 2007. Research directions in Terrier: a search engine for advanced retrieval on the web. *Novática/UPGRADE Special Issue on Web Information Access*, 8(1):49–56. Toolkit disponible en `http://www.terrier.org`.

Rehm, G. y H. Uszkoreit, eds. 2011. META-NET White Paper Series. Springer. Disponibles en `http://www.meta-net.eu/whitepapers`.

Robertson, A.M. y P. Willett. 1998. Applications of n-grams in textual information systems. *Journal of Documentation*, 54(1):48–69.

Savary, A. 2002. Typographical nearest-neighbor search in a finite-state lexicon and its application to spelling correction. *LNCS*, 2494:251–260.

Vilares, J., M.P. Oakes, y M. Vilares. 2007. A Knowledge-Light Approach to Query Translation in Cross-Language Information Retrieval. En *Proc. of RANLP 2007*, págs. 624–630.

Vilares, M., J. Otero, y J. Graña. 2004. On asymptotic finite-state error repair. *LNCS*, 3246:271–272.

Vilares, J., M. Vilares, y J. Otero. 2011. Managing Misspelled Queries in IR Applications. *Information Processing & Management*, 47(2):263–286.

Véronis, J. 1999. Multext-Corpora. An annotated corpus for five European languages. CD-ROM. Distributed by ELRA/ELDA.

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 33-40 recibido 30-04-2013 revisado 18-06-2013 aceptado 21-06-2013

Información Lingüística en Recuperación de Imágenes Multilingüe

Linguistic Information in Multilingual Image Retrieval

David Hernández-Aranda
NLP&IR group at UNED
daherar@lsi.uned.es

Víctor Fresno Fernández
NLP&IR group at UNED
vfresno@lsi.uned.es

Resumen: En este trabajo se evalúan diferentes modelos de indexación, así como la aplicación de técnicas de Procesamiento del Lenguaje Natural al texto asociado a las imágenes en un problema de Recuperación de Imágenes Multilingüe. Los resultados muestran que traducir el texto asociado y usar Entidades Nombradas, junto con sus categorías, permite mejorar el proceso de recuperación, mientras que la mejora obtenida con el uso de sintagmas nominales no compensa el coste computacional que conlleva.
Palabras clave: Recuperación de Imágenes, TBIR, Recuperación de Información Multilingüe, Procesamiento de Lenguaje Natural, Entidades Nombradas, Sintagmas Nominales.

Abstract: In this paper we evaluate several indexing models and the application of different Natural Language Processing techniques to textual descriptions associated with images in an Image Retrieval task. The results show that the use of a translation-based model and to take into account the Named Entities with their categories, improves the retrieval process. However, the improvement obtained with the use of noun phrases is not worth the computational cost involved.
Keywords: Image Retrieval, TBIR, Multilingual Information Retrieval, Natural Language Processing, Named Entities, Noun Phrases.

1 Introducción

Las principales aproximaciones utilizadas en la Recuperación de Imágenes Multilingüe se centran en el análisis del texto asociado a las imágenes (*Text-Based Image Retrieval, TBIR*) o de las características visuales de las mismas (*Content-Based Image Retrieval, CBIR*), y actualmente se están haciendo grandes esfuerzos en combinar ambos enfoques.

Las aproximaciones TBIR suelen aplican técnicas de IR ad-hoc sobre textos planos. En pocos casos se trata como un problema de IR estructurada, considerando de diferente modo distintos tipos de información textual. Si bien pueden encontrarse aproximaciones a la IR Multilingüe documental, hasta donde se ha revisado la literatura no se ha encontrado ningún estudio detallado sobre cómo tratar el texto asociado a las imágenes. En este trabajo se estudia qué y cómo realizar la indexación de esta información asociada, y sobre qué campos y cómo realizar el proceso de recuperación.

Los textos asociados a imágenes suelen ser textos cortos y de carácter descriptivo. De este modo, la presencia de entidades nombradas (*Named Entities*, NEs) cobra un papel destacado a la hora de describir el contenido de una imagen. Por otro lado, la aplicación de otras técnicas lingüísticas también puede ayudar a obtener buenos resultados. Entre estas técnicas se van a estudiar: el reconocimiento de NEs, junto con sus categorías, y el uso de sintagmas nominales. Aunque estos últimos necesitan de un mayor contexto para su detección, ya que en ocasiones se requiere de una fase de análisis sintáctico, se considera que pueden representar una unidad interesante de información a la hora de describir el contenido de una imagen.

Este trabajo se centra, por tanto, en tratar de explotar al máximo esa "pequeña cantidad de contenido multilingüe" de la que se dispone en las colecciones de imágenes anotadas y, para ello, se estudia la aplicación de técnicas de IR y Procesamiento de Lenguaje Natural con la idea de que cuanto más se mejoren las aproximaciones TBIR, mejores resultados se obtendrán después combinándolas con CBIR.

ISSN 1135-5948

2 Estado del Arte

En los sistemas TBIR, las imágenes se recuperan a partir de las anotaciones de texto (o metadatos) asociados con las imágenes (Zhang, Jiang y Zhang, 2009). Estas anotaciones pueden ser el texto que rodea a la imagen, el nombre de archivo, el hipervínculo o cualquier otro texto asociado con la imagen (Su et al., 2009). Los motores de búsqueda de imágenes de Google y Yahoo son sistemas que utilizan este enfoque.

Los metadatos de las imágenes se pueden dividir en dos partes (Styrman, 2005). Una se refiere a información sobre quién ha creado la imagen, las herramientas utilizadas en su creación, el estilo artístico, etc. La otra describe las propiedades implícitas que pueden ser entendidas por la percepción de la imagen en sí.

Por tanto, el primer paso es contar con anotaciones y, para ello, se suelen aplicar dos enfoques. En el primero se crean anotaciones manualmente; el segundo consiste en anotar automáticamente las imágenes a partir de un conjunto predefinido de categorías (Chatzilari et al., 2011). Para ello, utilizando aprendizaje automático se extraen relaciones entre las características de la imagen y palabras en imágenes anotadas (Zhang, Chai y Jin, 2005).

Los métodos automáticos no suelen ser fiables, ya que la anotación automática tiene funcionalidad limitada, y sólo unos pocos objetos pueden detectarse con seguridad a partir de imágenes generales (Uchihashi y Kanade, 2005). Por otro lado, las anotaciones manuales suelen ser costosas e incompletas, debido a la subjetividad, ambigüedad e imprecisión provocada por los contenidos semánticos de las imágenes. Este problema recibe el nombre de *semantic gap*, la diferencia que se produce entre dos (o más) descripciones de un objeto en diferentes representaciones (Nguyen, 2010). Esta diferencia suele venir dada por fenómenos del lenguaje natural tales como la sinonimia o la polisemia (Saenko y Darrell, 2008). Además, dependiendo del contexto, dos usuarios pueden anotar de manera diferente una misma imagen, p.ej., para una imagen de un coche, un usuario puede anotar "coche" y otro usuario puede anotar "Seat León de color rojo". Esto puede dar lugar a incoherencias e imprecisiones en la anotación (Pavlidis, 2008).

Otro hecho que hace que las anotaciones sean incompletas es que suelen ser cortas, por lo que es difícil representar completamente el contenido de la imagen. Las anotaciones no están siempre disponibles o no describen características visuales. Por otro lado, los conocimientos previos o influencias culturales pueden dar lugar a diferentes interpretaciones, por lo que también esto influye directamente en la calidad de las anotaciones. Además, el enfoque TBIR no es ajeno al multilingüismo, ya que las anotaciones pueden encontrarse en diferentes idiomas; por lo tanto, el éxito de la recuperación va a estar ligado a si un usuario conoce o no el idioma y no tanto por el contenido de la imagen en sí mismo.

Una vez extraídas las anotaciones asociadas a las imágenes, un sistema TBIR indexa esta información convirtiéndose en un problema de IR textual, por lo que solo se recuperarán imágenes que hayan sido anotadas con alguno de los términos usados en la consulta (Pérez-Iglesias, 2008). Sin embargo, estas técnicas no han sido muy aplicadas a los textos asociados a las imágenes considerando sus características. De hecho, tomando como referencia los sistemas TBIR presentados en el foro de evaluación ImageCLEF, desde 2003 a 2012, se puede observar que no se ha ido más allá de la detección de NEs. Por ejemplo, en el trabajo presentado por la UNED (Granados et al., 2011) se generan dos índices, uno solo con metadatos y otro solo con NEs, de manera que luego fusionan las dos listas de resultados obtenidas en la recuperación.

3 Experimentación

En esta sección se describen los experimentos así como el marco de evaluación en el que se llevan a cabo.

3.1 Colección de evaluación

La colección de evaluación utilizada en este trabajo es la empleada en la edición de ImageCLEF 2010 "Wikipedia Retrieval" (Popescu, Tsikrika y Kludas, 2010) para la recuperación de imágenes.

Esta colección está formada por 237.434 imágenes extraídas de la Wikipedia y sus correspondientes anotaciones cortas textuales realizadas por usuarios. Las anotaciones son multilingües (EN, DE y FR), y su distribución es la siguiente: 10% de imágenes con anotaciones en los 3 idiomas, 24% en 2 idiomas (11% EN+DE, 9% EN+FR, 4% FR+DE), 62% en un solo idioma (30% EN, 20% DE, 12% FR) y un 4% de imágenes en otros idiomas.

Las anotaciones de cada imagen se encuentran en los elementos <description>, <comment> y <caption>, identificados para cada idioma con el atributo "xml:lang" del elemento <text> del xml asociado. Además, existe un elemento común a todos los idiomas (<comment>) que contiene anotaciones entremezcladas en los tres idiomas, y sin formato fijo, lo que dificulta su procesado.

El método de evaluación se basa en juicios de relevancia. Se proporcionan 70 consultas en las tres lenguas, aportándose la traducción exacta de los términos de la consulta.

3.2 Modelos propuestos

Se evalúan tres modelos de indexación de información multilingüe. El primer modelo está formado por tres índices independientes que contendrán la información de cada idioma por separado. Un segundo modelo expande los índices anteriores traduciendo la información de las tres lenguas entre sí. El tercer modelo consiste en la creación de un único índice con la información conjunta en los tres idiomas.

Todos los índices estarán compuestos por los siguientes campos:

- **id**. Identificador del documento contenido en el atributo *id* del elemento <image>.
- **content**. Texto formado por la concatenación de los textos contenidos en los elementos <name>, <description>, <comment>, <caption> y el elemento común <comment> para cada idioma.
- **nes**. NEs detectadas en el campo *content*.
- **nes_person**. NEs de tipo PERSON detectadas en el campo *content*.
- **nes_organization**. NEs de tipo ORGANIZATION detectadas en *content*.
- **nes_location**. NEs de tipo LOCATION detectadas en el campo *content*.
- **nes_misc**. NEs de tipo MISC detectadas en el campo *content*.
- **sintagmas**. Sintagmas nominales detectados en el campo *content*.

La detección de NEs y sus categorías, tanto para el inglés como el alemán, se ha realizado con la herramienta *Stanford NER*[1], y para el francés con la herramienta *Stilus NER*[2]. Las categorías de las NEs consideradas son PERSON, ORGANIZATION, LOCATION y MISC, y para la detección de los sintagmas nominales, tanto en inglés como en francés, se ha utilizado la herramienta *Stilus Core*[3]; para el alemán, la herramienta *ParZu*[4].

En los tres modelos, el preprocesamiento del texto se ha llevado a cabo por los analizadores de *SnowBall*[5] implementados para cada idioma en *Lucene*[6], consistentes en la transformación del texto en minúsculas, eliminación de caracteres especiales, signos de puntuación y stemming. Sin embargo, la lista de stopwords ha sido sustituida por listas más completas proporcionadas por la Universidad de Neuchatel[7] (UniNE). Las consultas tendrán el mismo tratamiento que el de los documentos y para el caso de consultas multi-término, el operador utilizado para la búsqueda es OR.

3.2.1 Índices independientes

Este primer modelo consiste en crear tres índices independientes por idioma, que denominaremos 'EN', 'FR', 'DE' para el inglés, francés y alemán respectivamente. De esta manera, cada índice almacena únicamente la información extraída en su idioma. Se proponen dos modos de recuperación. Por un lado, se realizan búsquedas monolingües, consistentes en realizar una consulta por idioma, denominada 'en', 'fr' y 'de' para el inglés, francés y alemán respectivamente. Así, con una consulta en inglés se accederá sólo al índice en inglés y así sucesivamente para el resto de idiomas, obteniéndose tres listas de rankings. Estas consultas se denotan por las duplas EN-en, FR-fr y DE-de, siendo el primer término el idioma del índice y el segundo el de la consulta. Se propone una fusión de listas de documentos mediante el algoritmo de fusión *raw scoring* (Kwok, Grunfeld y Lewis, 1995).

3.2.2 Índices independientes expandidos

En este modelo se crean tres índices independientes por idioma, pero ahora la información extraída se expande con la información traducida de otros idiomas en el caso de no existir en el idioma original. Por ejemplo, si en el campo de un documento no existe información para el inglés, pero sí la hay en alemán, traduciríamos el texto en alemán al

[1] http://nlp.stanford.edu/software/CRF-NER.shtml
[2] http://www.daedalus.es/productos/stilus/stilus-ner/
[3] http://daedalus.es/productos/stilus/stilus-core/
[4] http://github.com/rsennrich/parzu
[5] http://snowball.tartarus.org/
[6] http://lucene.apache.org/java/docs/index.html
[7] http://members.unine.ch/jacques.savoy/clef/

inglés. Si no hubiera tampoco contenido en alemán, traduciríamos la del francés si la hubiera. Este proceso crea tres índices expandidos (ENexp, FRexp y DEexp). Teniendo en cuenta que en la colección la gran mayoría de la información está en inglés, seguida del alemán y francés, se establece este orden de precedencia para la búsqueda de información para traducir, y una vez encontrada en un idioma, no se seguiría buscando. La traducción se ha realizado por medio del API proporcionado por *Google Translate*[8]. Al igual que en el modelo anterior, se realizará la recuperación mediante búsquedas monolingües representadas por las duplas ENexp-en, FRexp-fr y DEexp-de y habrá que fusionar las listas de rankings obtenidas.

3.2.3 Índice único

Este modelo consiste en la creación de un único índice, que llamamos ALL, con toda la información extraída de los tres idiomas. Dada la dificultad para identificar el idioma de los textos en los elementos <name> y <comment>, donde se mezclan contenidos en diferentes lenguas, se asume que los contenidos están en inglés. La recuperación se realiza mediante búsquedas monolingües al índice conjunto, representadas por las duplas ALL-en, ALL-fr y ALL-de, y otra búsqueda compuesta por la concatenación de la consulta en los tres idiomas, representada por la dupla ALL-all, donde all es la unión de las consultas en 'en', 'fr' y 'de'. Con este modelo se obtendrá, por tanto, una única lista de ranking.

3.3 Funciones de ranking

Se considera el empleo de la función de ranking BM25, y para poder estudiar el impacto de la información lingüística contenida en los diferentes campos de los índices, se utiliza su extensión a documentos estructurados, BM25F (Robertson y Walker, 1994).

Como valores de *b* y k se establecen valores estándar, mientras que los valores de empuje para los diferentes campos en el caso de BM25F se fijan tras probar diferentes combinaciones, algo común cuando se usan estas funciones tan parametrizables (Hernández-Aranda, 2013).

[8] http://code.google.com/intl/es-ES/apis/language/translate/overview.html

3.4 Medidas de evaluación

Las medidas de evaluación que se van a emplear esta experimentación son: MAP (*Mean Average Precision*), que calcula una media de la precisión hallada a distintos niveles de cobertura; y P@k (*Precision at k*) que indica la precisión obtenida en el conjunto de las primeras k imágenes recuperados.

4 Análisis de Resultados

En esta sección se presentan y analizan los resultados de la experimentación llevada a cabo. Este análisis se estructura en torno a los modelos de indexación considerados y, para cada uno de ellos, los resultados se encuentran agrupados en los siguientes bloques:

- **Baseline**. La consulta se realiza sólo sobre el campo *content* y al ser un único campo, la función de ranking utilizada es BM25.
- **+ NEs**. La consulta se realiza también sobre el campo en el que se almacenan las NEs y se emplea la función de ranking BM25F.
- **+ Categorías**. La consulta se realiza sobre los campos *content* y los correspondientes a las diferentes categorías de las NEs.

A continuación, y para cada modelo de indexación, se presentan los resultados obtenidos cuando se añaden, a los experimentos anteriores, los sintagmas nominales detectados. En estos experimentos los resultados se agrupan en los siguientes bloques:

- **+ Sintagmas**. En este caso, además de lanzar la consulta sobre el campo *content* (*Baseline*), se hace también sobre el campo correspondiente a los sintagmas nominales.
- **+ Sintagmas + NEs**. Además de considerar los campos empleados en la aproximación anterior (*+Sintagmas*), la consulta se realiza ahora también sobre el campo de las NEs.
- **+ Sintagmas + Categorías**. Además de considerar los campos usados en *+Sintagmas*, la consulta se lanza también sobre los campos correspondientes a las diferentes categorías de las NEs.

Para cada uno de los experimentos se aplica el test de significancia estadística pareado de Wilcoxon, de modo que nos permita comprobar si las diferencias entre los valores obtenidos con un método u otro son estadísticamente significativas. El test se realiza tanto para valores de MAP ("p (MAP)" en las tablas), como de P@10 ("p (P@10)" en las tablas).

Antes de presentar los resultados, y para contextualizarlos de manera adecuada, se resumen los resultados obtenidos en el foro ImageCLEF 2010 y se comparan con los obtenidos en este trabajo. En ImageCLEF 2010 se presentaron 13 grupos de investigación que enviaron 113 experimentos, de los cuales 48 suponían un enfoque TBIR. Los mejores resultados obtenidos por cada grupo se correspondieron con experimentos realizados sólo en inglés o considerando todos los idiomas conjuntamente, tanto en las descripciones como en las consultas. Este hecho indica que la información textual disponible en inglés es más rica que la disponible para el resto de idiomas.

El mejor resultado para los enfoques TBIR, en el que nuestro trabajo se enmarca, lo obtuvo el grupo *XRCE* de Xerox (Clinchant et al., 2010), representando el texto por medio de modelos del lenguaje y con un modelo de información basado en la ley de potencias, combinándolo con una aproximación basada en retroalimentación. Los valores de MAP y P@10 que obtuvieron fueron 0.2361 y 0.4871 respectivamente, mientras que la media obtenida por los experimentos presentados en ImageCLEF 2010, calculada eliminando el 10% de los peores resultados, tiene como valor de MAP 0,1602 con una desviación estándar de 0,0505. Del mismo modo, en el caso de la P@10 el valor medio de los sistemas fue 0,4095, con una desviación estándar de 0,0713.

A continuación se muestran los resultados obtenidos con nuestros experimentos para cada uno de los tres modelos propuestos.

Índices independientes. La Tabla 1 muestra los resultados obtenidos para este modelo. En primer lugar, y como se concluyó en ImageCLEF 2010, se observa que para las búsquedas monolingües los mejores resultados se obtienen para el inglés. Una posible razón de este comportamiento es el hecho de que la información en este idioma es más extensa.

Observando estos resultados se puede concluir, de forma general, que el hecho de considerar NEs y sus categorías mejoran el baseline, excepto para el caso del alemán en el que puede no haber muchas NEs en su texto, obteniéndose mejores resultados para las NEs en conjunto en lugar de ponderar de diferente manera a cada una de las categorías.

En cuanto a los resultados de la P@10, en ningún caso se obtienen mejoras significativas, lo que quiere decir que las posibles diferencias que hubiera al considerar o no las NEs y sus categorías no modifican el conjunto de los 10 documentos más relevantes, aunque en algún caso el orden dentro de estos 10 primeros documentos pudiera ser diferente. Esto es lo que explica que exista algún caso donde se encuentren mejoras en MAP y no de P@10.

Baseline	**Map**	**P@10**		
BM25 (EN-en)	0.2244	0.5071		
BM25 (FR-fr)	0.1029	0.3500		
BM25 (DE-de)	0.0959	0.3186		
BM25 (con fusión)	0.2263	0.5114		
+ NEs	**Map**	**P@10**	**p(Map)**	**p(P@10)**
BM25F (EN-en)	**0.2314**	0.5014	0.0394	0.4400
BM25F (FR-fr)	0.1052	0.3457	0.0003	0.3735
BM25F (DE-de)	0.1002	0.3186	0.4614	0.4761
BM25F (con fusión)	**0.2359**	0.5071	0.0007	0.4353
+ Categorías	**Map**	**P@10**	**p(Map)**	**p(P@10)**
BM25F (EN-en)	0.2307	0.5000	0.0072	0.4259
BM25F (FR-fr)	0.1052	0.3600	0,0000	0.2758
BM25F (DE-de)	0.0985	0.3186	0,5954	0.6082
BM25F (con fusión)	0.2328	0.5086	0,0002	0.6026
+ Sintagmas	**Map**	**P@10**	**p(Map)**	**p(P@10)**
BM25F (EN-en)	0.2264	0.4971	0.1259	0.6639
BM25F (FR-fr)	0.1081	0.3643	0.0004	0.2970
BM25F (DE-de)	0.0918	0.2814	0.0051	0.0046
BM25F (con fusión)	0.2257	0.4786	0.9714	0.0183
+ Sintagmas + NEs	**Map**	**P@10**	**p(Map)**	**P(P@10)**
BM25F (EN-en)	0.2313	0.5000	0.0327	0.3994
BM25F (FR-fr)	0.1074	0.3557	0.0001	0.8717
BM25F (DE-de)	0.0954	0.3029	0.3950	0.3568
BM25F (con fusión)	0.2324	0.5000	0.0112	0.4242
+ Sintagmas + Categorías	**Map**	**P@10**	**p(Map)**	**p(P@10)**
BM25F (EN-en)	0.2306	0.5000	0.0073	0.4844
BM25F (FR-fr)	0.1053	0.3600	0.0000	0.2155
BM25F (DE-de)	0.0949	0.3143	0.8569	0.1670
BM25F (con fusión)	0.2298	0.4971	0.0010	0.0686

Tabla 1 - Resultados sobre el modelo de *Índices independientes*.

Un aspecto a resaltar es el hecho de que sí se encuentra mejora en el valor de MAP en el caso de la fusión de listas. Esta mejora puede deberse a que a la hora de fusionar se están introduciendo en la lista de ranking del inglés (por ser el idioma con el que se obtiene mejor resultado), documentos del resto de idiomas que en otro caso no se hubieran recuperado por no estar escritos en inglés. De hecho, el mejor resultado con este modelo, con un valor de MAP de 0.2359, se obtiene cuando se realiza la fusión de listas con la función BM25F y teniendo en cuenta las NEs. De nuevo no se observa mejora en los valores de P@10, lo que indica que la fusión está reordenando los 1000 primeros resultados devueltos por el sistema, pero sin variar el número de documentos relevantes en las primeras 10 posiciones.

Cuando se tienen en cuenta los sintagmas no se encuentran diferencias significativas, salvo en el único caso del francés. Sin embargo,

cuando se combinan los sintagmas con las NEs y categorías se consigue superar siempre el baseline excepto para el alemán. Incluso con la fusión de listas se consiguen mejoras significativas. Sin embargo, estos resultados no superan los resultados obtenidos cuando se tienen en cuenta solamente las NEs y categorías. Por tanto, se puede concluir que esta mejora se debe al uso de las NEs y sus categorías, y no al uso de los sintagmas.

En resumen, para este primer modelo el mejor resultado monolingüe ha sido obtenido por el inglés teniendo en cuenta las NEs, con un valor de MAP de 0.2314. Sin embargo, el mejor resultado total se ha obtenido con la fusión de listas de documentos en los tres idiomas teniendo en cuenta las NEs, con un valor de MAP de 0.2359. Por tanto, se observa un buen comportamiento con el uso de las NEs y sus categorías, mientas que los sintagmas no parecen ser una buena opción. Sin embargo, a la vista de los valores de P@10, la mejora solo se da en la ordenación de los resultados, pero no se consigue recuperar más imágenes relevantes en los 10 primeros resultados.

Índices independientes expandidos. La Tabla 2 muestra los resultados obtenidos para este modelo. Al igual que en el modelo anterior, los mejores resultados se obtienen para el inglés. Sin embargo, se observa una importante mejora en el caso de francés y alemán gracias a la expansión de información que llevan consigo las traducciones. Esto se debe a que las traducciones permiten recuperar documentos escritos en una o dos lenguas que de otra manera sólo se hubiera podido recuperar con la consulta en su propia lengua.

Se observa que los mejores resultados se obtienen, tanto con búsquedas monolingües como con fusión de resultados teniendo en cuenta a las NEs (a excepción del alemán) y Categorías. Además, con las categorías, para la búsqueda monolingüe en inglés se obtiene el mejor resultado en el mismo experimento con NEs. De hecho, con 0.2413 es el mejor resultado en búsquedas monolingües. Por tanto, dar un peso específico a las NEs y, sobretodo, dar más importancia a las NE de tipo Persona y Organización que a las de tipo Location y Miscelánea, parece dar buenos resultados.

El mejor resultado de este modelo, con un valor de MAP de 0.2434, se vuelve a obtener cuando se realiza la fusión de listas teniendo en cuenta las NEs. De nuevo, con respecto a los valores de P@10, en ningún caso se obtienen mejoras significativas, lo que quiere decir que no se modifica el conjunto de los 10 documentos más relevantes.

Baseline	**Map**	**P@10**		
BM25 (ENexp-en)	0.2323	0.4871		
BM25 (FRexp-fr)	0.1766	0.4129		
BM25 (DEexp-de)	0.1432	0.3557		
BM25 (con fusión)	0.2302	0.5043		
+ NEs	**Map**	**P@10**	**p(Map)**	**p(P@10)**
BM25F (ENexp-en)	0.2408	0.5057	0.0125	0.2309
BM25F (FRexp-fr)	0.1858	0.4157	0.0130	0.4966
BM25F (DEexp-de)	0.1498	0.3614	0.1286	0.4352
BM25F (con fusión)	**0.2434**	0.5114	0.0007	0.6823
+ Categorías	**Map**	**P@10**	**p(Map)**	**p(P@10)**
BM25F (ENexp-en)	0.2413	0.4971	0.0001	0.3310
BM25F (FRexp-fr)	0.1855	0.4200	0.0000	0.1779
BM25F (DEexp-de)	0.1490	0.3657	0.0006	0.9022
BM25F (con fusión)	0.2404	0.5057	0.0000	0.8367
+ Sintagmas	**Map**	**P@10**	**p(Map)**	**p(P@10)**
BM25F (ENexp-en)	0.2362	0.4714	0.4084	0,3093
BM25F (FRexp-fr)	0.1839	0.4200	0.0011	0.4928
BM25F (DEexp-de)	0.1436	0.3443	0.5791	0.1297
BM25F (con fusión)	0.2383	0.5029	0.0172	0.8833
+ Sintagmas + NEs	**Map**	**P@10**	**p(Map)**	**p(P@10)**
BM25F (ENexp-en)	0.2409	0.4986	0.0165	0.5743
BM25F (FRexp-fr)	0.1860	0.4229	0.0045	0.2195
BM25F (DEexp-de)	0.1494	0.3629	0.1282	0.6854
BM25F (con fusión)	0.2427	0.5086	0.0023	0.7186
+ Sintagmas + Categorías	**Map**	**P@10**	**p(Map)**	**p(P@10)**
BM25F (ENexp-en)	**0.2414**	0.4914	0.0001	0.8353
BM25F (FRexp-fr)	0.1858	0.4214	0.0000	0.1516
BM25F (DEexp-de)	0.1485	0.3657	0.0024	1.0000
BM25F (con fusión)	0.2405	0.5086	0.0000	1.0000

Tabla 2 - Resultados sobre el modelo de *Índices independientes expandidos*.

Con este modelo los resultados superan a los mejores obtenidos en ImageCLEF 2010 (Xrce) en términos de MAP y P@10, tanto con consulta monolingüe en inglés como con fusión de listas considerando las NEs y sus categorías.

En este modelo, cuando se tienen en cuenta los sintagmas, tanto solos como junto a las NEs y Categorías, en la mayoría de los casos se consiguen superar los resultados del baseline, tanto en las búsquedas monolingües como en la fusión de listas. Esta mejora puede deberse, con respecto al modelo anterior, a que con las traducciones se ha ampliado el contexto de las descripciones de los documentos y, por tanto, se detectan un mayor número de sintagmas. Con respecto al modelo de índices independientes, en el que solo en un caso con sintagmas se conseguía superar el baseline, parece que con la traducción se ha superado la falta de información que supuestamente hacía que la utilización de sintagmas no fuera útil.

Para las búsquedas monolingües, teniendo en cuenta Sintagmas y Categorías, se consigue superar al mejor resultado previo. Sin embargo

esta mejora es mínima, siendo el valor de MAP de 0.2414 frente al 0.2413 que se consigue con el mismo experimento pero sin tener en cuenta los sintagmas. Para esta mejora tan mínima no merece la pena el coste computacional que supone la detección de sintagmas nominales. En cuanto a la fusión de listas se obtiene el mejor resultado considerando NEs, pero sin superar el mejor resultado obtenido dentro de este modelo.

En resumen, el mejor resultado monolingüe de este modelo se ha obtenido con inglés teniendo en cuenta los sintagmas y categorías. Por otro lado, el mejor resultado global se ha obtenido con la fusión de listas de documentos en los tres idiomas teniendo en cuenta las NEs, con un valor de MAP de 0.2434.

Índice Único. La Tabla 3 muestra los resultados obtenidos para este modelo. El comportamiento observado en los modelos anteriores se mantiene para el inglés. Para el inglés y alemán en el baseline, los resultados superan al baseline del modelo de Índices Independientes, lo que indica que se están recuperando más documentos relevantes al estar toda la información almacenada conjuntamente, por lo que es posible que existan términos en la consulta y en las descripciones de las imágenes que compartan su misma forma canónica o su raíz en inglés y alemán (puede ser que se trate de cognados entre ambas lenguas que compartan raíz). Esto no ocurre para el francés.

Se observa que los resultados que presentan mejoras significativas se obtienen considerando las NEs (a excepción del alemán) y sus Categorías, tanto con búsquedas monolingües como con fusión de resultados. Además con las categorías, y para el inglés, se obtienen mejoras al mismo experimento realizado con NEs. Por tanto, dar un peso específico a las NEs, y sobretodo, dar más importancia a las entidades de tipo Person y Organization que a las de tipo Location y Misc, parece dar buenos resultados.

Es importante destacar que este modelo mejora cuando la consulta se realiza con la concatenación de las *queries* en los tres idiomas (consulta *all*) y se tienen en cuentas las NEs y sus categorías. Se está logrando aumentar la cobertura (al recuperar documentos en todos los idiomas) sin penalizar valores de precisión. De hecho, los valores de P@10 mejoran con este tipo de consultas, de modo que se están consiguiendo recuperar más documentos relevantes en los 10 primeros resultados, al intercalar documentos relevantes de los tres idiomas en esas 10 primeras posiciones.

Baseline	**Map**	**P@10**		
BM25 (ALL-en)	0.2285	0.5000		
BM25 (ALL-fr)	0.1019	0.3043		
BM25 (ALL-de)	0.1043	0.3000		
BM25 (ALL-all)	0.2300	0.5014		
+ NEs	**Map**	**P@10**	**P(Map)**	**p(P@10)**
BM25F (ALL-en)	0.2332	0.5057	0.0016	0.7485
BM25F (ALL-fr)	0.1060	0.3100	0.0000	0.3791
BM25F (ALL-de)	0.1025	0.2943	0.5749	1.0000
BM25F (ALL-all)	0.2401	0.5243	0.0298	0.0132
+ Categorías	**Map**	**P@10**	**p(Map)**	**p(P@10)**
BM25F (ALL-en)	**0.2336**	0.5057	0.0002	0.7112
BM25F (ALL-fr)	0.1046	0.3014	0.0000	0.7750
BM25F (ALL-de)	0.1022	0.2943	0.2785	0.9364
BM25F (ALL-all)	**0.2403**	0.5107	0.0103	0.0191
+ Sintagmas	**Map**	**P@10**	**p(Map)**	**p(P@10)**
BM25F (ALL-en)	0.2285	0.4686	0.9829	0.0098
BM25F (ALL-fr)	0.1019	0.3143	0.0935	0.1659
BM25F (ALL-de)	0.1022	0.2900	0.9492	0.6174
BM25F (ALL-all)	0.2300	0.5086	0.9117	0.4113
+ Sintagmas +NEs	**Map**	**P@10**	**p(Map)**	**p(P@10)**
BM25F (ALL-en)	0.2321	0.4900	0.0597	0.2046
BM25F (ALL-fr)	0.1034	0.3057	0.0310	0.4101
BM25F (ALL-de)	0.1024	0.2857	0.8130	0.2386
BM25F (ALL-all)	0.2400	0.5071	0.1617	0.6306
+ Sintagmas + Categorías	**Map**	**P@10**	**p(Map)**	**p(P@10)**
BM25F (ALL-en)	0.2333	0.5100	0.0030	0.3919
BM25F (ALL-fr)	0.1034	0.3014	0.0385	0.7346
BM25F (ALL-de)	0.1014	0.2914	0.5725	0.7495
BM25F (ALL-all)	0.2402	0.5086	0.8416	0.5900

Tabla 3 - Resultados sobre el modelo de *Índice Único*

Se puede observar también que, tanto para las búsquedas monolingües como cuando se concatenan las consultas, al tener en cuenta los sintagmas no se consigue superar el baseline, ni cuando consideramos las NEs, ni las categorías. Sin embargo, cuando se tienen en cuenta los sintagmas y las categorías para el inglés, se consigue superar el baseline, aunque no se supera el resultado obtenido cuando se tienen en cuenta solamente las categorías. Por tanto, la mejora vuelve a deberse al uso de las categorías y no de los sintagmas.

En resumen, el mejor resultado monolingüe se ha obtenido con el inglés teniendo en cuenta las categorías. Sin embargo, el mejor resultado se ha obtenido cuando la recuperación se realiza con la concatenación de las consultas en los tres idiomas y teniendo en cuenta las categorías, con un valor de MAP de 0.2403.

5 Conclusiones y Futuros trabajos

En este artículo se ha presentado un análisis y evaluación del uso de diferentes modelos de indexación/recuperación dentro de una tarea de recuperación multilingüe de imágenes basada

en el texto asociado a las imágenes. Se ha evaluado el uso de las NEs, sus categorías y los sintagmas nominales detectados.

Los resultados obtenidos muestran que, a pesar del elevado coste computacional, el mejor enfoque es el de traducir los documentos a todas las lenguas consideradas. Sin embargo, si no es posible traducir toda la colección, la creación de un índice único y la concatenación de consultas en diferentes idiomas ha resultado ser la mejor opción. Además, la fusión de listas de resultados es mejor solución que realizar búsquedas monolingües. El hecho de considerar las entidades nombradas y sus categorías ha resultado ser buena opción, pero sin embargo, el uso de sintagmas nominales quedaría descartado, ya que se obtiene poco beneficio para el coste computacional que requiere.

En cuanto a los futuros trabajos, sería interesante la aplicación de los enfoques presentados en otro tipo de colecciones multimedia; por ejemplo de vídeos o audios de los que se tuvieran transcripciones del habla o subtítulos. Por último, se estudiará el efecto de diferentes algoritmos de fusión de listas.

6 Agradecimientos

Este trabajo no sería posible sin la participación de los autores en los proyectos de investigación MA2VICMR (S2009/TIC-1542), HOLOPEDIA (TIN2010-21128-C02) y PROYECTO UNED (2012V/PUNED/0004).

Bibliografía

Chatzilari, .E, S. Nikolopoulos, Papadopoulos, C. Zigkolis y Y. Kompatsiaris. 2011. Semi-Supervised object recognition using flickr images. In *9th International Workshop on Content-Based Multimedia Indexing*, Madrid, Spain.

Clinchant, S., G. Csurka, J. Ah-Pine, G. Jacquet, F. Perronnin, J. Sánchez y K. Minoukadeh. 2010. XRCE's Participation in Wikipedia Retrieval, Medical Image Modality Classification and Ad-hoc Retrieval Tasks of ImageCLEF 2010. *CLEF*.

Granados, R., J. Benavent, X. Benavent, E. de Ves y A. García-Serrano. 2011. Multimodal information approaches for the Wikipedia collection at Image-CLEF 2011. In *CLEF 2011 Working Notes*.

Hernández-Aranda, D. 2013. Información Textual Multilingüe en Recuperación de Imágenes. *Tesis de Fin de Master en Lenguajes y Sistemas Informáticos*. UNED.

Kwok, K. L., L. Grunfeld y D. D. Lewis. 1995. TREC-3 Ad-Hoc, Routing Retrieval and Thresholding Experiments using PIRCS. In *Procs. of TREC3*, 47-56. NIST.

Nguyen, C. T. 2010. Bridging semantic gaps in information retrieval: Context-based approaches. ACM VLDB, 10.

Pavlidis, T. 2008. Limitations of cbir. In *ICPR*.

Pérez Iglesias, J. 2008. Función de ranking para documentos estructurados, basada en lógica borrosa. *DEA*. UNED.

Popescu, A., T. Tsikrika y J. Kludas. 2010. Overview of the Wikipedia Retrieval Task at ImageCLEF 2010. In *CLEF* (Notebook Papers/LABs/Workshops).

Robertson, S. E. y S. Walker. 1994. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proc. of the SIGIR '94*, W. Bruce Croft and C. J. van Rijsbergen (Eds.). Springer-Verlag. NY, USA, 232-241.

Saenko, K. y T. Darrell. 2008. Unsupervised Learning of Visual Sense Models for Polysemous Word. In *Proc. of the 22nd Annual Conference on Neural Information Processing Systems*. Vancouver, Canada, pp.1393-1400.

Styrman, A. 2005. Ontology-based image annotation and retrieval. Doctoral dissertation, Master thesis.

Su, J., B. Wang, H. Yeh y V. S. Tseng. 2009. Ontology-Based Semantic Web Image Retrieval by Utilizing Textual and Visual Annotations. In *Web Intelligence/IAT Workshops*, pp: 425-428.

Uchihashi, S. y T. Kanade. 2005. Content-Free Image Retrieval Based On Relations Exploited From User Feedbacks. IEEE.

Zhang, Ch., J. Y. Chai y R. Jin. 2005. User Term Feedback in Interactive Text-based Image Retrieval. SIGIR'05, August 15–19, Salvador, Brazil.

Zhang, H., M. Jiang y X. Zhang. 2009. Exploring image context for semantic understanding and retrieval. In *International Conference on Computational Intelligence and Software Engineering*, pp. 1 – 4.

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 41-48 recibido 30-04-2013 revisado 18-06-2013 aceptado 21-06-2013

Removing Noisy Mentions for Distant Supervision

Eliminando Menciones Ruidosas para la Supervisión a Distancia

Ander Intxaurrondo*, Mihai Surdeanu, Oier Lopez de Lacalle***, Eneko Agirre***
*University of the Basque Country. Donostia, 20018, Basque Country
**University of Arizona. Tucson, AZ 85721, USA
***University of Edinburgh. Edinburgh, EH8 9LE, UK
ander.intxaurrondo@ehu.es, msurdeanu@email.arizona.edu,
oier.lopezdelacalle@ehu.es, e.agirre@ehu.es

Resumen: Los métodos para Extracción de Información basados en la Supervisión a Distancia se basan en usar tuplas correctas para adquirir menciones de esas tuplas, y así entrenar un sistema tradicional de extracción de información supervisado. En este artículo analizamos las fuentes de ruido en las menciones, y exploramos métodos sencillos para filtrar menciones ruidosas. Los resultados demuestran que combinando el filtrado de tuplas por frecuencia, la información mutua y la eliminación de menciones lejos de los centroides de sus respectivas etiquetas mejora los resultados de dos modelos de extracción de información significativamente.
Palabras clave: Extracción de Información, Extracción de Relaciones, Supervisión a Distancia, Aprendizaje con Ruido

Abstract: Relation Extraction methods based on Distant Supervision rely on true tuples to retrieve noisy mentions, which are then used to train traditional supervised relation extraction methods. In this paper we analyze the sources of noise in the mentions, and explore simple methods to filter out noisy mentions. The results show that a combination of mention frequency cut-off, Pointwise Mutual Information and removal of mentions which are far from the feature centroids of relation labels is able to significantly improve the results of two relation extraction models.
Keywords: Information Extraction, Relation Extraction, Distant Supervision, Learning with Noise

1 Introduction

Distant Supervision (DS) is a semi-supervised alternative to traditional Relation Extraction (RE) that combines some of the advantages of different RE approaches. The intuition is that any sentence that contains a pair of entities that are recorded in a Knowledge Base (KB) such as DBpedia[1] or Freebase[2] to participate in a known relation (e.g., *born-in* or *film-director-of*) is likely to provide evidence for that relation. Using this approach, large training datasets of relation mentions can be automatically created by aligning entities that participate in known relations with sentences from large corpora where the entity pairs are mentioned. Such sentences are preprocessed to identify all named or numeric entities that are mentioned. Entities are identified using named entity recognizers, tagging them as persons, organizations, locations, dates, etc. If the KB specifies that a pair of entities appearing in the same sentence participates in a known relation, the corresponding textual context becomes a mention for the corresponding relation label. If the KB has no record of the two entities, the corresponding relation is marked as *unrelated* (i.e., a negative mention). Using this approach, a very large number of relation mentions can be gathered automatically, thus alleviating the sparse data problem plaguing supervised relation extraction systems, which ultimately causes overfitting and domain dependence.

In order to illustrate the method, let's consider some relations[3] and tuples from Free-

[1] http://dbpedia.org/About
[2] http://www.freebase.com/

[3] In order to improve readability, we will use intuitive tags instead of the actual Freebase relation names, i.e., *education* for */education/education/student*, *capital* for */lo-*

ISSN 1135-5948

base:

- <Albert Einstein, *education*, University of Zurich>
- <Austria, *capital*, Vienna>
- <Steven Spielberg, *director-of*, Saving Private Ryan>

Searching for the entity pairs in those tuples, we can retrieve sentences that express those relations:

- **Albert Einstein** was awarded a PhD by the **University of Zurich**.
- **Vienna**, the capital of **Austria**.
- Allison co-produced the Academy Award-winning **Saving Private Ryan**, directed by **Steven Spielberg**...

Although we show three sentences that do express the relations in the knowledge-base, distant supervision generates many noisy mentions that hurt the performance of the relation extraction system. We identified three different types of noise in the mentions gathered by distant supervision:

1. Sentences containing related entities, but which are tagged as 'unrelated' by DS. This happens because the KB we use, as all real-world Kbs, is incomplete, i.e., it does not contain all entities that participate in a given relation type.
2. Sentences containing unrelated entities, tagged as related. This happens when both participating entities that are marked as related in the KB appear in the same sentence, but the sentence does not support the relation.
3. Sentences containing a pair of related entities, but which are tagged as a mention of another, incorrect, relation. This type is the most common, and happens for entity tuples that have more than one relational label. These were previously called multi-label relations in the literature (Hoffmann et al., 2011).

Suppose that we have an incomplete KB according to whom the tuple <Brazil, Celso Amorim> is unrelated. In reality Celso is a minister of Brazil, and thus a mention of the *country-minister* relation. Mentions like *Celso Amorim , the Brazilian foreign minister , said the (...)* will be tagged by DS systems as unrelated at the training dataset, instead of appearing as *country-minister* as it should be. This is an example of Type 1.

Situations of Type 2 noise occur with tuples like <Jerrold Nadler, *born_in*, Brooklyn>. If the system extracts the following sentences from the corpora, *(...) Representative Jerrold Nadler, a Democrat who represents parts of Manhattan and Brooklyn, (...)* and *Nadler was born in Brooklyn, New York City.*, they both will be tagged as *born_in* and used later for training, although the entity tuple in the first sentence is not a positive mention of the relation under consideration.

Below we give an example of Type 3 noise. Consider the tuple <Rupert Murdock, News Corporation>. This is a multi-label relation with labels *founder* and *top-member*. Thus sentences in the training set such as *News Corporation was founded by Rupert Murdock* and *Rupert Murdock is the CEO of News Corporation* will be both considered as mentions for both *founder* and *top-member*, even though the first sentence is not a mention for the *top-member* relation and the second is not a mention for the *founder* relation.

We selected randomly 100 mentions respectively from single-label related mentions, multi-labeled related mentions and unrelated mentions which correspond to Freebase relations as gathered by (Riedel, Yao, and McCallum, 2010). We analyzed them, and estimated that around 11% of the unrelated mentions belong to Type 1, 28% of related single-labeled mentions belong to Type 2. Regarding multi-labeled mentions, 15% belong to Type 3 and 60% to Type 2, so only 25% are correct mentions. All in all, the dataset contains 91373 unrelated mentions, 2330 single-labeled and 26587 multi-labeled mentions, yielding an estimate of 29% correct instances for related mentions, and 74% correct instances overall.

Noisy mentions decrease the performance of distant supervision systems. However, because the underlying datasets are generally very large, detecting and removing noisy mentions manually becomes untenable. This paper explores several methods that automatically detect and remove noisy mentions generated through DS.

cation/country/capital, and *director-of* for */film/director/film*

2 Related Work

Distant Supervision was originally proposed by (Craven and Kumlien, 1999) for the biomedical domain, extracting binary relations between proteins, cells, diseases and more. Some years later, the approach was improved by (Mintz et al., 2009), making it available for different domains, such as *people, locations, organizations,...*, gaining popularity since then.

We can find many approaches that model the noise to help the classifier train on the respective datasets. (Riedel, Yao, and McCallum, 2010) model distant supervision for relation extraction as a multi-instance single-label problem, allowing multiple mentions for the same tuple, but it does not allow more than one label per object. (Hoffmann et al., 2011) and (Surdeanu et al., 2012) focus on multi-instance multi-label learning.

Distant supervision has also been the most relevant approach used to develop different relation extraction system at the *TAC-KBP Slot Filling* task[4] for the last years, organized by NIST. Nearly all the participants use distant supervision for their systems to extract relations for *people* and *organization* entities. The approach has improved slowly during the latest years, and working with noisy mentions to train the systems has been recognized as the most important hurdle for further improvements.

3 Distant Supervision for Relation Extraction

The methods proposed here for cleaning the textual evidence used to train a RE model are system independent. That is, they apply to any RE approach that follows the "traditional" distant supervision heuristic of aligning database tuples with text for training. As proof of concept, in this paper we use two variants of the *Mintz++* system proposed by (Surdeanu et al., 2012) and freely available at `http://nlp.stanford.edu/software/mimlre.shtml`. This algorithm is an extension of the original work of (Mintz et al., 2009) along the following lines:

- The *Mintz++* approach models each relation mention independently, whereas Mintz et al. collapsed all the mentions of the same entity tuple into a single datum. In other words, *Mintz++* constructs a *separate* classification data point from every sentence that contains a training tuple, whereas the original Mintz et al. algorithm merges the features extracted from all sentences that contain the same tuple into a single classification mention. The former approach was reported to perform better in practice by (Surdeanu et al., 2012).
- *Mintz++* allows multiple labels to be predicted for the same tuple by performing a union of all the labels proposed for individual mentions of the same tuple, whereas the Mintz et al. algorithm selected only the top-scoring label for a given entity pair. The multiple-label strategy was also adopted by other models ((Hoffmann et al., 2011); (Surdeanu et al., 2012)). This is necessary because the same pair of entities may express multiple relations, e.g., (*Balzac*, *France*) expresses at least two relations: *BornIn* and *Died*, which cannot be modeled by Mintz et al.'s algoritm.
- *Mintz++* implements a bagging strategy that combine five individual models. Each model is trained using four out of five folds of the training corpus. The final score is an unweighted average of the five individual scores. In this paper, we report results using two variants of the *Mintz++* model: when this ensemble modeling strategy is enabled (*Mintz++*) or disabled, i.e., using a single model trained over the entire training data (which we will call *Mintz**). This allows us to directly compare the effects of bagging with the impact of the data-cleanup proposed in this paper.

The results reported here are generated over the corpus created by (Riedel, Yao, and McCallum, 2010) and used by many other IE researchers like (Hoffmann et al., 2011), (Surdeanu et al., 2012), inter alia. This corpus uses Freebase as the source for distant supervision and the New York Times (NYT) corpus by (Sandhaus, 2008) for the source of textual evidence. The corpus contains two partitions: a training partition, containing 4700 relation mentions from the 2005–2006 portion of the NYT corpus, and a testing

[4]Task definition for 2013 available at `http://surdeanu.info/kbp2013/KBP2013_TaskDefinition_EnglishSlotFilling_1.0.pdf`

partition, containing 1950 more recent (2007) relation mentions. Because this corpus does not have a reserved development partition, we tuned our models over the training partition using cross-validation. In both partitions, negative mentions were automatically generated from pairs of entities that co-occur in the same sentence and are not recorded in Freebase with any relation label. Crucially, the corpus authors released only a random subsample of 5% of these negative mentions for the training partition. This means that any results measured over the training partition will be artificially inflated because the systems have fewer chances of inferring false positive labels.

4 Methods to Remove Noise

We tried three different heuristics to clean noisy mentions from the dataset. We experimented removing tuples depending on their mention frequency (MF), their pointwise mutual information (PMI), and the similarity between the centroids of all relation mentions and each individual mention (MC). We also built several ensemble strategies that combine the most successful individual strategies, as parametrized over development data. Note that none of these methods uses any additional manual annotation at all.

4.1 Mention Frequency (MF)

For our first heuristic, we consider that tuples with too many mentions are the most probable to contain noisy mentions, so we remove all those tuples that have more than a predefined number of mentions. Our system removes both positive tuples that appear in Freebase, and negative (unrelated) tuples which contain more than X mentions. We experimented with different thresholds and chose the limit that gave the highest F-Measure on the development set[5]. The chosen value was $X = 90$, i.e., all tuples with more than 90 mentions where removed, around 40% of the positive mentions, and 15% of the total dataset considering both positive and negative mentions.

For example, the tuple <European Union, *has-location*[6], Brussels> appears with 95 mentions. This tuple contains good mentions like *The European Union is headquartered in Brussels.* but also many noisy mentions like *The European Union foreign policy chief, Javier Solana, said Monday in Brussels that (...)* or *At an emergency European Union meeting of interior and justice ministers in Brussels on Wednesday, (...)* which do not explicitly say that Brussels is in the European Union, and can thus mislead the supervised RE system. This heuristic removes all instances of this tuple from the training data.

4.2 Tuple PMI

The second heuristic calculates the PMI between each entity tuple and the a relation label. Once we calculate the PMI for each tuple, we consider that the tuples which have a PMI below a defined threshold have noisy mentions, and remove them. Empirically, we observed that our system performs better if we remove only positive mentions with low PMI and keep the negative ones, regardless of their PMI value. Our system performed better with a threshold of 2.3, removing around 8% of the positive training tuples. This heuristic is inspired by the work of (Min et al., 2012).

As an example, this approach removed the tuple: <Natasha Richardson, *place-of-death*[7], Manhattan>. This tuple has only one mention: *(...) Natasha Richardson will read from 'Anna Christie,' (...) at a dinner at the Yale Club in Manhattan on Monday night.*. This mention does not support the place-of-death relation. That is, even though Natasha Richardson died in Manhattan, the mention is unrelated to that fact.

4.3 Mention Centroids (MC)

This heuristic calculates the centroid of all mentions with the same relation label, and keeps the most similar mentions to the centroids. We hypothesize that the noisy mentions are the furthest ones from their label centroids. For this experiment, we consider each mention as a vector and the features as space dimensions. We use the same features used by the DS system to build the vectors, with the frequency as the value of the feature. The centroid is built from the vectors as described in equation 1 below.

$$\vec{C_i} = \left(\frac{feat_1}{mentions_i}, \frac{feat_2}{mentions_i}, \ldots, \frac{feat_N}{mentions_i} \right) \quad (1)$$

[5] Throughout the paper, development experiments stand for cross-validation experiments on Riedel's training partition.

[6] /location/location/contains

[7] /people/deceased_person/place_of_death

where mentions_i = number of mentions for label i ($1 \leq i \leq M$), feat_j = number of appearances of feature j ($1 \leq j \leq N$) and C_i = Centroid for label i.

The similarity between a centroid and any given mention is calculated using the cosine:

$$cosine(C, M) = \frac{\vec{C} \cdot \vec{M}}{\sqrt{\vec{C} \cdot \vec{C}} \cdot \sqrt{\vec{M} \cdot \vec{M}}} \quad (2)$$

where C = Centroid and M = Mention.

We select a percentage of the most similar mentions to each centroid, and discard the rest. Our system returned the best results on development when we kept 90% of the most similar mentions of each relational label.

We do not use this heuristic for negative mentions. Empirically, we observed that this heuristic performs better if we kept all negative mentions rather than deleting any of them. This could be an artifact of the fact that only 5% of the negative mentions are included in Riedel's training dataset. Thus, sub-sampling negative mentions further yields datasets with too few negative mentions to train a discriminative model. This methods removes around 8% of the positive mentions.

As an example of the method, if we take the centroid for relation *company-founders*, the mention appearing in the sentence *(...) its majority shareholder is Steve Case, the founder of AOL* of the tuple <Steve Case, company-founders, AOL> is the most similar to the centroid of the same label. On the contrary, the mention *Ms. Tsien and Mr. Williams were chosen after a competition that began with 24 teams of architects and was narrowed to two finalists, Thom Mayne's Morphosis being the other* of the tuple <Thom Mayne, Morphosis> was correctly excluded, as the mention does not explicitly say that Thom Mayne is the founder of Morphosis.

4.4 Ensemble Models

We experimented with several ensemble models that combine the above individual strategies, in different order. The best results on development, as shown in Section 5.1, were different for *Mintz** and *Mintz++*. For the first, we first filtered using PMI, then run the MF filter, and finally applied the centroid-based filter. For the second, the best combination was to run PMI and then MF. The MC method did not provide any additional gain.

	Rec.	Prec.	F1
Mintz*	34.98	39.44	37.07
MF 90	33.19	44.49	**38.01**
PMI 2.3	34.49	40.64	37.31
MC 90%	34.81	40.31	37.33
PMI+MF+MC	32.72	46.36	**38.53**

Table 1: Development experiments using Mintz*, showing the results of each filtering method and the best combination.

	Rec.	Prec.	F1
Mintz++	34.85	41.45	37.86
MF 180	33.65	44.48	**38.45**
PMI 2.4	34.00	42.97	37.95
PMI+MF	33.25	45.57	**38.45**

Table 2: Development experiments using Mintz++, showing the results of each filtering method and the best combination.

5 Experiments

We evaluated the methods introduced in the previous section with the dataset developed by (Riedel, Yao, and McCallum, 2010). This dataset was created by aligning Freebase relations with the New York Times (NYT) corpus. They used the Stanford named entity recognizer to find entity mentions in text and constructed relation mentions only between entity mentions in the same sentences. We used the same features as (Riedel, Yao, and McCallum, 2010) for the mention classifier.

The development set was created using a three-fold cross-validation technique, similarly to (Surdeanu et al., 2012). For the formal evaluation on the test set, we only used the best ensemble models, instead of applying each method individually.

5.1 Results on the Development Corpus

The initial experiments where done using the *Mintz++* system in (Surdeanu et al., 2012) without any ensemble at the classifier. From now on, the *Mintz++* without the ensemble will be denoted as *Mintz** in this paper. Table 1 shows the results we obtained with each method. If we execute our methods individually, we get the best results with the *Mention frequency* experiment (Section 4.1), where

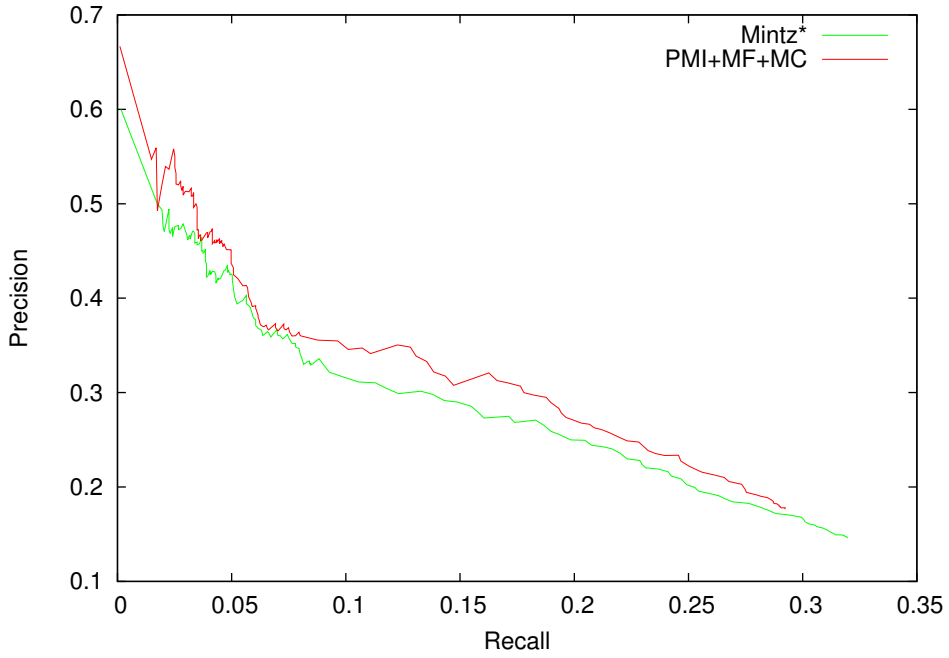

Figure 1: Precision/recall curves for the Mintz* system on the test partition. The red line is our best filtering model.

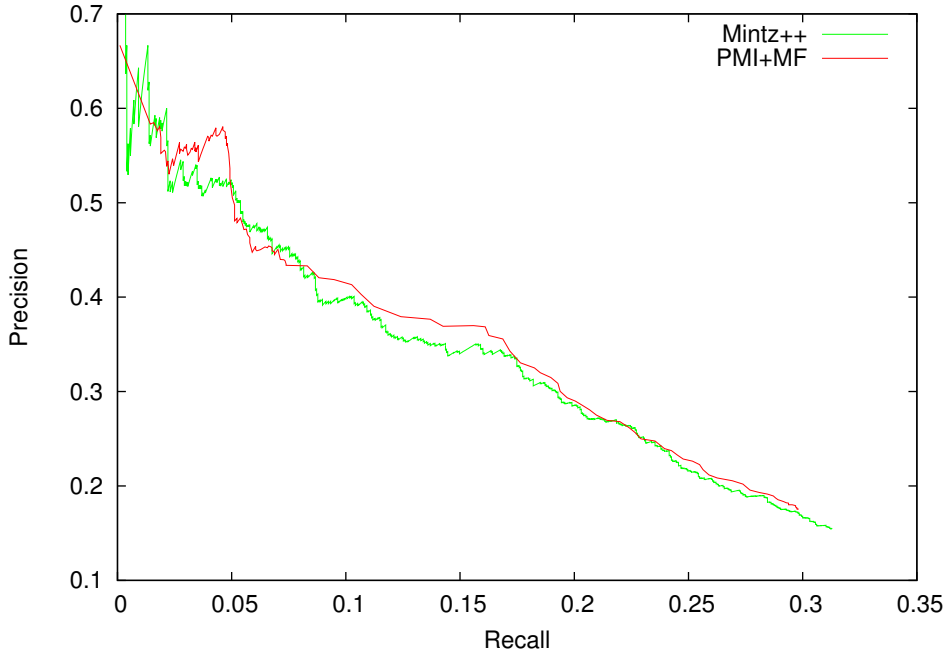

Figure 2: Precision/recall curves for the Mintz++ system on the test partition. The red line is our best filtering model.

our system's F-Measure improves nearly 1%. The PMI (Section 4.2) and the *Mention centroids* models (Section 4.3) both yield a small improvement over the baseline. For the ensemble models, we obtain the best performance by combining PMI with *Mention frequency* and the *Mention centroids*, improving the F-Measure nearly 1.5 absolute points. Our system improves the precision in each experiment, but not the recall, this scoring

	Rec.	Prec.	F1
Mintz*	31.95	14.57	20.02
PMI+MF+MC	29.23	17.64	**22.00**

Table 3: Results on the test partition for Mintz* (without bagging).

	Rec.	Prec.	F1
Mintz++	31.28	15.43	20.67
PMI+MF	29.79	17.48	**22.03**

Table 4: Results for Mintz++ (with bagging).

parameter generally decreases slightly. This is to be expected, since the models built using filtered data train on fewer positive mentions, thus they will be more conservative in predicting relation labels.

We applied the same heuristics to the original *Mintz++* system at (Surdeanu et al., 2012), and optimized them. The optimal parameters are 180 mention maximum for *Mention frequency* (4.1), and 2.4 for the *PMI* heuristics (Section 4.2). Unfortunately the *Mention centroids* (Section 4.3) heuristic did not yield an improvement here. Finally, we combined the *PMI* heuristic with the *Mention frequency* experiment to improve our results. Table 2 shows the results we obtained for each heuristic. Surprisingly, *MF 180* and *PMI+MF* give the same F-Measure.

5.2 Results on the Test Partition

For the formal evaluation on the test set, we only chose the ensemble models that performed best with the development set for *Mintz**, with the same optimal parameters obtained on development. On the test set, the F-Measure improves approximately 2 points. The results are shown in Table 3.

Figure 1 shows the precision/recall curves for our best system relative to the *Mintz** baseline. The figure shows that our approach clearly performs better.

Table 4 shows the results on the test partition of the original *Mintz++* system of (Surdeanu et al., 2012) and the *Mintz++* extended with our best ensemble filtering model (tuned on development).

Figure 2 shows the precision/recall curves of the two systems based on *Mintz++*. The models trained using filtered data perform generally better than the original system, but the differences are not as large as for the previous model that does not rely on ensemble strategies. This suggests that ensemble models, such as the bagging strategy implemented in *Mintz++*, are able to recover from some of the noise introduced by DS. However, bagging strategies are considerably more expensive to implement than our simple algorithms, which filters the data in a single pass over the corpus.

To check for statistical significance, we used the bootstrapping method proposed by (Berg-Kirkpatrick, Burkett, and Klein, 2012) verifying if the improvement provided by mention filtering is significant[8]. This bootstrapping method concluded that, although the difference between the two models is small, it is statistically significant with p-values below 0.001, thus supporting our hypothesis that data cleanup for DS algorithms is important.

6 Conclusions

Motivated by the observation that relation extraction systems based on the distant supervision approach are exposed to data that includes a considerable amount of noise, this paper presents several simple yet robust methods to remove noisy data from automatically generated datasets. These methods do not use any manual annotation at the datasets. Our methods are based on limiting the mention frequency for each tuple, calculating the Pointwise Mutual Information between tuples and relation labels, and comparing mention vectors against the mention centroids of each relation label.

We show that these heuristics, especially when combined using simple ensemble approaches, outperform significantly two strong baselines. The improvements hold even on top of a strong baseline that uses a bagging strategy to reduce sensitivity to training data noise.

References

Berg-Kirkpatrick, Taylor, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and*

[8]The statistical significance tests used the points at the end of the P/R curves.

Computational Natural Language Learning, EMNLP-CoNLL '12, pages 995–1005, Stroudsburg, PA, USA. Association for Computational Linguistics.

Craven, Mark and Johan Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 77–86. AAAI Press.

Hoffmann, Raphael, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 541–550, Stroudsburg, PA, USA. Association for Computational Linguistics.

Min, Bonan, Xiang Li, Ralph Grishman, and Sun Ang. 2012. New york university 2012 system for kbp slot filling. In *Proceedings of the Fifth Text Analysis Conference (TAC 2012)*. National Institute of Standards and Technology (NIST).

Mintz, Mike, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 1003–1011, Stroudsburg, PA, USA. Association for Computational Linguistics.

Riedel, Sebastian, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD '10)*.

Sandhaus, Evan. 2008. The new york times annotated corpus. In *Linguistic Data Consortium, Philadelphia*.

Surdeanu, Mihai, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 455–465, Stroudsburg, PA, USA. Association for Computational Linguistics.

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 49-56 recibido 30-04-2013 revisado 18-06-2013 aceptado 21-06-2013

Extracting Drug-Drug Interaction from Text Using Negation Features

Estudio del efecto de la negación en la detección de interacciones entre fármacos

Behrouz Bokharaeian, Alberto Díaz
NIL Group
Universidad Complutense de Madrid
Madrid, Spain
behroubo@ucm.es, albertodiaz@fdi.ucm.es

Miguel Ballesteros
Natural Language Processing Group
Universitat Pompeu Fabra
Barcelona, Spain
miguel.ballesteros@upf.edu

Resumen: La extracción de relaciones entre entidades es una tarea muy importante dentro del procesamiento de textos biomédicos. Se han desarrollado muchos algoritmos para este propósito aunque sólo unos pocos han estudiado el tema de las interacciones entre fármacos. En este trabajo se ha estudiado el efecto de la negación para esta tarea. En primer lugar, se describe cómo se ha extendido el corpus DrugDDI con anotaciones sobre negaciones y, en segundo lugar, se muestran una serie de experimentos en los que se muestra que tener en cuenta el efecto de la negación puede mejorar la detección de interacciones entre fármacos cuando se combina con otros métodos de extracción de relaciones.
Palabras clave: Interacciones entre fármacos, negación, funciones kernel, máquinas de vectores de soporte, funciones kernel.

Abstract: Extracting biomedical relations from text is an important task in BioMedical NLP. There are several systems developed for this purpose but the ones on Drug-Drug interactions are still a few. In this paper we want to show the effectiveness of negation features for this task. We firstly describe how we extended the DrugDDI corpus by annotating it with the scope of negation, and secondly we report a set of experiments in which we show that negation features provide benefits for the detection of drug-drug interactions in combination with some simple relation extraction methods.
Keywords: Drug-Drug interaction, Negation, Support vector machines, kernel-based methods

1. Introduction

A drug-drug interaction (DDI) occurs when one drug affects the level or activity of another drug, this may happen, for instance, in the case of drug concentrations. This interaction can result on decreasing its effectiveness or possibly altering its side effects that may even the cause of health problems to patients (Stockley, 2007).

There is a great amount of DDI databases and this is why health care experts have difficulties to be kept up-to-date of everything published on drug-drug interactions. This fact means that the development of tools for automatically extracting DDIs from biomedical resources is essential for improving and updating the drug knowledge and databases.

There are also many systems on the extraction of biomedical relations from text; however the research on studying the effect of negation in biomedical relation extraction is still limited. On the other hand, negation is very common in clinical texts and it is one of the main causes of making errors in automated indexing systems (Chapman et al., 2001); the medical personnel is mostly trained to include negations in their reports. Particularly when we are detecting the interaction between drugs, the presence of negations can produce false positives classifications, for instance, the sentence *Coadministration of multiple doses of 10 mg of lenalidomide had no effect on the single dose pharmacokinetics of R- and S- warfarin* a DDI between *lenalidomide* and *warfarin* could be detected as a practicable fact if negation is not taken into account. We therefore believe that detecting the words that

ISSN 1135-5948

are affected by negations may be an essential part in most biomedical text mining tasks that try to obtain automatically the accurate knowledge from textual data.

In order to avoid errors derived of using automatic negation detection algorithms such as NegEx (Amini et al., 2011), we annotated a DDI corpus - previously developed- with the scope of negations. The corpus is called DrugDDI corpus (Segura-Bedmar et al., 2011b), and it was developed for the Workshop on Drug-Drug Interaction Extraction (Segura-Bedmar et al., 2011a) that took place in 2011 in Huelva, Spain. The DrugDDI corpus contains 579 documents extracted from the DrugBank database. We analyzed the corpus and we annotated the sentences within with the scope of negation in order to find the effect of negation features in the detection of DDIs. We annotated it using the same guidelines of the BioScope corpus (Vincze et al., 2008), that is, we annotated cues and scopes affected by negation statements into sentences in a linear format.

For detecting the DDIs we used a fast version of a support vector machine (henceforth, SVM) classifier with a linear kernel based on a bag of words (henceforth, BOW) representation obtained from the extracted features. We carried out some experiments with different kernels (global context, subtree, shortest path), with and without negation information. The results presented in this paper show that negation features can improve the performance of relation extraction methods.

The rest of the paper is structured as follows. In Section 2 we discuss previous related work about biomedical relation extraction and relevant information about the DrugDDI corpus. In Section 3 we explain how we annotated the corpus with the scope of negation. In Section 4 we explain how we used the obtained information from negation tags to improve the DDI detection task. In Section 5 we discuss the results obtained. Finally, in Section 6, we show our conclusions and suggestions for future work.

2. Related work

In this Section we describe the DrugDDI corpus and we present some related work on kernel-based relation extraction.

2.1. DrugDDI corpus

There are some annotated corpora that were developed with the intention of studying biomedical relation extraction, such as, Aimed (Bunesu et al., 2005), LLL (Nedellec et al., 2005), BioCreAtIvE-PPI (Krallinger et al., 2008) on protein-protein interactions (PPI) and DrugDDI (Segura-Bedmar et al., 2011b), on drug-drug interactions. In particular, the DrugDDI corpus is the first annotated corpus on the phenomenon of interactions among drugs and it is the one that we used for our experiments. It was designed with the intention of encouraging the NLP community to conduct further research on this type of interactions. The DrugBank database (Wishart et al., 2008) was used as source to develop this corpus. This database contains unstructured textual information on drugs and their interactions.

The DrugDDI corpus is available in two different formats: (i) the first one contains the information provided by MMTX (Aronson, 2001) and the unified format adapted from PPI corpora format proposed in (Pyysalo et al., 2008). The unified XML format (see Figure 1) does not contain any linguistic information; it only provides the plain text sentences, their drugs and their interactions. Each entity (drug) includes reference (origId) to the sentence identifier in the MMTX format corpus. For each sentence contained in the unified format, the annotations correspond to all the drugs entities and the possible DDI candidate pair that represents the interaction. Each DDI candidate pair is represented as a *pair ID* node in which the identifiers of the interacting drugs are registered on its *e1* and *e2* attributes. If the pair is a DDI, the *interaction* attribute is set to *true*, otherwise this attribute is set to *false*. Table 1 shows related statistics of the DrugDDI corpus (Segura-Bedmar et al., 2011b).

2.2. Biomedical Relation Extraction

Nowadays, there are many systems developed for extracting biomedical relations from text that can be categorized in (i) feature based and (ii) kernel-based approaches. Feature-based approaches transform the context of entities into a set of features; this set is used to train a data-driven algorithm. On the other hand, kernel-based approaches are

```
−<sentence id="DrugDDI.d346.s0" origId="s0" text="Uricosuric Agents: Aspirin may decrease the effects of probenecid,
 sulfinpyrazone, and phenylbutazone.">
   <entity id="DrugDDI.d346.s0.e0" origId="s0.p0" charOffset="0-17" type="drug" text="Uricosuric Agents"/>
   <entity id="DrugDDI.d346.s0.e1" origId="s0.p2" charOffset="19-26" type="drug" text="Aspirin"/>
   <entity id="DrugDDI.d346.s0.e2" origId="s0.p6" charOffset="55-65" type="drug" text="probenecid"/>
   <pair id="DrugDDI.d346.s0.p0" e1="DrugDDI.d346.s0.e0" e2="DrugDDI.d346.s0.e1" interaction="false"/>
   <pair id="DrugDDI.d346.s0.p1" e1="DrugDDI.d346.s0.e0" e2="DrugDDI.d346.s0.e2" interaction="false"/>
   <pair id="DrugDDI.d346.s0.p4" e1="DrugDDI.d346.s0.e1" e2="DrugDDI.d346.s0.e2" interaction="true"/>
</sentence>
```

Figure 1: The unified XML format in the DrugDDI corpus.

	No.
Documents	579
Sentences	5,806
Drugs	8,260
Sentences with at least two drug	3,775
Sentences with at least one DDI	2,044
Sentences with no DDI	3,762
Candidate drug pairs	30,757
Positive interactions	3,160
Negative interactions	27,597

Table 1: Basic statistics for the DrugDDI corpus.

based on similarity functions. This idea provides the option of checking structured representations, such as parse trees and computing the similarity between different representations directly. Combing kernel based and feature based approaches were investigated by Thomas et al. (2011), they developed a voting system (based on majority) that benefits from the outcomes of several methods.

So far, the sequence and tree kernels are the ones that have shown a superior performance for the detection of biomedical relations from text (Bunesu et al., 2005). In particular, global context kernel, subtree and shortest path kernels are three important kernel methods that were applied successfully for biomedical relation extraction task. For instance, Giuliano et al. (2005) applied by considering three different patterns and they calculated the similarity between two sentences by computing common n-grams of two different patterns.

The shortest path kernel (Bunescu y Mooney, 2005) uses the shortest path between two entities (or drugs) in a phrase structure tree. The subtree kernel (Moschitti, 2006) counted the number of common subtrees in whole parse trees by comparing two different sentences. Moreover, a comparative survey about different kernel-based approaches and their performances can be found in (Frunza y Inkpen, 2010).

More recent research on tree kernels were carried out by Guodong et al. (2010). They introduced a "context-sensitive" convolution tree kernel, which specifies both "context-free" and "context-sensitive" sub-trees by traversing the paths of their ancestor nodes as their contexts to capture structural information in the tree structure. Another motivating work was reported by Chen et al. (2011), that presented a proteing-protein interaction pair extractor, it consists on a SVM classifier that exploits a linear kernel with a complete set of features.

Finally, Simões et al. (2013) introduced an approach for Relation Extraction (RE) based on labeled graph kernels, they proposed an implementation of a random walk kernel (Neuhaus y Bunke, 2006) that mainly explores two characteristics: (i) the words between the candidate entities and (ii) the combined information from different sources.

3. Annotating the DrugDDI corpus with negations

We followed the Bioscope guidelines in order to annotate the corpus (Vincze et al., 2008). The main idea is based in the detection of a set of negation cues, like 'no' or 'not'. After this, the scope of the cue is calculated based on its syntactic context. There are several systems that annotate the scope of negation, in our approach we used the one published by Ballesteros et al. (Ballesteros et al., 2012), which is publicly available,[1] rule-based system that works on biomedical literature (Bioscope) and the input is just the sentence without any required annotation, which serves very well for our purposes.

We used as input all the sentences of the DrugDDI corpus, containing 5,806 sentences and 579 files. The output was therefore a set of sentences annotated with the scope of negation. After applying the system, we observed that there were a set of 1,340 sentences containing negations in the corpus,

[1] `http://minerva.fdi.ucm.es:8888/ScopeTagger/`

which conforms 23% of the corpus.

Taking into account that the negation scope detection system is fully automatic, we manually checked the outcome correcting the annotations when needed. In order to do so, we divided the annotated corpus in 3 different sets that were assigned to 3 different evaluators. The evaluators checked all the sentences contained in each set and corrected the sentences that contained annotation errors. After this revision, a different evaluator revised all the annotations produced by the other 3 evaluators. Finally, we got the whole set of 1,340 sentences (correctly) annotated with the scope of negation.

The algorithm produced errors -according to the evaluators- in 350 sentences from the 1,340, including false positives matches (there were 16 cases). Which means that 74% of sentences was annotated correctly in an automatic way, when considering a full scope match. The main errors produced by the algorithm were related with the processing of passive voice sentences, commas, and copulative keywords (and, or). In particular the problem of passive voice sentences was related with the pattern *It + to be + not + past participle*, which seems that it was not captured by the system, at least in all cases. The false positives were related with the cue *failure*, which is not a negation when it is a noun modified by an adjective, for instance, *renal failure* or *heart failure*. In the DrugDDI corpus these words appear always as nouns, and therefore all of the performed annotations were incorrect.

The following paragraph shows some examples and corrections made by the evaluators:

- Scope closed in an incorrect way containing words from two different clauses such as: Example: *It is [{not} clear whether this represents an interaction with TIKOSYN or the presence of more severe structural heart disease in patients on digoxin;]*. The scope should be closed in or.
- Scope closed in an incorrect way in copulative coordinated sentences: Example: *The following medications have been administered in clinical trials with Simulect? with [{no} increase in adverse reactions: ATG/ALG] , azathioprine, corticosteroids, cyclosporine, mycophenolate mofetil, and muromonab-CD3.*
- Scope opened incorrectly in coordinated copulative sentences: Example: *In an in vitro study, cytochrome P450 isozymes 1A2, 2A6, 2C9, 2C19, 2D6, 2E1, [and 3A4 were{not} inhibited by exposure to cevimeline].*
- Some passive voice sentences were not detected. In particular, as it is already mentioned, sentences with the format 'It (this and that) + finite form of to be + not + past participle'. Example: *[Concomitant use of bromocriptine mesylate with other ergot alkaloids is{not} recommended].*

We also carried out some analysis concerning the number of different cues in the corpus and the number of different errors observed. Table 2 shows that *not* and *no* are by far the most frequent cues in the corpus. It can be observed that the most problematic cue is *neither ... nor*.

Cue	No.	MODFs	Rate
Not	855	266	31.1%
No	439	58	13.2%
without	47	8	17.0%
Neither ... nor	14	12	85.7%
Absence	10	5	50%
Lack	8	1	12.5%
cannot	7	4	57.1%

Table 2: Statistics of negations cues in the corpus and modifications for each cue in the manual checking process.

We finally explored the sentences that are not automatically annotated but they indeed show a negative statement in order to find false negatives. We looked into several negations cues that are not detected by the system such as *unaffected, unchanged* or *nonsignificant*. We detected and corrected 75 different sentences that belong to this problem.

Here we show some examples of false negatives:

- *[The pharmacokinetics of naltrexone and its major metabolite 6-beta-naltrexol were {unaffected} following co-administration with Acamprosate].*
- *[Mean T max and mean plasma elimination half-life of albendazole sulfoxide were {unchanged}].*

- *Monoamine Oxidase Inhibition: Linezolid is a reversible, [{nonselective} inhibitor of monoamine oxidase].*

Therefore, the corpus finally contains 1,399 sentences annotated with the scope of negation, of which 932 correspond to sentences in which there are at least two drugs mentioned. It is worth mentioning that there are 1,731 sentences with 2 or more drug mentions but no DDI, and 2,044 with 2 or more drugs and at least one interaction.

Finally, the extension of the DrugDDI corpus consists of adding a new tag in the annotation of each sentence with the scope of negations. Figure 3 shows an example. The produced corpus is available for public use.[2]

4. DDI detection

In this Section, we explain in detail the experiments we carried out by using negation features. First, we illustrate in detail the methods we used without negation features, and then we present our proposed combined negation method, see figure 4. All the experiments were carried out by using the Stanford parser[3] for tokenization and constituent parsing (Cer et al., 2010), and the SVMs provided by Weka as training engine.

4.1. DDI detection without negation features

The DDI extraction method consists of four different processes: (1) initial prepossessing, (2) feature extraction, (3) Bag of Words computation and (4) classification. The preprocessing step (1) consists of removing some stop words and tokens, for instance removing question marks at the beginning of the sentence. We also carried out a normalization task for some tokens due to the usage of different encoding and processing methods, mainly HTML tags. In the feature extraction step (2), we extracted three different feature sets corresponding to different used relation extraction methods. The feature extraction step for global context kernel consists of extracting *fore-between, between*, and *between-after* tokens that we mentioned in Section 2. The feature extraction step for shortest path kernel method included constituent parsing of the sentence and then extracting shortest path between two drugs in the generated parse tree. And for the subtree kernel we also extracted all subtrees from the mentioned constituent parse tree. After extracting features, we applied the BOW method (3) to generate new feature sets that the SVM classifier uses. The aim of this step is producing a new representation of the instances which is used in the classification step. And finally in the classification step (4), we applied the Weka SVM classifier (Platt, 1998) (SMO), with a linear composition of features produced by the BOW method to detect the interactions among drugs. The Inner product of new features was used as kernel function between two new representations.

4.2. DDI detection with negation features

In this section, we explain our proposed method that merges negation features with the features mentioned in Section 4.1. We divided the corpus in instances affected by negation and instances without negation statements. The last ones were classified in the same way as in Section 4.1, while for the instances with negations we added negation features to the representation. The positive instances were classified in the same way as previous approaches but the sentences containing negations were categorized using negation features in addition to the other previous features. As in previous subsection, the combined method for instances containing negations consists of 4 steps. After a simple preprocessing step we carried out a feature extraction process. In this step, we generated six negation features in addition to three feature sets corresponding to global context kernel (GCK), Shortest Path and Subtree kernel methods. Negation feature consists of tokens inside the negation scope, left side tokens outside of the negation scope and right side tokens, and the negation cue tokens, negation cue, and position of open and closed negation scope. For instance in the sentence shown in Figure 3: tokens inside brackets create middle scope features, right side tokens construct right features and tokens in the left side of the negation scope form left scope features. As in the previous subsection, we used a BOW method to convert negation string features to word features. Finally, the new feature set is used to classify the drug-drug interactions by

[2] `http://nil.fdi.ucm.es/sites/default/files/NegDrugDDI.zip`

[3] `http://nlp.stanford.edu/software/lex-parser.shtml`

```
<sentence origId="s0" id="DrugDDI.d291.s0" text="Zidovudine: There is no significant
pharmacokinetic interaction between ZDV and zalcitabine which has been confirmed
clinically.">
        <entity .... />
        <pair .... />
        <negationtags>Zidovudine: There is <xcope><cue>no</cue> significant pharmacokinetic
            .... clinically</xcope>.</negationtags>
</sentence>
```

Figure 2: A sentence annotated with the Scope of Negation in our version of the DrugDDI corpus.

making use of the Weka SVM.

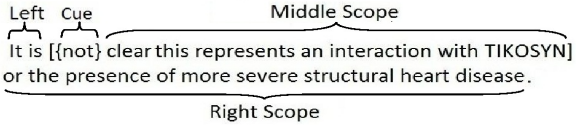

Figure 3: Left, middle and right side scope and negation cue in a negative sentence.

In summation, our approach is a feature based method that uses a bag of word kernel utilizing basic features to compute simple basic kernels and negation features. We applied a fast implementation of the support vector machine provided in Weka, which uses sequential minimal optimization. By carrying out some experiments we also limited the size of the words in each feature bag in the BOW approach to 1000 words per feature class.

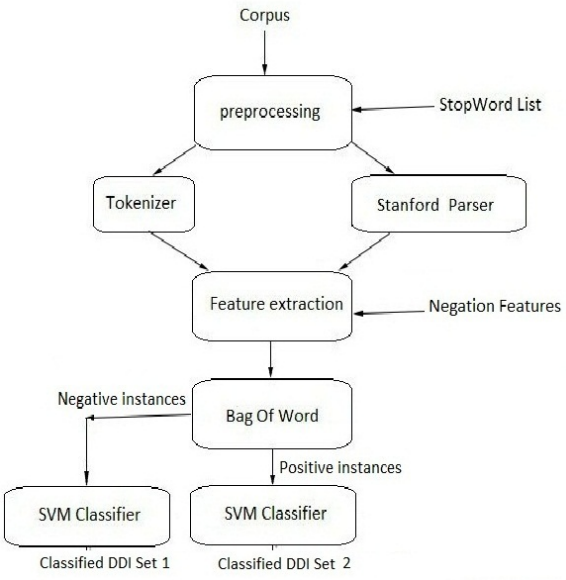

Figure 4: The different processes followed by our proposed approach.

5. *Evaluation*

5.1. Evaluation Setup

In order to demonstrate the improvements provided by using negation features, our experiments consisted of a 10-fold cross validation over the training part of the DrugDDI corpus. Therefore, our results are not directly comparable to the ones provided in the DDI challenge (Segura-Bedmar et al., 2011a). The training DrugDDI corpus contains 437 documents extracted from DrugBank database. It consists of 4267 sentences with average of 9.8 sentences per document and 25,209 instances with 2,402 interactions between different drugs.

Our measurement metrics included true positive, false positive, false negative, total number of positive instances, Precision, Recall and F-1 score.

5.2. Results

Table 3 shows the outcomes of the experiments by computing the metrics and by training over the DDI corpus. The table shows results for Global context kernel (GCK), Subtree and shortest path kernel (SubtreeK) and corresponding combined negation methods (GCKNS = GCK with negation features; SubtreeKNS = Subtree kernel with negation features). The first three rows of the table show the performance of the three basic kernels and the last three ones (with NS postfix) show the outcomes for the combined version that includes negation features. The best result was obtained with GCKNS, and the worst result was obtained by the shortest path tree approach. Moreover, the best improvement was obtained by the GCK approach; it improves 3.8 percentual points of the F score.

As we can see in the table, there is an improved performance when we applied the negation features for classification. This fact

Method	TP	FP	FN	Total	P	R	F1
GCK:	902	1094	1500	2402	0.452	0.376	0.410
SubtreeK:	818	1105	1584	2402	0.425	0.341	0.378
ShortestPathTK:	795	1066	1607	2402	0.427	0.331	0.373
GCKNS:	987	1021	1415	2402	0.492	0.411	0.448
SubtreeKNS:	919	1280	1483	2402	0.418	0.383	0.399
ShortestPathTKNS:	936	1240	1466	2402	0.430	0.390	0.409

Table 3: 10- cross validation results for the methods that do not use negation features and the methods that use negation features.

demonstrates our hypothesis and the emphasizes the purpose of the present work.

6. Conclusions and Future Work

Due to the huge amount of drug related information in bio-medical documents and the importance of detecting dangerous drug-drug interactions in medical treatments, we believe that implementing automatic Drug-Drug interaction extraction methods from text is critical. The DrugDDI corpus is the first annotated corpus for Dug-Drug interaction tasks used in the DDI Extraction 2011 challenge.

In this paper, after reviewing related work on biomedical relation extraction, we first explained the process of annotating the DrugDDI corpus with negation tags; and then we explored the performance of combing negation features with three simple relation extraction methods. Our results show the superior performance of the combined method utilizing negation features over the three basic experimented relation extraction methods.

However, the experiments also show that the application of negation features can indeed improve the relation extraction performance but the obtained improvement clearly depends on the number and rate of positive and negative relations, rate of negative cues in the corpus, and other relation extraction features. It is also true that combining negation features with a huge number of other features may not improve the performance and even may hurt the final result, and this is why we used a limited number of features. It is therefore obvious that corpora having more sentences with negation cues can benefit more from using negation features.

For further work, we plan to use a different type of annotation such as negation events instead of scopes, and also handling hedge cues and speculative statements in conjunction with negations.

References

Amini, I., M. Sanderson, D. Martinez y X. Li. 2011. Search for clinical records: Rmit at trec 2011 medical track. En *Proceedings of the 20th Text REtrieval Conference (TREC 2011)*.

Aronson, A. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. En *Proceedings of the AMIA Symposium*, páginas 17–27. URL http://metamap.nlm.nih.gov/.

Ballesteros, M., V. Francisco, A. Diaz, J. Herrera y P. Gervas. 2012. Inferring the scope of negation in biomedical documents. En *Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2012)*.

Bunescu, R., R. Ge, R. Kate, E. Marcotte, R. Mooney, A. Ramani y Y. Wong. 2005. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33(2):139–155.

Bunescu R. y R. Mooney. 2005. A shortest path dependency kernel for relation extraction. En *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, páginas 724–731.

Cer, D., M. de Marneffe, D. Jurafsky y C. D. Manning. 2010. Parsing to Stanford Dependencies: Trade-offs between speed and accuracy. En *Proceedings of the 7th In-*

ternational Conference on Language Resources and Evaluation(LREC 2010) .

Chapman, W., W. Bridewell, P. Hanbury, G. F. Cooper y B. G. Buchanan. 2001. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informactics*, 34(5):301-310.

Chen, Y., F. Liu y B. Manderick. 2011. Extract Protein-Protein Interactions From the Literature Using Support Vector Machines with Feature Selection. *Biomedical Engineering, Trends, Researchs and Technologies*.

Frunza, O. y D. Inkpen. 2010. Extraction of disease-treatment semantic relations from biomedical sentences. *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, páginas 91–98.

Giuliano, C., A. Lavelli y L. Romano. 2005. Exploiting shallow linguistic information for relation extraction from biomedical literature. En *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, páginas 5–7.

Guodong, Z., Q. Longhua y F. Jianxi. 2010. Tree kernel-based semantic relation extraction with rich syntactic and semantic information. *International Journal on Information Sciences*, 180(8):1313–1325.

Krallinger, M., A. Valencia y L. Hirschman. 2008. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biology* , 9(Suppl 2):S8.

Moschitti, A. 2006. Making Tree Kernels Practical for Natural Language Learning. En *Proccedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*.

Nedellec, C. 2004. Machine Learning for Information Extraction in Genomics - State of the Art and Perspectives. *Text Mining and its Applications*, Springer Verlag.

Neuhaus, M. y H. Bunke. 2006. A Random Walk Kernel Derived from Graph Edit Distance. *Lecture Notes in Computer Science*, 4109(5):191-199.

Platt, J. 1998. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. *Advances in kernel methods - Support vector learning*.

Pyysalo, S., A. Airola, J. Heimonen, J. Bjorne, F. Ginter y T. Salakoski. 2008. Comparative analysis of five protein-protein interaction corpora. *BMC bioinformatics*, 9(Suppl 3):S6.

Segura-Bedmar, I., P. Martínez y D. Sánchez-Cisneros. 2011. En *Proceedings of the 1st Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011)*. CEUR Workshop Proceedings, Vol. 761.

Segura-Bedmar, I., P. Martínez y C. de Pablo Sánchez. 2011. Using a shallow linguistic kernel for drug-drug interaction extraction. *Journal of Biomedical Informatics*, 44(5):789–804.

Simões, G., D. Matos y H. Galhardas. 2013. A Labeled Graph Kernel for Relationship Extraction. *CoRR*, abs/1302.4874.

Stockley, I. H. 2007. *Stockley's Drug Interaction*. Pharmaceutical Press.

Thomas, P., M. Neves, I. Solt, D. Tikk y U. Leser. 2011. Relation extraction for drug-drug interactions using ensemble learning. En *Proceedings of the First Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011)*, pp:11–17.

Vincze, V., G. Szarvas, R. Farkas, G. Mora y J. Csirik. 2008. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. *BMC Bioinformatics*, 9(Suppl 11):S9.

Wishart, D. R., C. Knox , A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali. 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, 36(Suppl 1):D901-D906.

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 57-64 recibido 18-04-2013 revisado 18-06-2013 aceptado 21-06-2013

Characterising social media users by gender and place of residence

Caracterización de los usuarios de medios sociales mediante lugar de residencia y género

Óscar Muñoz-García
Havas Media Group
Madrid - Spain
oscar.munoz@havasmg.com

Jesús Lanchas Sampablo, David Prieto Ruiz
Acceso
Madrid - Spain
jlanchas@acceso.com, dprieto@acceso.com

Resumen: La caracterización de los usuarios mediante atributos sociodemográficos es un paso necesario previo a la realización de estudios de opinión a partir de información publicada por dichos usuarios en los medios sociales. En este trabajo se presentan, comparan y evalúan diversas técnicas para la identificación de los atributos "género" y "lugar de residencia", a partir de los metadatos asociados a dichos usuarios, así como el contenido publicado y compartido por los mismos, y sus redes de amistad. Los resultados obtenidos demuestran que la información proporcionada por la red social es muy útil para identificar dichos atributos.
Palabras clave: demografía, género, lugar de residencia, usuarios, análisis de medios sociales

Abstract: Characterising users through demographic attributes is a necessary step before conducting opinion surveys from information published by such users in social media. In this paper, we describe, compare and evaluate different techniques for the identification of the attributes "gender" and "place of residence" by mining the metadata associated to the users, the content published and shared by themselves, and their friendship networks. The results obtained show that the social network is a valuable source of information for obtaining the sociodemographic attributes of single users.
Keywords: demographics, genre, place of residence, social media analysis, users

1. Introduction

Social media has revolutionized the way in which organizations and consumers interact. Users have adopted massively these channels to engage in conversations about content, products, and brands, while organizations are striving to adapt proactively to the threats and opportunities that this new dynamic environment poses. Social media is a knowledge mine about users, communities, preferences and opinions, which has the potential to impact positively marketing and product development activities (Weber, 2007).

Social media monitoring tools are being used successfully in a range of domains (including market research, online publishing, etc.). Most of these tools generate its reports from metrics based on volume of posts and on opinion polarity about the subject that is being studied. Although such metrics are good indicators of subject popularity and reputation, these metrics are often inadequate for capturing complex multi-modal dimensions of the subjects to be measured that are relevant to business, and must be complemented with ad-hoc studies such as opinion polls.

The validity of these social metrics depends to a large extent on the population over which they are applied. However, social media users cannot be considered a representative sample until the vast majority of people regularly use social media. Therefore, until then, it is necessary to identify the different strata of users in terms of socio-demographic attributes (e.g., gender, age or geographical precedence), in order to weight their opinions according to the proportion of each stratum in the population (Gayo-Avello, 2011). Author and content metadata is not enough for capturing such attributes. As an example, not all the social media channels qualify their

ISSN 1135-5948

users neither with gender nor with geographical location. Some channels, such as Twitter, allow their authors to specify their geographical location via a free text field. However, this text field is often left empty, or filled with ambiguous information (e.g., Paris - France vs. Paris - Texas), or with other data that is useless for obtaining real geographical information (e.g., "Neverland"). For these cases, the friendship networks and the content shared and produced by social media users can be used for estimating their socio-demographic attributes, applying techniques such as geographical entity recognition.

This paper explores different techniques for obtaining the place of residence and gender attributes. Such techniques exploit social users' metadata, the content published and shared by the users to be categorised, and their friendship networks.

The paper is structured as follows. Section 2 summarises related work. Section 3 describes techniques for the identification of the "place of residence" attribute. Section 4 describes techniques for gender recognition. Section 5 evaluates and compare the techniques. Finally, Section 6 presents the conclusions and future lines of work.

2. Related work

The identification of the geographical origin of social media users has been tackled in the past by several research works.

In (Mislove et al., 2011) geographical location is estimated for Twitter users by exploiting the self-reported location field in the user profile, which correspond to the technique described in Subsection 3.1.

Regarding content-analysis approaches, in (Cheng, Caverlee, and Lee, 2010) the authors propose to obtain user location based on content analysis. The authors use a generative probabilistic model that relates terms with geographic focuses on a map, placing 51 % of Twitter users within 100 miles of their actual location. This probabilistic model was previously described in (Backstrom et al., 2008). In (wen Chang et al., 2012) a similar approach is followed, consisting in estimating the city distribution on the use of each word.

In addition, in (Burger et al., 2011) the authors describe a method for obtaining user regional origin from content analysis, testing different models based on Support Vector Machines (Cortes and Vapnik, 1995), achieving a 71 % of accuracy when applying a model of socio-linguistic-features.

With respect to gender identification, in (Burger et al., 2011) the use of profile metadata to identify the gender of the authors is proposed. Using only the full name of the author, an accuracy of 0.89 is reached. Using the author description, the screen name and the tweet text the obtained accuracy is 0.92.

Another relevant related work regarding gender identification is described in (Rao et al., 2010). In this case the proposed method, based on SVM, tries to distinguish the author gender exclusively from the content and style of their writing. This solution needs an annotated seed corpus with authors classified as male or female, to create the model used by the SVM classifier. In this case the accuracy of the best model is 0.72, lower than considering the full name of the author.

3. Place of residence recognition

We have tested different techniques for identifying the place of residence of users, defining "place of residence of a user" as the geographical location where a user lives usually. Each technique is described next.

3.1. Technique based on metadata about locations of users

This technique makes use of the location metadata in the user profile, as for example, the *location* attribute returned by Twitter API when querying user details (Twitter, 2013). Users may express their location in different forms through this attribute, such as geographical coordinates, or the name of a location (e.g., a city, a country, a province, etc.). Therefore, a normalization stage is required in order to obtain a standard form for each location.

For normalising the location this technique makes use of a geocoding API. Our implementation uses Google Maps Web services. This technique invokes a method of the geocoding API that analyses a location and return a normalised tuple composed by a set of components that define the location, including *latitude*, *longitude*, *locality*, and *country*, among others. As for example, if the request "santiago" is sent to the Web service, the response will be a tuple containing "Chile" as the country and "Santiago" as the locality, among other location components. The complete list of components is listed in the API

```
1 function ResidenceFromLocationData(user)
2     return GeoCode(location(user))
```

Listing 1: Technique based on metadata about locations of users

documentation (Google, 2013). Please note that this query does not provide enough information for disambiguating locations, i.e., "santiago" may refer to many geographical locations, including *Santiago de Chile* and *Santiago de Compostela (Spain)*. Therefore the precision of this technique depends on how users describe their location when filling in their profiles. For example, geographical coordinates will define locations accurately, while combinations of city and country (e.g., "Villalba, Spain") will enhance disambiguation (although not completely). In addition, this technique does not return a place of residence when users have not filled in the location field contained in user's profile form of the social network. The technique described next deals with these precision and coverage issues.

Listing 1 summarises the steps executed by this technique.

3.2. Technique based on friendship networks

This technique exploits the homophily principle in social networks (McPherson, Smith-Lovin, and Cook, 2001) for obtaining the place of residence of users. Listing 2 summarises the steps executed by this technique, which are described next.

1. Firstly, we execute the previous technique for obtaining the place of residence of a given user. If a result is obtained, the process finishes. If not, the steps described next are executed (line 2).
2. Secondly, the friends of the user in her online community are collected. After that, the location of each friend is obtained by using the geocoding API. The normalised locations obtained are appended to a list (lines 5-6).
3. Finally, the list obtained in the previous step is filtered iteratively selecting on each iteration the locations that contain the value with the most frequency for a given location component, starting from the country and finishing in the street number, until there is only one location in the set. First the locations whose country are the most frequent are selected, then the locations whose first-order civil entity (e.g., a state in USA or an autonomous community in Spain) is the most frequent, and so forth. The location that remains in the list after completing the filtering iterations is selected as the place of residence of the user. This approach ensures that the most frequent regions in the friendship network of the user are selected(line 7).

```
1 function ResidenceFromFriends(u)
2     l ⇐ ResidenceFromLocationData(u)
3     if l = ∅ then
4         L ⇐ ∅
5         for each f in friends(u) then
6             L ⇐ L ∪ {GeoCode(location(f))}
7         l ⇐ MostFrequentLocation(L)
8     return l
```

Listing 2: Technique based on friendship networks

3.3. Technique based self-descriptions of users

This technique exploits the description published by users about themselves in their profiles for obtaining their place of residence, as for example, the *description* attribute returned by Twitter API when querying a user profile (Twitter, 2013). Listing 3 summarises the steps executed by this technique, which are described next.

1. Firstly, we execute the technique described in Subsection 3.1). If a result is obtained, the process finishes. If not, the steps described next are executed (line 2).
2. Secondly, we obtain the user self-description attribute. Such attribute usually consist on a sentence that have to be processed for extracting the geographical locations mentioned in the text (line 4).
3. After obtaining the description of the user, we identify the language in which user self-description is written. For doing so, we make use of the Freeling (Padró and Stanilovsky, 2012) language recognition feature (line 5).

```
1  function ResidenceFromDescription(u)
2      l ⇐ ResidenceFromLocationData(u)
3      if l = ∅ then
4          desc ⇐ description(u)
5          lang ⇐ IdentifyLanguage(desc)
6          E ⇐ NamedEntities(desc, lang)
7          L ⇐ ∅
8          for each entity in E do
9              if isLocation(entity) then
10                 L ⇐ L ∪ {GeoCode(entity)}
11         l ⇐ MostFrequentLocation(L)
12     return l
```

Listing 3: Technique based on self-descriptions of users

```
1  function ResidenceFromtTexts(u)
2      l ⇐ ResidenceFromDescription(u)
3      if l = ∅ then
4          L ⇐ ∅
5          for each text in publications(u) do
6              norm ⇐ Normalise(text)
7              lang ⇐ IdentifyLanguage(norm)
8              E ⇐ NamedEntities(norm, lang)
9              for each ent in E do
10                 if isLocation(ent) then
11                     L ⇐ L ∪ {GeoCode(ent)}
12         l ⇐ MostFrequentLocation(L)
13     return l
```

Listing 4: Technique based on content

4. Once the language of the user's description has been identified, we perform an entity detection and classification process. As we use Freeling for this purpose, this step is restricted to the languages for which Freeling is able to perform named entity classification (i.e., English, Spanish, Galician and Portuguese). We enable Freeling's multi-word detection (line 6).

5. After that, we filter the named entities obtained in the previous step taking only the entities that correspond to a location. Such entities are sent one by one to the geocoding API for obtaining a set of normalised locations (lines 8-10).

6. As several locations may be obtained in the previous step, once the normalised locations have been obtained, we select only one location by following the same selection approach described in step 3 of the technique explained previously, returning one location as the place of residence of the user (line 11).

3.4. Technique based on content

This technique mines the content published (e.g., tweets and posts) and shared (e.g., retweets and links) by the users to obtain their place of residence. Listing 4 summarises the steps executed by this technique, which are described next.

1. Firstly, we attempt to obtain the place of residence by using the location metadata, as explained in Subsection 3.1. If a result is obtained, the process finishes with a location. Otherwise, the process continues in the following step.

2. Secondly, we use the user self-description as explained in Subsection 3.3. Is a result is obtained, the process finishes, otherwise, the process continues.

3. If the previous steps do not return a location, we obtain the textual contents published and shared by the user. We process each document obtaining a list of normalized locations mentioned in user's generated content. The process followed for obtaining the locations from the content is explained in Subsection 3.4.1.

4. Finally, we select the place of residence of the user from the list of locations obtained in the previous step, by applying the same location selection criteria used for the techniques previously described.

3.4.1. Extracting locations from content

For obtaining the locations from the textual content, we firstly identify the language of the post by applying the method explained in Subsection 3.3.

Secondly, if the content processed is a micro-post (i.e., content posted on Twitter), we perform a syntactic normalisation. This step converts the text of the tweet, that often includes metalanguage elements, to a syntax more similar to the usual natural language. Previous results demonstrate that this normalisation step improves the accuracy of the part-of-speech tagger (Codina and Atserias, 2012), of which the named entity classification module depends. Specifically, we have implemented several rules for syntactic normalization of Twitter messages. Some of these rules haven described in (Kaufmann and

Jugal, 2010). The rules executed by the content normaliser are the following:

1. Transform to lower-case the text completely written with upper-case characters;
2. Delete the sequence of characters "RT" followed by a mention to a Twitter user (marked by the symbol "@") and, optionally, by a colon punctuation mark;
3. Delete mentions to users that are not preceded by a coordinating or subordinating conjunction, a preposition, or a verb;
4. Delete the word "via" followed by a mention to a user at the end of the tweet;
5. Delete the hash-tags found at the end of the tweet;
6. Delete the "#" symbol from the hash-tags that are maintained;
7. Delete the hyper-links contained within the tweet;
8. Delete ellipses points that are at the end of the tweet, followed by a hyper-link;
9. Delete characters that are repeated more than twice (e.g., "maaaadrid" is converted to "madrid");
10. Transform underscores to blank spaces;
11. Divide camel-cased words in multiple words (e.g., "FutbolClubBarcelona" is converted to "Futbol Club Barcelona").

After normalising the text, we use Freeling to extract the locations, as described in Subsection 3.3. We have evaluated this step by using the training set published by the Concept Extraction Challenge of the #MSM2013 Workshop (MSM, 2013). Such training set consist of a corpus of 2.815 micro-posts written in English. The precision obtained is 0.52, while the recall is 0.43 (F_1 =0.47).

Finally we invoke the geocoding API for obtaining the normalized list of locations.

3.5. Hybrid technique

This technique combines the ones described previously, executing one after the other, ordered by computational complexity. Listing 5 summarises the steps executed by this technique, which are described next.

1. Firstly, we execute the technique based on content, which has been described previously (line 2).

```
1 function ResidenceHybrid(u)
2     l ⇐ ResidenceFromTexts(u)
3     if l = ∅ then
4         L ⇐ ∅
5         for each f in friends(u) do
6             L ⇐ L ∪ {ResidenceFromTexts(f)}
7         l ⇐ MostFrequentLocation(L)
8     return l
```

Listing 5: Hybrid technique

2. Finally, if the previous step does not return a place of residence, we make use of the friendship network of the user, by applying this hybrid technique to the list of friends of the users, and selecting the location as described in Subsection 3.2 (lines 3-7).

4. Gender Recognition

We have tested two techniques for gender recognition which are described next.

4.1. Technique based on user name metadata

This technique exploits publicly available metadata associated with the user profile. Such metadata may include the user name, as for example, the *name*, and *screen_name* Twitter attributes (Twitter, 2013).

The technique makes use of two lists of first names that have been previously classified by gender (one list for male names, and one list for female names). The lists have been curated, so unisex names have been excluded for classification purposes, given the ambiguity that they introduce. Specifically, we have generated the lists of first names from the information published by the Spanish National Institute of Statistics (INE, 2013). The initial list contains 18,697 first names (single and composite) for males and 19,817 first names for females. After the curation process (removing the first names that appear in both lists) the male first names list is reduced to 18,391 entries and the female names list to 19,511. Some examples of removed first names are Pau, Loreto and Reyes, as they are valid for either males and females in Spain.

Given a user account, its name metadata is scanned within the lists and, if a match is found in one of the lists, we propose the gender associated to the list where the first name has been found as the gender of the user. Our

technique not only takes the current value for the name metadata, but also the previous values for each attribute, as our data collection system stores historical data.

The proposed method is mostly language independent, being the only language-dependent resource the lists of first names. Those lists could be manually created from scratch, but there are plenty resources readily available, such as population censuses that can be used to build them.

4.2. Technique based on mentions to users

This techniques exploits the information provided by mentions to users. As for example, if someone post in Twitter *"I'm going to visit to my uncle @Daureos to Florida"*, she is providing explicit information about the gender of the user mentioned. We know that *@Daureos* is male because of the word "uncle" written before the user identifier.

We propose a technique for the Spanish language that performs a dependency parsing of the text with the aim of determining the gender of the terms related with the user mentioned. Therefore, for each tweet in which the user is mentioned, we attempt to estimate the gender of the user. Note that not all mentions to users provide information for estimating their genders (e.g., *"via @user"* and *"/cc @user"* at the end of the tweet). The dependency parser used is TXALA (Atserias, Comelles, and Mayor, 2005).

The steps executed by this technique are the following:

1. Firstly, we execute technique based on user name metadata described previously. If a gender is obtained, the process finishes.
2. If a gender is not identified in the previous step, we obtain all the tweets that mention the user.
3. For each tweet, we perform a dependency parsing. Once obtained the dependency tree, we assign a gender to the user for the tweet analysed according to the following heuristics: (1) if the gender of the term in the parent node, of the branch where the user is mentioned, is male or female, we consider that the user is male or female accordingly (e.g., *"Mi tío Daureos"*); (2) if some of the child nodes of the node corresponding to the user mention correspond to a term with a specific gender, we consider that the gender of the user correspond to the gender of such terms (e.g., *"Vio a Daureos enfermo y triste"*); (3) If there is a noun adjunct as the predicate of an attributive sentence where the user is the subject, we assign the gender of the noun adjunct as the gender of the user (e.g., *"Daureos es trabajador"*).
4. Finally, we select the gender that is associated the most to the tweets analysed for the user being analysed.

5. Evaluation

5.1. Place of residence recognition

We have evaluated the place of residence recognition techniques with an evaluation set of 1,080 users extracted from Twitter whose place of residence is known. Users in the evaluation set are distributed among 11 different countries (Argentina, Chile, Colombia, Spain, USA, Japan, Mexico, South Africa, Switzerland, Uruguay and Venezuela). Such users share and publish content in different languages (mainly in Spanish and English).

For evaluating the techniques that make use of the friendship networks, for practical reasons we have restricted the number of friends for each user to 20 (10 followers plus 10 persons followed by the user to be characterised), since Twitter limits the number of calls to its API. With respect to the techniques that make use of the content published and shared, we have restricted the number of tweets analysed to 20, for the same practical reasons, including tweets authored by the user and retweets.

All the techniques achieve a similar accuracy ($\approx 81\,\%$), with the exception of the technique based on friendship networks, which improves de accuracy to 86 %.

5.2. Gender recognition

To evaluate the techniques described we have considered an aleatory sample consisting on authors who have written a tweet in Spanish, as well as tweets that mention those authors between 29^{th} May 2012 and 27^{th} March 2013. The language of each tweet has been identified using LingPipe (Alias-i, 2008). The error of the language identification task causes the inclusion of authors in the evaluation corpus that might not be Spanish speakers, penalising the method recall.

	Predicted class		
Actual class	Male	Female	No gender
Male	530	42	49
Female	10	528	20
No gender	130	97	103

Table 1: Confusion matrix with the results of the technique based on mentions to users.

The evaluation set obtained for gender recognition contains 69,261 users. From these users, the technique based on profile metadata has been able to classify 46,030 users (9,284 female users and 36,746 male users), achieving a coverage of 66 % of the corpus. By contrast, the technique based on mentions to users has classified 46,396 users (9,386 female users and 37,010 male users), improving the coverage up to 67 %.

For evaluating the accuracy, we have annotated by hand the gender of 1,509 users (558 female users, 621 male users and 330 neutral users), and checked the automatic classification with respect to the manual annotation, obtaining an overall accuracy[1] of 0.9 for the technique based on user names, and of 0.84 for the technique based on mentions to users. By gender, for the technique based on user names, the precision obtained is 0.98 for male users and 0.97 for female users, while the recall is 0.8 and 0.87 respectively. For the technique based on mentions to users, the precision obtained is 0.8 for male users and 0.79 for female users, while the recall is 0.85 and 0.95 respectively. Therefore, the technique based on mentions to users achieves a smaller precision, but increases the recall with respect to the technique that only makes use of user names.

Table 1 shows the confusion matrix for the technique based on mentions to users. Users manually annotated as "no gender" correspond to non-personal Twitter accounts (e.g., a brand or a corporation), while those automatically classified as "no gender" are the users for which the algorithm was not able to identify a gender. Mainly, the confusions are produced between the male and female classes and the residual class.

In (Mislove et al., 2011) the authors propose techniques to compare Twitter population to the US population along three axes (geography, gender and race). Regarding the gender identification task, the method proposes a gender for 64.2 % of the authors (in our experiment this percentage is 66.45 %). In addition the 71.8 % of the users identified are males, while our experiment identifies the 79.8 %, obtaining similar distributions by gender.

[1] Accuracy $= \frac{tp+tn}{tp+tn+fp+fn}$

6. Conclusions

In this paper, we have described different techniques for obtaining the demographic attributes "place of residence" and "gender".

The evaluation results obtained for the techniques for identifying the place of residence of Twitter users show that the techniques that make use of the user's community achieve better performance than the techniques based on the analysis of the content published and shared by the user. While the major part of the community of a user uses to share the place of residence (because of the homophily principle in social networks), the mentions to locations included in the content published by the users are not related necessarily with their place of residence. Therefore, the hybrid technique does not perform better than the other techniques based on content.

We have achieved very satisfactory results for gender identification by just making use of user profile metadata, since the precision obtained is high and the technique used is very simple with respect to computational complexity, which leads to a straightforward set up in a production environment. The technique based on mentions to users increases the recall in the cases where the previous technique is not able to identify the gender of a given user, because for the Spanish language there exists grammatical agreement with respect to gender between nouns and other part-of-speech categories (e.g., adjectives and pronouns). However, such technique requires a language-depending dependency parser.

Future lines of work include experimenting with the detection of more demographic and psycho-graphic user characteristics which are relevant to the marketing and communication domains, including: age, political orientation and interests, among others.

7. Acknowledgement

This research is partially supported by the Spanish Centre for the Development of Industrial Technology under the CENIT program, project CEN-20101037, "Social Media" (`http://www.cenitsocialmedia.es`).

References

Alias-i. 2008. LingPipe 4.1.0. `http://alias-i.com/lingpipe`. [Online; accessed 8-April-2013].

Atserias, Jordi, Elisabet Comelles, and Aingeru Mayor. 2005. Txala: un analizador libre de dependencias para el castellano. *Procesamiento del Lenguaje Natural*, (35):455–456, September.

Backstrom, Lars, Jon Kleinberg, Ravi Kumar, and Jasmine Novak. 2008. Spatial variation in search engine queries. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 357–366, New York, NY, USA. ACM.

Burger, John D., John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1301–1309, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cheng, Zhiyuan, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geolocating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 759–768, New York, NY, USA. ACM.

Codina, Joan and Jordi Atserias. 2012. What is the text of a tweet? In *Proceedings of @NLP can u tag #user_generated_content?! via lrec-conf.org*, Istanbul, Turkey, May. ELRA.

Cortes, Corinna and Vladimir Vapnik. 1995. Support-vector networks. *Mach. Learn.*, 20(3):273–297, September.

Gayo-Avello, Daniel. 2011. Don't turn social media into another 'literary digest' poll. *Communications of the ACM*, 54(10):121–128, October.

Google. 2013. The Google Geocoding API. `https://developers.google.com/maps/documentation/geocoding/`. [Online; accessed 8-April-2013].

INE. 2013. INEbase: Operaciones estadísticas: clasificación por temas. `http://www.ine.es/inebmenu/indice.htm`. [Online; accessed 8-April-2013].

Kaufmann, Max and Kalita Jugal. 2010. Syntactic normalization of twitter messages. In *Proceedings of the International Conference on Natural Language Processing (ICON-2010)*.

McPherson, Miller, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444.

Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. 2011. Understanding the Demographics of Twitter Users. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, Barcelona, Spain, July.

MSM. 2013. Making Sense of Microposts (#MSM2013) – Concept Extraction Challenge. `http://oak.dcs.shef.ac.uk/msm2013/challenge.html`. [Online; accessed 8-April-2013].

Padró, Lluís and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Rao, Delip, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, SMUC '10, pages 37–44, New York, NY, USA. ACM.

Twitter. 2013. REST API v1.1 (GET users/show). `https://dev.twitter.com/docs/api/1.1/get/users/show`. [Online; accessed 8-April-2013].

Weber, Larry. 2007. *Marketing to the Social Web: How Digital Customer Communities Build Your Business*. Wiley, June.

wen Chang, Hau, Dongwon Lee, M. Eltaher, and Jeongkyu Lee. 2012. @phillies tweeting from philly? predicting twitter user locations with spatial word usage. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, pages 111–118.

Gramáticas y Formalismos para el Análisis Morfológico y Sintáctico

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 67-74 recibido 30-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Corrección no Supervisada de Dependencias Sintácticas de Aposición mediante Clases Semánticas*

Unsupervised Correction of Syntactic Dependencies of Apposition through Semantic Classes

Bernardo Cabaleiro, Anselmo Peñas
NLP & IR Research Group
ETSI Informática, UNED, Spain
{bcabaleiro,anselmo}@lsi.uned.es

Resumen: En este artículo exploramos de qué manera el conocimiento semántico adquirido de manera automática a partir de grandes colecciones de texto permite mejorar el resultado de un analizador sintáctico. Para ello proponemos un método no supervisado que corrige dependencias de aposición buscando en su contexto candidatos con mayor compatibilidad semántica que el proporcionado por el analizador sintáctico en primera instancia.
Palabras clave: Analizadores sintácticos, Adquisición de clases semánticas

Abstract: In this article we explore how automatic acquired semantic knowledge from large text collections allows to improve a syntactic parser. We propose an unsupervised method that corrects appositions dependencies by searching candidates in its context more semantically suitable than the one proposed by the syntactic parser in the first place.
Keywords: Syntactic parsers, Semantic classes acquisition

1. Introducción

El análisis de dependencias es un proceso clave para el procesamiento de lenguaje natural. Sin embargo, encontrar la estructura sintáctica correcta depende en muchas ocasiones de conocer las relaciones semánticas entre las palabras involucradas.

En este trabajo mostramos cómo un analizador sintáctico podría aprovechar información semántica extraída automáticamente para mejorar la identificación y corrección de errores en las dependencias de aposición.

Para ello presentamos un método no supervisado de corrección de aposiciones basado en clases semánticas. Consideramos como aposición aquellas construcciones gramaticales con dos sintagmas nominales adyacentes, donde uno define o modifica al otro(House, 2005). Cuando uno de los sintagmas nominales está gobernado por un sustantivo y el otro por una entidad nombrada, generalmente el sustantivo se utiliza para denotar una clase semántica de la entidad. Por ejemplo:

David meets his wife, ***Julia****.*[1]

Los analizadores sintácticos suelen determinar que existe una aposición entre el nombre común "wife" y la entidad nombrada "Julia", dando lugar a interpretar que "Julia" es de la clase "wife".

Cuando la entidad no admite el sustantivo como clase se puede inferir que la dependencia dada por el analizador podría ser incorrecta. Por ejemplo:

David supports the team of his wife, ***The Vikings****.*

Un analizador habitualmente determina una dependencia entre "wife" y "Vikings' por lo que se interpreta que Vikings es de la clase wife. Sin embargo, existe una incompatibilidad entre el tipo de la entidad nombrada y la clase, ya que "Vikings" es una organización y "wife" una clase característica de personas, lo que supone una evidencia de que tanto la dependencia como la interpretación que de ella se deriva son incorrectas. En cambio, en el contexto próximo se encuentra la

* This work has been partially supported by the Spanish Ministry of Science and Innovation, through the project Holopedia (TIN2010-21128-C02), and the Regional Government of Madrid, through the project MA2VICMR (S2009/TIC1542)

[1] En negrita la entidad nombrada, y subrayado los sustantivos candidatos a gobernante de la aposición.

ISSN 1135-5948

clase "team", que sí es válida para organizaciones.

Otro motivo de incompatibilidad es el género. Por ejemplo:

David meets a friend of his wife, ***Peter****.*

Aquí el analizador podría identificar una dependencia entre "wife" y "Peter". Saber que "wife" es una clase exclusivamente femenina y "Peter" una entidad nombrada masculina nos hace considerar más adecuada la clase "friend".

De acuerdo con el escenario descrito, nos formulamos las siguientes preguntas de investigación:

- ¿Puede detectarse en qué casos el analizador sintáctico no asigna correctamente el gobernante de una aposición debido a una incompatibilidad semántica?
- En caso de encontrar un error, ¿se puede determinar la estructura correcta y corregirlo?

Para responder a estas preguntas tratamos de identificar posibles errores buscando aposiciones con varios sustantivos candidatos a gobernar la relación. Después decidimos cuál de ellos es más apropiado utilizando como soporte el conocimiento antecedente sobre las clases.

Adquirimos el conocimiento antecedente a partir de un corpus de 1.5 millones de documentos, empleando patrones sintácticos muy sencillos para la identificación de clases semánticas.

El resto del artículo se organiza de la siguiente manera. Primero explicamos los dos pasos que sigue el sistema: adquisición de conocimiento antecedente (sección 2) e identificación y corrección de errores (sección 3). En la sección 4 mostramos la evaluación, hacemos un breve repaso a trabajos similares en la sección 5. Terminamos con las conclusiones y el trabajo futuro en la sección 6.

2. Adquisición de conocimiento antecedente

El reconocimiento de entidades es un problema muy estudiado en el procesamiento de lenguaje natural y los sistemas alcanzan buenos resultados (Finkel, Grenager, y Manning, 2005), especialmente para los tipos básicos (persona, localización y organización).

Sin embargo, conocer las clases semánticas asociadas a entidades (por ejemplo, una persona puede ser "bombero", "hijo", "dueño", etc) es un problema abierto que resulta interesante para múltiples tareas de NLP. El objetivo de esta fase es extraer relaciones entre clases e instancias con una probabilidad asociada.

Para adquirir esta información representamos en forma de grafo los documentos de una colección. Empleamos como base el paquete Stanford CoreNLP, que incluye uno de los analizadores sintácticos más populares (Klein y Manning, 2003), basado en una gramática libre de contexto probabilística. También utilizamos el etiquedator de partes de la oración (Toutanova y Manning, 2000), el reconocedor de entidades nombradas (Finkel, Grenager, y Manning, 2005) y el sistema de resolución de correferencia (Lee et al., 2011). Además, colapsamos los nodos en referentes de discurso. La descripción completa del método de obtención de la representación se puede ver en (Cabaleiro y Peñas, 2012).

Para obtener un conjunto de clases a partir de los grafos utilizamos una serie de patrones extremadamente sencillos basados en dependencias sintácticas «*governor, dependency, dependant*» en los que participa un sustantivo que será la clase y una entidad nombrada, que a su vez tiene un tipo de entidad asociado (persona, organización o localidad). La lista completa de patrones sintácticos está detallada en la tabla 1. En caso de coincidencia, asignamos el nombre común como clase semántica del nombre propio y obtenemos una instancia $EN - Clase - Tipo$.

Patrón Sintáctico		
EN	nn	NN
EN	appos	NN
EN	abbrev	NN
NN	appos	EN
NN	abbrev	EN
EN	such_as	NN
EN	like	NN

Tabla 1: Patrones para la asignación de clases semánticas. Cada entrada se corresponde con una tripleta «*governor, dependency, dependant*» donde NN es un nombre común y EN es una entidad nombrada.

Con estos patrones no se pretende obtener todas las relaciones clase-instancia expresadas en la colección, sino adquirir un número

representativo y suficiente de clases con las que evaluar los sustantivos candidatos.

Tras obtener las asignaciones de clases semánticas, agregamos la información de toda la colección para obtener las frecuencias de las coocurrencias entre entidades nombradas, clases y tipos. La Tabla 2 contiene las 5 clases más comunes asociadas al tipo de entidad *person*.

Clase	Tipo de EN	Frecuencia
spokesman	person	140229
president	person	102877
director	person	98182
leader	person	79716
coach	person	55511

Tabla 2: Clases más comunes asociadas al tipo de entidad *person*.

En la fase de identificación y corrección de aposiciones necesitaremos la probabilidad conjunta entre clases y entidades nombradas, así como entre clases y tipos de entidad nombrada (ver sección 3). Para obtenerlas utilizaremos un estimador de máxima verosimilitud:

$$p(c, en) = \sum_t p(en, c, t) \quad (1)$$

$$p(c, t) = \sum_{en} p(en, c, t) \quad (2)$$

Donde c es la clase, en la entidad nombrada y t el tipo de la entidad nombrada.

3. Identificación y corrección de errores en dependencias de aposición

En esta fase se consideran sustantivos candidatos a gobernar la relación de aposición. Para ello, primero identificamos dependencias formadas por dos sintagmas nominales que cumplan dos premisas: el primer sintagma nominal debe contener más de un sustantivo y el núcleo del segundo debe ser una entidad nombrada. Se considerarán candidatos todos los sustantivos contenidos en el primer sintagma nominal.

Utilizaremos dos evidencias distintas para determinar cuál de los sustantivos candidatos es más adecuado:

1. Considerar la información mutua normalizada entre la entidad nombrada y el sustantivo candidato $npmi(en; c)$.
2. Considerar la información mutua normalizada entre el tipo de la entidad nombrada y el sustantivo candidato $npmi(t; c)$.

La fórmula que describe la información mutua normalizada es:

$$npmi(x; y) = \frac{pmi(x; y)}{-log(p(x, y))} \quad (3)$$

Donde:

$$pmi(x; y) = log\frac{p(x, y)}{p(x)p(y)} \quad (4)$$

$p(x, y)$ es la probabilidad estimada en la fase anterior, y $p(x)$ y $p(y)$ el resultado de marginalizar. Utilizar la información mutua puntual normalizada sitúa los resultados en un rango de $(-1, 1)$, siendo -1 el valor cuando no se ha observado ninguna ocurrencia conjunta y 1 el valor cuando siempre se observan conjuntamente.

4. Evaluación

En esta sección describiremos los datos utilizados en la experimentación, presentaremos los resultados de la misma y haremos un análisis de los diferentes casos encontrados.

4.1. Datos

Para esta tarea se ha utilizado la colección perteneciente a la tarea TAC KBP (Ji, Grishman, y Dang, 2011), compuesta por alrededor de 1.5 millones de documentos pertenecientes a diferentes categorías, incluyendo noticias, blogs y transcripciones de conversaciones telefónicas.

Este corpus se utiliza para generar una colección con todas las oraciones que contienen una asignación de dependencia de aposición candidata a ser corregida (ver sección 3). Encontramos un total de 284845 oraciones, para las cuales alguna de las dos evidencias sugiere otra alternativa en el 47,7 % de las ocasiones.

Para evaluar el rendimiento del sistema tomamos una muestra de 300 ejemplos que etiquetamos manualmente indicando si los gobernantes escogidos por el analizador sintáctico y por las evidencias son correctos o no.

4.2. Definición del experimento

Utilizamos como línea base los resultados que obtiene el analizador sintáctico en primera instancia. Para poder comparar la aportación

a la corrección según las diferentes configuraciones del método realizaremos las siguientes pruebas:

Configuración 1: En la primera prueba medimos cuantas dependencias son correctas si se escoge el resultado de aplicar la evidencia 1 ($npmi(ne; c)$). Así comprobaremos en qué medida la entidad nombrada es indicativa de la clase.

Configuración 2: La segunda prueba medirá el acierto del sistema si se escoge el resultado de aplicar la evidencia 2 ($npmi(t; c)$). En esta prueba se comprobará en qué medida el tipo de la entidad nombrada es indicativo de la clase.

Configuración 3: En esta prueba mostramos cuantas dependencias serían correctas si combinamos las configuraciones 1 y 2 escogiendo siempre el resultado correcto: El objetivo de esta prueba es comprobar si existe la posibilidad de agregar la evidencia de las entidades nombradas y sus tipos para mejorar los resultados.

4.3. Resultados

En los 300 ejemplos encontramos 84 casos (28 %) que descartamos porque no se tratan de verdaderas aposiciones (ver sección 4.5). Trabajaremos por tanto con un conjunto de 216 ejemplos, en 150 (69,4 %) de los cuales ninguna de las configuraciones propone un sustantivo diferente al dado inicialmente por el analizador, mientras que en 66 (30,4 %) sí se propone alguna alternativa.

La línea base que hemos definido es el número de dependencias de aposición correctas del analizador sintáctico. En esta muestra la tasa de acierto es de 84.2 %, lo que supone 34 ejemplos clasificados incorrectamente.

En la configuración 1 probamos la utilidad de las entidades nombradas. Nos hemos encontrado el problema de que algunas entidades nombradas no coocurren ninguna vez con sus clases candidatas. En este caso, hemos supuesto que no se corregía la dependencia original por lo que el resultado es correcto si originalmente ya lo era. Podemos comprobar que con esta estrategia se empeoran los resultados un 0,9 % con respecto a la prueba base.

En la configuración 2 probamos la utilidad de los tipos de entidad nombrada. Con esta estrategia mejoramos ligeramente (1 %) los resultados de la prueba base.

En la configuración 3 mostramos cuantas dependencias serían correctas si combinamos las configuraciones 1 y 2 de manera óptima. Con este sistema ideal mejoraríamos el sistema base en un 9,8 %. Este resultado motiva el desarrollo de un modelo conjunto.

En la Tabla 3 mostramos los resultados de cada configuración. Las filas se corresponden con el número de dependencias de aposición en las que se mantiene el sustantivo original (no corregidas) y el número para el que se propone un candidato alternativo (corregidas). En la configuración 1 también se muestran los casos donde no se han encontrado coocurrencias entre la entidad nombrada y ninguna de las clases (Sin información). Consideramos correctas las aposiciones que coinciden con las etiquetas manuales.

La tabla 4 muestra ejemplos de oraciones correctamente corregidas. En el primer ejemplo, los dos métodos propuestos escogen el gobernador mejor que el analizador sintáctico. En el segundo no se encuentran coocurrencias en el conocimiento antecedente entre la entidad y las clases, la información que aporta el tipo de la entidad sirve para corregir la dependencia.

En la tabla 5 podemos ver ejemplos en los que alguno de los métodos ha fallado. En el primer ejemplo el tipo de la entidad reconocido es incorrecto, lo que induce a error a la configuración 2. Es interesante que las puntuaciones de todos los candidatos son negativas, lo que nos hace pensar que detectando casos en los que las clases candidatas no son compatibles con el tipo de la entidad podemos corregir errores en la clasificación de entidades nombradas (ver sección 6)

En el segundo ejemplo la evaluación de los sustantivos según la configuración 2 falla. En oraciones así, considerar el género gramatical ayudaría a mejorar el resultado.

4.4. Casuística

En este apartado analizamos los diferentes casos encontrados que afectan a las dependencias de aposición: sintagmas nominales con múltiples sustantivos, sintagmas nominales compuestos por conjunciones, sintagmas nominales ambiguos semánticamente y casos de no aposición.

4.4.1. Sintagmas nominales con múltiples sustantivos

Generalmente, el núcleo del primer sintagma nominal debería ser el gobernante de la aposición, sin embargo, determinarlo no es un

		Correctas	Incorrectas	Total
Línea Base	Total	186 (84,2 %)	34 (15,8 %)	216
Configuración 1 $npmi(en; c)$	No corregidas	118 (54,6 %)	7 (3,2 %)	125 (57,9 %)
	No corregidas (Sin información)	46 (21,3 %)	11 (5,1 %)	57 (26,4 %)
	Corregidas	16 (7,4 %)	18 (8.3 %)	34 (15,7 %)
	Total	180 (83.3 %)	36 (16,7 %)	216
Configuración 2 $npmi(t; c)$	No corregidas	150 (69,4 %)	0 (0 %)	150 (69,4 %)
	Corregidas	34 (15,8 %)	32 (14,8 %)	66 (30,6 %)
	Total	184 (85,2 %)	32 (14,8 %)	216
Configuración 3 $npmi(en; c)$ & $npmi(t; c)$	No corregidas	169 (78,2 %)	0 (0 %)	169 (78,2 %)
	Corregidas	34 (15,8 %)	13 (6 %)	47 (21,8 %)
	Total	203 (94 %)	13 (6 %)	216

Tabla 3: Línea base y resultados del sistema para las tres configuraciones.

Oración	Tipo de EN	Inicial			Alternativa		
		Clase	Conf 1	Conf 2	Clase	Conf 1	Conf 2
... the chief of the Shin Bet security agency , **Yuval Diskin**, ...	person	agency	0.13	-0.25	chief	**0.28**	**0.09**
... the head of the Pellervo economic research institute, **Paula Horne**	person	institute	-1	-0.13	head	-1	**0.08**

Tabla 4: Aposiciones correctamente corregidas.

problema trivial. En ocasiones contiene varios sustantivos que son candidatos válidos a ser el gobernante de la relación.

En algunos casos existen compuestos nominales cuyos sustantivos tienen la misma importancia, como en los siguientes ejemplos:

... used by its domestic subsidiary airline, ***Eagle Air****, ...*

... by the IOC 's chief Beijing organiser, ***Hein Verbruggen****, ...*

En otras ocasiones el sintagma nominal incluye una oración subordinada, que también tiene un sustantivo apto para gobernar la aposición, como en el siguiente ejemplo:

Another passenger who gave only his surmane, ***Chen*** *...*

Además también hay casos en los que el núcleo es menos discriminativo como clase semántica que el otro sustantivo, por ejemplo:

Henry is grounded by his illustrator partner, **Rudy** ...

Tanto escoger el gobernador de la dependencia de aposición como asignar la clase a la entidad nombrada depende del objetivo final del sistema. En este trabajo hemos decidido que a efectos de evaluación el último sustantivo válido del sintagma será el gobernante correcto de la relación de aposición.

4.4.2. Sintagmas nominales compuestos por conjunciones

Este caso se produce cuando el primer sintagma nominal contiene una o varias conjunciones, con lo que hay dos sustantivos válidos para gobernar la relación de aposición. Por ejemplo:

... a prominent Jewish writer and Holocaust survivor, ***Ralph Giordano*** *...*

En este caso consideramos que la solución ideal sería mantener la aposición en el sustantivo en el que se haya encontrado, y añadir una aposición nueva para los demás sustantivos candidatos. A efectos de puntuación del sistema, hemos considerado que escoger cualquiera de los dos sustantivos es una elección correcta. Hemos encontrado frases con esta casuística en cuatro ocasiones (1,85 %).

4.4.3. Sintagmas nominales ambiguos semánticamente

En algunas oraciones existen dos clases referidas a dos entidades diferentes, pero que pre-

Oración	Tipo de EN	Inicial			Alternativa		
		Clase	Conf 1	Conf 2	Clase	Conf 1	Conf 2
The bronze medalist, China's **Ma Yuxi**, set ...	organization (erróneo)	medalist	-1	-0.23	bronze	-1	**-0.07**
... leaving with his father and his mother, **Linda**.	person	mother	**0.2**	0.06	father	0.01	**0.07**

Tabla 5: Aposiciones mal corregidas.

cisan de conocimiento extra-lingüístico para decidir cómo están relacionadas. Por ejemplo:

... at least one brother of another defendant, ***Ali Dayeh****.*

En el ejemplo anterior existen dos clases, "brother" y "defendant", que se refieren a dos entidades distintas, "Ali Dayeh Ali" y una entidad desconocida. Sin conocimiento externo no puede decidirse si Ali Dayeh es de clase "brother" o de clase "defendant".

4.5. Casos de no aposición

En la inspección manual hemos encontrado múltiples ejemplos de situaciones en las que el sistema detecta una relación de aposición errónea. Corregir estos errores es por sí mismo una línea de trabajo para mejorar el análisis sintáctico, pero en este trabajo nos limitaremos a indicar qué casos hemos identificado.

- Conjunción entre dos oraciones. Por ejemplo:

  Guo Wenjun won the women's 10-meter air pistol, ***Guo Jingjing*** *and Wu Minxia the women's synchronized 3-meter springboard, ...*

- Relación incorrecta de aposición entre un sintagma nominal y un sintagma verbal. En muchos casos, el análisis correcto sería considerar el sintagma nominal como objeto del sintagma verbal. Por ejemplo:

  ... will serve as the incoming president's chief of staff, President-elect ***Ma Ying-jeou****'s office announced.*

- Estructuras que denotan la relación localidad-región. Por ejemplo:

  ... ESA mission control in Toulouse, southwestern ***France****, ...*

- Enumeraciones. Por ejemplo:

  A spa tucked away in the basement includes a large pool, a ***whirlpool****, workout room, saunas, a solarium ...*

- Texto sin sentido. Ejemplo:

  ... Anti-Muslim Bigots , V-for-Vendicar, fruitella, Zionism equal Racism, ***The Chemical Oil Nazi****, LORD RAMA RANTER , Muslim With Mission, ...*

5. Trabajo relacionado

La relación que existe entre la ambigüedad sintáctica y el análisis semántico de un texto, así como sus aplicaciones a la desambiguación sintáctica es un tema de estudio en PLN desde hace años (Church y Patil, 1982; Resnik, 1993).

En (Ciaramita, 2007) se experimenta si añadir rasgos semánticos en un analizador sintáctico pueden mejorar su rendimiento. Para ello etiquetan las entidades nombradas de un corpus mediante un reconocedor de entidades y tratan las etiquetas semánticas como partes de la oración. En (Agirre, Baldwin, y Martinez, 2008; Agirre et al., 2011) el foco está en utilizar WordNet para generalizar palabras relacionadas, como, por ejemplo, tomar la clase "herramienta" en vez de las instancias "tijeras" o "cuchillo", para mejorar analizadores sintácticos.

Más en la línea de nuestro trabajo, (Clark y Harrison, 2009) tratan de explotar la redundancia de instancias adquiridas mediante patrones en un corpus previo. Nosotros además combinamos clases con tipos de entidades nombradas. Así encontramos evidencia de que, por ejemplo, las clases "spokesman" y "leader" se relacionan con más frecuencia con entidades de tipo "person", y "group" y "company" a "organization".

Existen múltiples ontologías o diccionarios con información sobre clases semánticas que podríamos utilizar como conocimiento antecedente, ya sea creados a mano como WordNet (Miller et al., 1990) o de manera semisupervisada como DBPedia (Mendes, Jakob, y

Bizer, 2012), pero su cobertura es insuficiente, especialmente para clases abiertas, y no incorporan fácilmente conocimiento nuevo.

Para solucionar este problema se han propuesto múltiples métodos no supervisados de adquisición de clases semánticas, lo que se conoce entre otros nombres "semantic-class learning" o "semantic class induction" (Lin y Pantel, 2001). Una técnica común es procesar los textos mediante un analizador morfosintáctico y seleccionar uno o varios patrones superficiales para extraer un conjunto de clases semánticas, y posteriormente refinar los resultados obtenidos (Hearst, 1992; Kozareva, Riloff, y Hovy, 2008).

En nuestro caso, hemos optado por no restringir el número de clases, escogiendo varios patrones comunes y aplicándolos a un corpus de millones de documentos. De esta manera pretendemos aprovechar la idea de (Schubert, 2002), que sostiene que los textos contienen conocimiento general en forma de aserciones, que puede ser explotado mediante el procesamiento de grandes cantidades de texto, como en KNEXT (Schubert, 2002) o DART (Clark y Harrison, 2009). De esta manera, nos centramos más en encontrar redundancias que en perfeccionar el análisis intra-documento.

6. Conclusiones y trabajo futuro

En este artículo hemos estudiado si considerar clases semánticas podría mejorar el análisis de dependencias de aposición. Para ello evaluamos la compatibilidad semántica de los sintagmas nominales que participan en la dependencia.

Para caracterizar los sintagmas nominales utilizamos conocimiento antecedente de clases semánticas adquirido automáticamente. Este conocimiento está dividido en dos evidencias distintas, una que relaciona clases semánticas con entidades nombradas y otra que relaciona clases semánticas con tipos de entidad nombrada.

El conocimiento antecedente que relaciona las clases con las entidades no mejora los resultados, ya que la dispersión de las instancias clase-entidad hace que la probabilidad estimada de coocurrencia no sea robusta. Esto podría mejorarse sustituyendo los estimadores de máxima verosimilitud por otros como modelos basados en entropía condicional o en información mutua.

Es destacable que en el conjunto de test utilizado no hay ningún ejemplo que sea respondido correctamente únicamente con la evidencia clase-entidad ya que o bien coincide con la evidencia del tipo de la entidad o con la respuesta del analizador sintáctico en primera instancia.

Al considerar la información aportada por los tipos de las entidades nombradas el problema es que algunas clases son muy dominantes ("chief", "business") y tienden a sobreasignarse.

La combinación óptima de la información de las entidades nombradas y de sus tipos mejora notablemente los resultados, pasando de un 84 % a un 94 % de dependencias correctas. El siguiente paso es por tanto generar un modelo más robusto que considere conjuntamente esta información, e incluso otra adicional, como el género gramatical de las entidades nombradas.

Este modelo tendría varias utilidades: (1) mejorar la corrección de aposiciones, (2) mejorar la propia asignación de clases, utilizando aposiciones corregidas para la adquisición de conocimiento antecedente y (3) mejorar el reconocimiento de entidades nombradas, identificando casos en los que las clases candidatas no sean adecuadas al tipo de la entidad nombrada.

Este trabajo está centrado en las dependencias de aposición, pero es interesante estudiar si se puede extrapolar esta técnica para resolver otros tipos, como abreviaturas o verbos copulativos, o incluso resolución de correferencia.

Bibliografía

Agirre, Eneko, Timothy Baldwin, y David Martinez. 2008. Improving parsing and PP attachment performance with sense information. En *Proceedings of ACL-08: HLT*, páginas 317–325, Columbus, Ohio, June. Association for Computational Linguistics.

Agirre, Eneko, Kepa Bengoetxea, Koldo Gojenola, y Joakim Nivre. 2011. Improving dependency parsing with semantic classes. En *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, páginas 699–703, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cabaleiro, Bernardo y Anselmo Peñas. 2012.

Representación gráfica de documentos para extracción automática de relaciones. *Procesamiento del Lenguaje Natural*, 49(0).

Church, Kenneth y Ramesh Patil. 1982. Coping with syntactic ambiguity or how to put the block in the box on the table. *Comput. Linguist.*, 8(3-4):139–149, Julio.

Ciaramita, Massimiliano. 2007. Dependency parsing with second-order feature maps and annotated semantic information. En *Proc. of the 12th International Workshop on Parsing Technologies (IWPT)*.

Clark, Peter y Phil Harrison. 2009. Large-scale extraction and use of knowledge from text. En *Proceedings of the fifth international conference on Knowledge capture*, K-CAP '09, páginas 153–160, New York, NY, USA. ACM.

Finkel, Jenny Rose, Trond Grenager, y Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. ACL '05, páginas 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hearst, Marti A. 1992. Automatic acquisition of hyponyms from large text corpora. En *Proceedings of the 14th conference on Computational linguistics - Volume 2*, COLING '92, páginas 539–545, Stroudsburg, PA, USA. Association for Computational Linguistics.

House, Random. 2005. *Random House Kernerman Webster's College Dictionary*.

Ji, Heng, Ralph Grishman, y Hoa Dang. 2011. Overview of the TAC2011 Knowledge Base Population Track. En *TAC 2011 Proceedings Papers*.

Klein, Dan y Christopher D. Manning. 2003. Accurate unlexicalized parsing. En *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, páginas 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kozareva, Zornitsa, Ellen Riloff, y Eduard Hovy. 2008. Semantic class learning from the web with hyponym pattern linkage graphs. En *Proceedings of ACL-08: HLT*, páginas 1048–1056, Columbus, Ohio, June. Association for Computational Linguistics.

Lee, Heeyoung, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, y Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. CONLL Shared Task '11, páginas 28–34, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lin, Dekang y Patrick Pantel. 2001. Induction of semantic classes from natural language text. En *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, páginas 317–322, New York, NY, USA. ACM.

Mendes, Pablo N., Max Jakob, y Christian Bizer. 2012. Dbpedia for nlp: A multilingual cross-domain knowledge base. En *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, Mayo.

Miller, George A, Richard Beckwith, Christiane Fellbaum, Derek Gross, y Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database*. *International journal of lexicography*, 3(4):235–244.

Resnik, Philip. 1993. Semantic classes and syntactic ambiguity. En *Proceedings of the workshop on Human Language Technology*, HLT '93, páginas 278–283, Stroudsburg, PA, USA. Association for Computational Linguistics.

Schubert, Lenhart. 2002. Can we derive general world knowledge from texts? En *IN PROC. HLT 2002*, páginas 24–27.

Toutanova, Kristina y Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. EMNLP '00, páginas 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 75-82 recibido 30-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Reglas de formación de palabras compuestas en español para la automatización de su reconocimiento

Formation rules of compound words in Spanish to automate recognition

Octavio Santana Suárez, José Pérez Aguiar
Universidad de Las Palmas de Gran Canaria
Edificio Departamental de Informática y Matemáticas
Campus Universitario de Tafira
35017 Las Palmas de Gran Canaria
{osantana, jperez}@dis.ulpgc.es

Virginia Gutiérrez Rodríguez, Isabel Sánchez Berriel
Universidad de La Laguna
Edificio de Física y Matemáticas
Campus Universitario Anchieta
C/Astrofísico Francisco Sánchez s/n
38271 La Laguna
{vgutier, isanchez}@ull.es

Resumen: En el presente trabajo se recogen las reglas de formación y los criterios de aplicación que se deberían llevar a cabo en cada situación para permitir la identificación automatizada de uno de los procesos de formación de palabras que tiene el español: la composición, se estudian sólo aquellos casos en los que se ha producido dicho fenómeno mediante la unión gráfica de los elementos que participan, con miras a su automatización. A tal fin, se extraen de diferentes diccionarios los distintos compuestos con el propósito de garantizar un conocimiento suficiente de los diferentes casos que se pueden prever de la materia y se estudian los mecanismos de unión aplicados según la categoría gramatical del compuesto y las de sus elementos componentes.
Palabras clave: Composición, reglas de formación, neologismos, procesamiento del lenguaje natural, lingüística computacional

Abstract: The following work presents the formation rules and application criteria that should be carried out in each situation in order to obtain the automated identification of one of the word's formation processes in Spanish. In the composition are studied only those cases in which this phenomenon has been produced through the graphic link of the elements involved, aimed to automation. At this aim, different compounds are extracted from dictionaries with the only purpose of guarantee a basic knowledge of the cases that are provided by the materia. The linking mechanisms are also studied depending on the grammatical category of the compounds and the elements which form part of it.
Keywords: Composition, formation rules, neologisms, natural language processing, computational linguistic

1 Introducción

Uno de los medios con los que cuenta la lengua española para ampliar el conjunto de voces del idioma es utilizar mecanismos de tipo morfológico —también llamados neologismos morfosintácticos— para formar nuevas palabras como son la composición, la derivación o la parasíntesis, entre otros. Sin duda, la composición es uno de los procesos de formación de palabras con mayor importancia que dispone la lengua para la renovación y enriquecimiento de su léxico.

En estos mecanismos morfológicos se parte de elementos ya presentes en el lenguaje o de otros tomados de fuera para crear nuevos vocablos; en la composición mediante la unión de dos o más de estos elementos, en contraposición a la derivación donde existe un elemento gramatical que no está libre, es decir, consiste en la creación de elementos léxicos nuevos mediante la adición o supresión a palabras ya existentes de elementos inseparables ―afijos. Por último, en la parasíntesis se combinan los mecanismos anteriores, bien por afijación que simultanea

ISSN 1135-5948

dos procedimientos derivativos —sufijación y prefijación— o por combinación de elementos de la composición y de la derivación como trata Serrano Dolader (1995). En el presente trabajo se estudian aquellos casos de compuestos que se han consolidado como la unión gráfica de los elementos que intervienen[1], además se tratan los pseudoprefijoides o pseudosufijoides como elementos compositivos no como morfemas derivativos, y se incluye la parasíntesis por composición, desde un punto de vista morfológico. Se trata, en suma, de procedimientos que pueden crear neologismos, lo que constituye una alternativa moderna para el enriquecimiento de la lengua.

2 La composición en español

En los estudios realizados sobre composición en español, no se ha llegado a dar una definición de forma unánime, sobre todo a la hora de establecer la frontera entre composición y derivación. Para M. Lang (1997) la composición consiste en la unión de palabras ya se trate de formas libres o morfemas léxicos; en la misma línea puede resultar la definición de M. Alvar (2002) donde en la composición participan dos o más unidades léxicas que pueden aparecer libres en la lengua; pero la más próxima a nuestro estudio es la definida por el Diccionario de la Real Academia de Lengua Español —DRAE— "*procedimiento por el cual se forman vocablos agregando a uno simple una o más preposiciones o partículas u otro vocablo íntegro o modificado por eufonía" —coyotomate, quitaipón, cagalaolla, paraguas.* Se considera que los elementos componentes que forman una palabra compuesta pueden ser de dos tipos: palabras[2] "castellanas enteras" (P) y temas cultos de origen grecolatino llamados elementos compositivos prefijales o sufijales (EC). Al fijar que los elementos compositivos son "temas cultos" se establece la diferencia con lo que se entiende por verdaderos afijos, tanto sean prefijos como sufijos[3]. La Tabla 1 muestra ejemplos de palabras compuestas formadas por las combinaciones posibles de estos dos tipos de unidades léxicas.

Composición	Palabra	Elemento compositivo
Palabra	*hojalata* *malgastar uñalbo*	*germanófilo* *timbrología*
Elemento compositivo	*ecosistema* *cardioprotector*	*logopeda* *filántropo*

Tabla 1. Combinaciones de distintas unidades léxicas para formar palabras compuestas

La utilización de raíces cultas greco-latinas es frecuente en los procesos de formación de nuevas palabras, especialmente en los campos científicos y técnicos. Las voces en cuya formación intervienen estos elementos podrían no considerarse propiamente compuestas, pues la mayoría de estas raíces no pueden aparecer de forma aislada, pero tampoco pueden considerarse derivadas puesto que tienen un comportamiento peculiar que los aleja de los auténticos afijos; es más, tienen un significado léxico más próximo al de las raíces que al de los afijos. A este tipo de raíces se les da el nombre de *elementos prefijales* o *sufijales —elementos compositivos—*, en función de si se anteponen o se posponen a otra raíz.

Hay que tener en cuenta que aunque, por norma general —incluido en muchas definiciones de composición— el número de constituyentes que forma una palabra compuesta es de dos, existen casos de tres —*tiraiafloja, bienteveo, cenaaoscuras*—, o incluso cuatro componentes —*correveidile, ahimelollevas, tentenelaire*—, aunque, en la mayoría de ellas, suelen insertarse elementos monosilábicos átonos como preposiciones, conjunciones copulativas, pronombres, artículos, entre otros. Las palabras compuestas han sufrido un proceso de aglutinación sin pérdida de dicho elemento del sintagma nominal u oración original del que proviene.

En cuanto a las combinaciones de palabras (P+P), estas pueden crearse a partir de

[1] Hay que tener en cuenta que la composición puede dividirse en composición de tipo ortográficamente unidos y composición de tipo sintagmático, donde los componentes han alcanzado una coherencia semántica sin fusión ortográfica. El presente trabajo sólo se centrará en el primer tipo de composición.

[2] Entendiéndose como palabras aquellas que tienen un solo lexema o un lexema unido a morfemas flexivos o las que están formadas por un morfema libre o unido a morfemas flexivos.

[3] Un reconocimiento automatizado de compuestos formados por elementos compositivos tiene que considerar la distinción entre elementos compositivos y afijos: un elemento prefijal no equivale a la derivación por prefijación de una palabra (*bienvenir* es una palabra compuesta, no derivada del verbo *venir* con los prefijos *bi-en-* sino proviene del elemento compositivo *bien-*).

conceptos que estén relacionados, o mejor dicho, categorías gramaticales como sustantivos, adjetivos y verbos; por tanto, la estructura de los compuestos puede presentar múltiples formas, según la categoría gramatical de los componentes y la categoría del resultado final. Con la variedad de formantes y de resultados que presentan, las relaciones entre los elementos participantes son de muy diversa índole. La composición más común es aquella que sirve para denominar objetos como la composición nominal, seguida de la composición adjetiva y, en menor medida, la composición verbal. En función de la clase funcional de los elementos componentes se obtienen los siguientes esquemas de combinación más típicos:

Sustantivo+Sustantivo
Verbo+Complemento
Sustantivo+Adjetivo
Adjetivo+Sustantivo
Sustantivo+Preposición+Sustantivo

donde E. Bustos (1986) plantea una clasificación según las categorías gramaticales de los elementos componentes, así como la categoría del compuesto como se indica en la Tabla 2.

Se entiende por composición *sintagmática* aquel tipo de composición que tiene como resultado unos compuestos que se aproximan bastante a los sintagmas nominales correspondientes, pero que no pueden ser considerados como sintagmas nominales[4], sino como unidades léxicas pues el compuesto se utiliza como una única palabra en la que han quedado cohesionados todos sus componentes; donde la composición *adjetiva* y la *propia* sustantiva —también conocida como *composición léxica*— está formada por dos o más palabras o bases con alguna modificación fónica, generalmente con unión gráfica de los elementos que la componen. Desde un punto de vista histórico, no existe una separación tajante entre los compuestos sintagmáticos y determinados ejemplos de compuestos propios, por cuanto que estos, en ciertos casos, pueden ser el resultado de un proceso de aglutinación —*hoja de lata* → *hojalata, hidalgo o aguanafa*. Desde un punto de vista sincrónico, el no reconocimiento de grupos sintácticos en alguno de estos tipos de compuestos es un problema de tradición lingüística o de realidad de los hechos del lenguaje. Se dejan al margen la mayoría de los compuestos de tipo sintagmático, ya que los componentes han alcanzado una coherencia semántica pero sin fusión gráfica y por lo tanto no se van a considerar en el presente estudio.

Composición Adjetiva	Adverbio+Adjetivo Adjetivo+Adjetivo Sustantivo+Adjetivo	
Composición Nominal	**Composición Propia**	Sustantivo+Sustantivo Verbo+Verbo Verbo+Complemento
	Composición Sintagmática	Sustantivo+Adjetivo o Adjetivo+Sustantivo Sustantivo+Preposición +Sustantivo

Tabla 2. Clasificación de la composición según E. Bustos Gisbert

Según la frecuencia de aparición de la categoría gramatical del compuesto, se pueden detallar las siguientes combinaciones:

- Uno de los compuestos más productivos son los que dan como resultado un *sustantivo*. Dentro de este tipo, los más caudalosos son los de *Verbo+Complemento* debido a la constitución interna que presentan así como a la comodidad que le produce a un hablante esta estructura por la semántica que sus compuestos implican, ya que caracterizan al referente a través de su actividad, como indica E. Bustos (1986). En la Tabla 3 se aprecian diferentes formaciones de compuestos nominales.

  Existen compuestos que representan sistemas marginales o bien resulta escaso el número de casos que producen o son casos particulares de los anteriores:

  PREPOSICIÓN+CONJUNCIÓN CONDICIONAL +ADVERBIO DE MODO: *porsiacaso.*
  NUMERAL+PREPOSICIÓN+SUSTANTIVO: *cientoemboca, milenrama.*
  VERBO+CONJUNCIÓN: *creique, penseque.*

  Los compuestos que constan de más de dos piezas léxicas dan como resultado un sustantivo formado por las combinaciones, entre otras:
  VERBO+Y+VERBO: *tiraiafloja, quitaipón*.
  VERBO+VERBO+Y+VERBO: *correveidile*.
  ADVERBIO+PRON.PERSONAL+VERBO: *bienmesabe*.
  VERBO+PRONOMBRE PERSONAL+ PRON. IMPERSONAL: *sabelotodo*.
  VERBO+PRON.PERSONAL+VERBO: *hazmerreír*.
  VERBO+PRON.PERSONAL+PREPOSICIÓN+ PRON.IMPERSONAL: *metomentodo*.

[4] No pueden adquirir la categoría de compuesto aunque semánticamente estén unidos como *farola de jardín* o *casa hospital* frente a palabras que sí lo son como *libro de cocina* o *restaurante espectáculo*.

Categoría	Sustantivo	Adjetivo
Sustantivo	*casatienda bocamanga carricoche telaraña*	*aguardiente hierbabuena artimaña pelirrojo*
Adjetivo	*ciempiés malasangre mediodía malpaís*	
Verbo	*hincapié matamoscas quemasangres saltarrostro quitapenas trotamundos*	
Adverbio	*bienandanza menoscuenta bienquerencia*	

Categoría	Verbo	Adverbio
Verbo	*compraventa duermevela ganapierde*	*bogavante atalejo mandamás*
Adverbio	*bienmesabe*	

Tabla 3. Compuestos nominales según la categoría gramatical de sus elementos componentes

- El segundo caso en productividad son los compuestos *adjetivos***.** Los más caudalosos son los de *Sustantivo+ Adjetivo* donde existe una marcada relación entre los elementos del compuesto ya que el segundo miembro predica una cualidad del primero, sustantivo[5] que generalmente designa partes exteriores del cuerpo humano o animal.
- Componentes que parecen admitir dos interpretaciones diferentes —*adjetivos y sustantivos*—, y no se posee información suficiente para decidir cual resulta más correcta —*cardocuco "cardo silvestre"* no está claro si *cuco* es adjetivo o sustantivo.

Por regla general, el resultado de fusionar dos palabras para formar un compuesto es un sustantivo o un adjetivo. Pero existen, aunque en menor medida, diversas categorías gramaticales adicionales:

Cuando dan lugar a *verbo*. No se ha detectado que esta categoría gramatical combine dos constituyentes de su misma categoría gramatical, es decir, *Verbo+Verbo*, constatándose esta misma restricción en otras lenguas.
SUSTANTIVO+VERBO: *maniatar,aliquebrar*.
ADVERBIO+VERBO: *maldecir, malcomer, bienpensar*.

- Que dan lugar a *adverbio*
  CONJUNCIÓN+VERBO: siquiera (adverbio o conjunción).
  ADVERBIO DE CANTIDAD+ADJETIVO O ADVERBIO: *tampoco*.
  ELEMENTO COMPOSITIVO+ADVERBIO DE TIEMPO: *anteayer*.
  ADJETIVO+ADVERBIO DE MODO: *otrosi*.
  Raros: ADVERBIO+VERBO: *dondequiera*.
- Que dan lugar a *pronombres*:
  PRONOMBRE RELATIVO+VERBO: *cualquiera, quienquiera*.
  PRON.PERSONAL+ADJETIVO: *nosotros*.
- Que dan lugar a *conjunciones*:
  ADVERBIO DE TIEMPO O DE MODO + PRONOMBRE RELATIVO: *aunque*.
  PREFIJO+SUSTANTIVO U.T.C. CONJUNCIÓN ADVERSATIVA: *sin embargo*.
  CONJUNCIÓN+ADVERBIO DE MODO: *sino*.
- Que dan lugar a *numerales*:
  veinticinco→veinte+cinco, dieciséis→diez+seis

Categoría	Adjetivo (o Participio)
Sustantivo	*alicaído, cejijunto, pelirrojo, teticiega, patidifuso*
Adjetivo	*tonticiego, grandilocuente, agridulce, hispanohablante*
Adverbio	*bienintencionado, malaconsejado*

Tabla 4.- Compuestos adjetivales según la categoría gramatical de sus elementos componentes

Se tiene en cuenta no tratar la formación del tipo: ADJETIVO + -MENTE → ADVERBIO —*rápidamente, compulsivamente*— debido a que, aunque algunos autores (Pérez, 2002) la reconocen como un proceso de composición, está consolidada como una formación derivativa; así lo recoge el Diccionario General de la Lengua Española (VOX, 2003) ya que considera -mente como un sufijo, no elemento sufijal, que entra en la formación de adverbios de modo pospuesto a los adjetivos en su forma femenina[6] —*buenamente*.

Por otro lado, según se recoge en el trabajo de M. Alvar (2003), la composición se sirve de procedimientos para la creación de palabras nuevas, como son: sinapsia, disyunción, contraposición, yuxtaposición, prefijos vulgares y acortamiento. El más caudaloso de los procesos de composición es la *yuxtaposición o lexías compuestas*, aquí la fusión gráfica de los elementos participantes en el compuesto es total, así como su lexicalización y su gramaticalización —*carnicol, malqueda, cochitril, hincapié*. Sin embargo, la unión de los miembros en la *sinapsia* es de naturaleza sintáctica, no morfológica, por lo que es difícil determinar si se ha producido lexicalización o no. Suele

[5] Existe una marcada tendencia a colocar en el primer miembro sustantivos bisílabos

[6] El adjetivo adopta siempre la forma femenina, si la tiene, pues -mente es femenino en latín y conserva el acento.

existir un nexo de unión entre las dos palabras que dan lugar al nuevo término, generalmente con las preposiciones *de* y *a* —*pan de azúcar, paso a nivel, cuerda sin fin*—, correspondiendo este compuesto con la clasificación sintagmática que hace E. Bustos Gisbert (1986) de *Sustantivo + Preposición + Sustantivo*. Por otro lado, la *disyunción* da origen a un tipo de lexías en la que los dos elementos participantes no se han soldado gráficamente, por más que la lexicalización sea un hecho —*alta mar, peso pluma, pájaro mosca*— correspondiéndose con diversos compuestos *Sustantivo+Sustantivo* o *Sustantivo+Adjetivo* o *viceversa* de E. Bustos. Se puede incluso llegar, en las dos composiciones anteriores, a la unión gráfica de los elementos —*tela de araña→telaraña, agua nieve→aguanieve, ave fría→avefría.*

En un grado más alto de unión gráfica que los dos anteriores, está la *contraposición* donde los elementos que participan en ella se escriben unidos por un guión, aunque generalmente el resultado aparecerá sin él debido a las restricciones del uso del guión que tiene el español —*falda pantalón → falda-pantalón*—, considerándola como uno de los compuestos anteriores.

Se hace una mención especial a la combinación de una preposición con otra categoría gramatical, como por ejemplo: PREPOSICIÓN + SUSTANTIVO —*sobredosis, entreacto, contradanza*—, PREPOSICIÓN + VERBO —*sobredimensionar, contradecir, entresacar*—, PREPOSICIÓN + ADJETIVO —*sobreabundante, contrachapado, entremedio*—, pues generalmente se suelen confundir estos casos con composición cuando en realidad se trata de un proceso derivativo como indica S. Varela (1990) en su libro *Fundamentos de Morfología*, aunque M. Alvar (2002) considera la formación de palabras mediante prefijos vulgares como parte de la composición cuando estos prefijos coinciden con las preposiciones, esto es, se unen dos elementos independientes de la lengua.

Muchos autores consideran el *acortamiento* como un procedimiento de formación de nuevas palabras o neologismos que, por su naturaleza, escaparían, en principio, a una teoría morfológica, no como parte de la composición, pero la frontera entre ellos no está todavía muy clara sobre todo en el caso de abreviamiento.

Existen palabras que se forman a partir de una combinación de dos palabras más una sufijación, según D. Serrano (1995) —siguiendo la estructura *A+B+Sufijación*—, pero no existe la combinación *A+B* ni *B+Sufijación* —se considerarían derivación de palabras compuestas o derivados por sufijación, respectivamente—, y en caso de existir no son el origen de la palabra final. A este tipo especial de palabras se le conoce con el nombre de parasintéticos en composición.

misacantano —no existe *misacanta* ni *cantano*—, *ropavejero* —no existe *vejero*—, *doceañista* —no existe *añista*—, *aguamanil* —no existe *mano+il*

Es, por todo lo anterior, que en el presente artículo se procede a estudiar los compuestos yuxtapuestos o lexías compuestas en composición nominal y adjetiva, según la clasificación de E. Bustos (1986), al igual que algunos casos especiales de acortamiento, elementos compositivos y parasintéticos por composición, desde un punto de vista morfológico, ya que en los restantes tipos se han de tener en cuenta factores sintácticos y semánticos para poder justificar que es un verdadero compuesto en español y esto se escapa actualmente del dominio informático.

3 Reglas de formación de palabras compuestas

Las reglas de formación de compuestos que se estudian son de carácter léxico y no sintáctico, semántico o fonológico, lo que no excluye que, en ocasiones, sea necesario hacer referencia a dichos aspectos dado que son características inseparables de este proceso de formación de palabras.

Las primitivas, a las cuales se le aplican las reglas de formación de palabras compuestas que se buscan, deben ser palabras consolidadas en nuestra lengua o neologismos, pero nunca palabras incorrectas —ortografía irregular o palabras bloqueadas como **grabamiento* por *grabación*. La existencia en el lexicón[7] de innumerables particularidades no debería suponer una barrera para la automatización de este proceso, tanto en el reconocimiento como en la generación, dado que algunas de las palabras que presentan irregularidades admiten un reglado y el resto pueden llevar un tratamiento especial.

Para la deducción de las reglas de formación que permiten la identificación automatizada de palabras compuestas yuxtapuestas, se extraen

[7] Los ejemplos utilizados pueden ser formaciones neologísticas, no necesariamente documentadas en la base de referencia.

de diferentes diccionarios los distintos compuestos y se estudian los cambios gráficos que han sufrido las diferentes palabras que los forman así como las categorías gramaticales de éstas. Se obtiene un conjunto de reglas que, junto a las excepciones encontradas, permitan el reconocimiento y posterior generación automática de palabras compuestas.

3.1 Obtención de la base de estudio

Se parte de un conjunto de tamaño suficientemente significativo de compuestos, alrededor de 11000, y se clasifican en grupos según la categoría gramatical de sus constituyentes. Para ello, se toma como base documental de partida los compuestos recogidos en los principales repertorios lexicográficos de la lengua española (Bibliograf, 2003; RAE, 2001; Larousse, 1996; Clave, 1997; Moliner, 1996) y del glosario de compuestos del libro *La composición nominal en español* de E. Bustos (Bustos, 1986), donde para su elaboración se utilizaron obras de carácter general ―DRAE― y obras de carácter regional o dialectal ―hablas leonesas, aragonesas, meridionales, español de América.

3.2 Obtención de las reglas de formación

A partir de los conjuntos obtenidos se buscan las reglas de formación de los mismos. Se parte de la palabra y se prueba que cumpla unas ciertas condiciones ―tamaño de la palabra, por norma general, mayor a cinco caracteres, o formado por al menos tres sílabas[8] ―*uñalbo*. El proceso de reconocimiento empieza haciendo cortes a la palabra hasta que se encuentra algún vocablo íntegro o bien ambos, a los cuales se les aplica la regla correspondiente o bien se trata como excepción, utilizando para ello el "*Flexionador y Lematizador de palabras del español*" del *Grupo de Estructura de Datos y Lingüística Computacional* (Santana et al., 1997, 1999, 2006). Al ir generando cortes se pueden obtener múltiples soluciones entre las que se pueden encontrar varias que no lo sean, pero esta primera aproximación permite añadir otro tipo de condicionantes[9].

Hay que tener en cuenta que en un estudio cuyo objetivo sea la automatización de la morfología con medios informáticos, los aspectos formales o teóricos no tienen por qué coincidir con los estrictamente lingüísticos. Así, **altobajo* ―falsa composición pues lo correcto sería *altibajo*― no tendría por qué tratarse de una mala formación al no contravenir ninguna regla fonotáctica del lenguaje, ni siquiera la norma de la estructura silábica del español.

En la Figura 1 muestra el funcionamiento del procedimiento de reconocimiento de compuestos, donde una vez superada las condiciones de partida, como no contener ningún error ortográfico, se lematiza la palabra identificando sus diversas formas canónicas, categorías gramaticales y las flexiones o derivaciones que las produce. Por lo general, si se tratara de una palabra compuesta, las diversas formas canónicas se reducirían a una, coincidiendo con la palabra en si. Además, la categoría gramatical suele ser simple: sustantivo ―en la mayoría de los casos―, adjetivo o verbo. Se analiza si la palabra es compuesta "pura", derivada ―en cuyo caso se estaría hablando de derivación por composición, pasándose a reconocer la forma canónica de la misma―, o pudiera ser parasintética por composición (Serrano, 1995). A continuación, se aplican las reglas de formación de compuestos si no pertenece a la base de estudio, se comprueba cuáles son buenas, ordenados por criterios de idoneidad, en caso contrario, no existe solución en el reconocimiento. Por todo lo anterior, se buscan las reglas de formación a partir del estudio del comportamiento de los vocablos constituyentes del compuesto, algunas coinciden con las tratadas por lingüistas, aunque con una adaptación informática justificada por el comportamiento mayoritario observado. Además, se debe tener en cuenta que cuando se aplican las reglas hay que considerar los cambios gráficos que se pueden producir como consecuencia de aplicarla, siendo esto necesario para un correcto tratamiento informático.

3.3 Reglas de formación

Los procedimientos por los que se forman palabras compuestas pueden dividirse en dos grandes grupos atendiendo al grado de modificación que sufran los elementos

[8] Generalmente los vocablos en español son bisilábicos o trisilábicos, por lo que se puede decir que los compuestos nominales contienen de cinco a seis sílabas.

[9] Las palabras que forman el compuesto no son derivadas sino forman una única unidad léxica, es decir, no tienen prefijos, en especial el primer constituyen del compuesto, o bien no se admite la flexión de diminutivo en la segunda palabra del compuesto, entre otras.

originales: o bien por la mera adición de dos o más términos sin que ninguno de ellos se modifique —*rompeolas*, *abrelatas*, *mediodía*— o por la unión que conlleve algún tipo de modificación gráfica en alguno de los elementos que intervienen en la composición —generalmente ocurre en el primero de los componentes— o de adición al resultado final —*agridulce, rojiblanco, coliflor, balompié*.

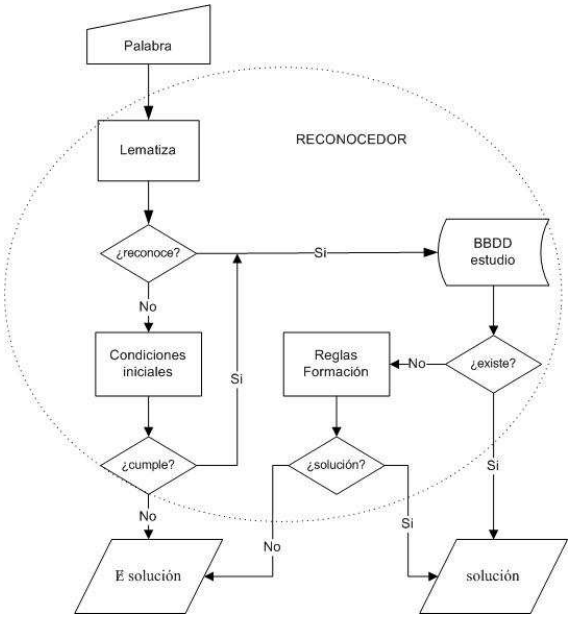

Figura 1.-Diagrama para el reconocimiento de palabras compuestas

Se presentan las reglas de formación de palabras compuestas detectadas en la base de estudio.

- Unión sin pérdida ni adición ni cambio. Esta regla se aplica de forma general a todas las categorías gramaticales. Generalmente, en la composición adjetiva de *Adjetivo+Adjetivo* se utiliza con adjetivos de nacionalidad y en la composición nominal de *Verbo+Verbo* se duplica el lexema verbal. La más productiva de la lengua española es la composición nominal de *Verbo + Complemento,* coincide en que ésta es, precisamente, la regla que más se utiliza para este tipo de compuesto —*dedododama, bienintencionado, iberoamericano, tragaavemarías.*
- Cuando las palabras a unir terminan y empiezan con la misma vocal *a*, *e* u *o*, se elimina una de ellas —sinalefas o reducciones de los hiatos. En *Verbo+Complemento* la pérdida que se produce es de la vocal *a,* pues normalmente el elemento verbal aparece en forma imperativa y la flexión de los verbos de la primera conjugación acaba en *a* —los más utilizados en esta regla—*claroscuro, uñalbo, caridelantado, ajolio, ceracate, aguardiente.*

  Caso adicional (*ha*, *ho*): Se elimina la *h* del punto de unión si se encuentra entre dos vocales iguales, como resultado de la unión de los constituyentes.

  matambre←mata+ambre←mata+hambre
- Sustitución de las vocales átonas finales *a, e* y *o* correspondientes al primer constituyente por el infijo compositivo *i*, o en el caso de *Verbo+Complemento* sustituir la terminación verbal, acabada en *a* para los verbos de la 1ª conjugación y en *e* para los de la 2ª y 3ª, por el infijo compositivo *i*, y llevar a cabo la unión sin pérdida. Generalmente, en la composición *Adjetivo+Adjetivo* se utiliza con adjetivos de color. Por la construcción del tiempo verbal en los compuestos *Verbo+Verbo,* la mayoría de los casos son con la vocal átona *e*. Sin embargo, son más los que hay que sustituir el morfema *o* en los compuestos de tipo *Sustantivo+Adjetivo* o *Adjetivo+ Sustantivo.* Debido a la estructura peculiar del compuesto *[Sustantivo+Adjetivo]*$_{Adjetivo}$ hay que tener en cuenta que existen palabras que aplicarían la regla anterior si no es porque previamente han incorporado el infijo compositivo *i* —*altibajo, cortiancho, rojinegro, dulciagrio, ardiviejas, batiaguas.*
- Si el primer elemento acaba en consonante y el segundo empieza en consonante, introducir la *i* en medio y llevar a cabo la unión. No se aprecia que se produzca en todos los tipos de composición, sino en aquellos casos en los que ambos componentes tienen igual categoría, a excepción de la combinación *Verbo + Verbo* —*azuliblanco, calicanto, paniqueso.*

  El papel que parece desempeñar el morfema compositivo *i* es de función conjuntiva que, aunque parece probable desde la perspectiva semántica —*carricoche, coliflor*—, no se justifica en otros casos, especialmente, en algunos compuestos adjetivales, porque en ellos no se atisba claramente ninguna razón de tipo semántico o fonológico.
- Se elimina la última vocal fuerte del primer constituyente, si acaba en vocal, o bien cuando es final vocálico el primer

componente y principio vocálico el segundo y tienen la posibilidad de formación de un diptongo o de una contracción, se procede a la unión con pérdida de la vocal del primer elemento —*eurasiático, papalbo, liquidámbar*..

Existen casos en que parece que más que aplicar esta regla, se podría haber utilizado la de sustituir el morfema por el infijo compositivo *i*, como puede verse en:

agridulce←agrio+dulce (≠ *agro+dulce)*

En otros casos, se dice que ha sufrido un proceso de pérdida del infijo compositivo *i:*

manvacío←manivacío←mano+vacío

- Los elementos compositivos pueden estar formados por la agregación de raíces cultas greco-latinas a una palabra española, antepuesta o pospuesta a la misma —*cornialto, denticonejudo, petrolífero, carnívoro*— o por la combinación de raíces cultas: raíz prefija y sufija griegas o latinas —*teléfono, filiforme*—, raíz prefija griega y sufija latina —*automóvil*— o raíz prefija latina y sufija griega —*hispanofilia*—, o por otro tipo de raíces —*arisblanco, euroasiático, galicursi.*

4 Conclusiones

A pesar de ser la composición, sin duda, uno de los procesos de formación de palabras de mayor importancia —pese al escaso tratamiento recibido por parte de la bibliografía— se ha contribuido a llenar en parte el vacío informático existente en su tratamiento, ya que han resultado infructuosas las búsquedas de referencias sobre procesamiento automático de la composición en español.

Uno de los procesos de formación de palabras compuestas más productivos, con respecto a los demás tipos compositivos, son las formaciones de *Verbo+Sustantivo* debido a la gran expresividad, simplicidad —el grado de modificación que sufren los elementos originales, en la mayoría de los casos, es nulo— y al frecuente uso que se hace en el lenguaje publicitario de este sistema compositivo. El auge en el uso de nuevos medios de comunicación social, el lenguaje periodístico, entre otros factores, han hecho que se creen neologismos compositivos debido a la rápida evolución de los acontecimientos y a su inmediata incorporación al mundo de las Tecnologías de la Información. Por ello, son imprescindibles procesos automáticos que sean capaces de identificar estas palabras situándolas en un contexto lingüístico adecuado, tanto desde un punto de vista morfológico como semántico.

Bibliografía

Alvar Ezquerra, M. 2002. "La formación de las palabras en español". Cuadernos de lengua española, Ed. Arco/Libros, Madrid.

Bibliograf, S.A. 2003. *"*Diccionario General de la Lengua Española VOX*"* Barcelona.

Bustos Gisbert, E. 1986. "La composición nominal en español", Universidad de Salamanca.

Clave SM. 1997. "Diccionario de Uso del Español Actual". Clave SM, Madrid.

Lang, M. 1992. "Formación de palabras en español. Morfología derivativa productiva en léxico moderno". Cátedra, Madrid.

Larousse Planeta, S.A. 1996. "Gran Diccionario de la Lengua Española". Larousse Planeta, S.A., Barcelona

María Moliner. 1996. "Diccionario de Uso del Español",Ed. Gredos, Madrid.

Pérez Cino, W. 2002. "Manual Práctico de formación de palabras en español I", Ed. Verbum.

Real Academia Española y Espasa-Calpe. 2001. "Diccionario de la Lengua Española", edición electrónica. 22ª edn. Madrid.

Santana, O.; Carreras, F.; Pérez, J.; Gutiérrez, V. 2006. "El reconocimiento automático de la composición en español". Conference Abstracts of the First International Conference of the Alliance of Digital Humanities Organizations.

Santana, O.; Pérez, J.; Carreras, F.; Duque, J.; Hernández, Z.; Rodríguez, G. 1999. "FLANOM: Flexionador y lematizador automático de formas nominales". Lingüística Española Actual XXI, 2, Ed. Arco/Libros, S.L.

Santana, O.; Pérez, J.; Hernández, Z.; Carreras, F.; Rodríguez, G. 1997. "FLAVER: Flexionador y lematizador automático de formas verbales". Lingüística Española Actual XIX, 2, Ed. Arco/Libros, S.L.

Serrano Dolader, D. 1995. "Las formaciones parasintéticas en español", Ed. Arco/Libros.

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 83-90 recibido 19-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Reutilización del Treebank de Dependencias del Euskera para la Construcción del Gold Standard de la Sintaxis Superficial de la Gramática de Restricciones (CG)

Reusability of the Basque Dependency Treebank for building the Gold Standard of Constraint Grammar Surface Syntax

José María Arriola, María Jesús Aranzabe, Iakes Goenaga

IXA NLP Group, University of the Basque Country (UPV/EHU)
Manuel Lardizabal 1 48014 Donostia
josemaria.arriola@ehu.es, maxux.aranzabe@ehu.es, iakesg@gmail.com

Resumen: El objetivo del trabajo consiste en reutilizar el *Treebank* de dependencias EPEC-DEP (BDT) para construir el *gold standard* de la sintaxis superficial del euskera. El paso básico consiste en el estudio comparativo de los dos formalismos aplicados sobre el mismo corpus: el formalismo de la Gramática de Restricciones *(Constraint Grammar*, CG) y la Gramática de Dependencias (*Dependency Grammar*, DP). Como resultado de dicho estudio hemos establecido los criterios lingüísticos necesarios para derivar la funciones sintácticas en estilo CG. Dichos criterios han sido implementados y evaluados, así en el 75% de los casos se derivan automáticamente las funciones sintácticas para construir el *gold standard*.
Palabras clave: reutilización recursos lingüísticos, creación *gold standard*, sintaxis superficial

Abstract: The aim of the work is to profit the existing dependency Treebank EPEC-DEP (BDT) in order to build the gold standard for the surface syntax of Basque. As basic step, we make a comparative study of both formalisms, the Constraint Grammar formalism (CG) and the Dependency Grammar (DP) that have been applied on the corpus. As a result, we establish some criteria that will serve us to derive automatically the CG style syntactic function tags. Those criteria were implemented and evaluated; as a result, in the 75 % of the cases we are able to derive the CG style syntactic function tags for building the gold standard.
Keywords: reusability of linguistic resources, gold standard creation, surface syntax

1 Introducción

La principal motivación de este trabajo es la construcción del *gold standard* de la sintaxis superficial del euskera reutilizando el *Treebank* EPEC-DEP (BDT) (Aranzabe, 2008). La premisa fundamental de la que parte el trabajo es la imposibilidad de generar el *gold standard* para evaluar la Gramática de Restricciones (Constraint Grammar, CG) de modo exclusivamente manual (Atro, 2012). La idea es la de agilizar dicho trabajo aprovechando los recursos existentes con el menor coste posible. En esta línea existen trabajos similares, entre los que cabría destacar los siguientes: Gelbukh et al., 2005; Nilson et al., 2008; Aldezabal et al., 2008 y Mille et al., 2009.

La reutilización de este recurso lingüístico nos permitirá obtener el *gold standard* de la sintaxis superficial del euskera correspondiente a las 300.000 palabras que constituyen el corpus EPEC (Corpus de Referencia para el Procesamiento del Euskera) (Aduriz et al., 2006). Este gold standard es un recurso indispensable para evaluar las gramáticas de restricciones del euskera (Aduriz et al., 2000) a nivel de las funciones sintácticas del estilo Constraint Grammar (CG) (Karlsson et al., 1995). Por tanto, al hablar de sintaxis superficial nos referimos al análisis de las funciones sintácticas que guardan las palabras agrupándose entre sí en sintagmas, oraciones simples y compuestas (Aduriz & Ilarraza, 2003). El análisis superficial de la oración de la Figura 1 (*Zure gorputza mapa bat zen non ez*

ISSN 1135-5948

nekien herrialde bakoitza non kokatu (Tu cuerpo era un mapa en el que no sabía dónde ubicar cada país)) muestra un ejemplo del análisis superficial que emplearemos para ilustrar los pasos seguidos en este trabajo. Nos centraremos en el análisis de los sintagmas nominales que aparecen resaltados en negrita en dicha oración, es decir, los sintagmas *Zure gorputza* (Tu cuerpo) y *mapa bat* (un mapa).

En la Figura 1 se presenta el análisis morfológico en formato CG. Básicamente, se puede observar que para cada palabra (representada entre los símbolos "< >") de la oración[1] el analizador morfológico ofrece un análisis por línea. El conjunto de los distintos análisis de cada palabra constituye la cohorte donde se recoge la siguiente información y en este orden: el lema, la categoría, la subcategoría, el caso, el número y por último la función sintáctica. Ciñéndonos a las palabras que tomamos como base para ilustrar el proceso, tenemos los siguientes análisis: *zure* (@IZLG>: complemento del nombre; *gorputza* y *mapa* que presentan las siguientes funciones sintácticas: (@SUBJ: sujeto; @OBJ: objeto; @PRED: predicado y @KM>: elemento modificador de la palabra portadora del caso).

```
"<$.>"<PUNT_PUNT>"
"<Zure>"<HAS_MAI>"
 "zu"  PRON 2ª PERSON S GEN  @IZLG>
"<gorputza>"
   "gorputz" N C ABS S @SUBJ
   "gorputz" N C ABS S @OBJ
   "gorputz" N C ABS S @PRED
   "gorputz" N C @KM>
"<herrialde>"
"<bakoitza>"
"<non>"
"<kokatu>"
"<ez>"
"<nekien>"
"<mapa>"
  "mapa"  N C @KM>
  "mapa"  N C ABS S @SUBJ
  "mapa"  N C ABS S @OBJ
  "mapa"  N C ABS S @PRED
"<bat>"
  "bat"  DET INDET  ABS S @SUBJ
  "bat"  DET INDET  ABS S @OBJ
  "bat"  DET INDET  ABS S @PRED
"<zen>"
"<$.>"<PUNT_PUNT>"
```

Figura 1: Ejemplo análisis superficial.

La idea principal es derivar las funciones sintácticas en estilo CG (etiquetas precedidas del símbolo @, ver Figura 1) partiendo del esquema de anotación basado en la Gramática de Dependencias que ha sido utilizado para el etiquetado manual de EPEC-DEP (BDT). El esquema de dependencias está constituido por 29 etiquetas y se basa fundamentalmente en el trabajo de Carroll *et al.*, (1998). Dichas etiquetas de dependencias representan las relaciones, siempre binarias, que se establecen entre los elementos terminales de las oraciones donde una palabra es el núcleo y la otra el dependiente. Así, en las dependencias se destacan las relaciones que se establecen entre las palabras de la oración (Figura 2) en oposición a otras aproximaciones que se basan en constituyentes o estructuras de frase. Por ejemplo, las relaciones de dependencia que corresponden a las palabras que sirven de ejemplo ilustrativo de la Figura 2 se definen de esta manera:

- El adjetivo posesivo *Zure* (tu) en función de complemento del nombre depende del nombre *gorputza* (cuerpo)
- El nombre *gorputza* (cuerpo) en función de sujeto depende del verbo *zen* (era).
- El nombre *mapa* (mapa) en función de predicado depende del verbo *zen* (era).
- El determinante *bat* (un) en función de modificador depende del nombre *mapa* (mapa).

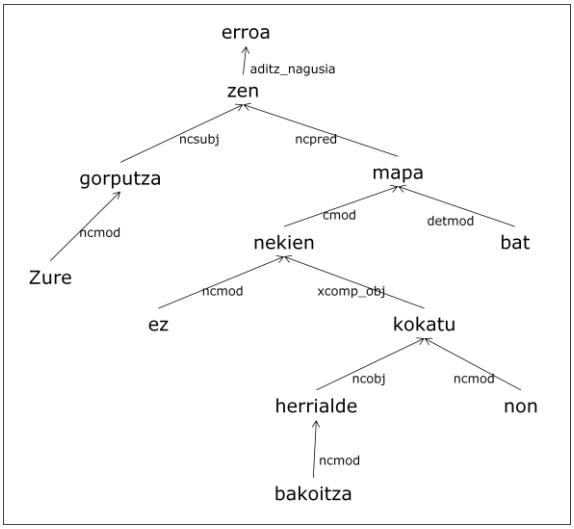

Figura 2: Análisis de dependencias.

El hecho de que el formalismo de dependencias esté basado en palabras al igual

[1] Sólo se ofrece el análisis de aquellas palabras de la oración utilizadas a modo de ejemplo.

que el formalismo de la gramática de restricciones hace factible el poder aprovechar el trabajo de etiquetado no sólo para las gramáticas de dependencias sino también para las gramáticas de restricciones.

En este artículo presentamos los resultados obtenidos a través del estudio comparativo de ambos formalismos de cara a la construcción del *gold standard*. Tras explicar los motivos del trabajo, en la segunda sección presentamos brevemente los recursos lingüísticos en los que se basa el mismo. En la tercera sección, describimos los pasos seguidos en nuestro estudio, centrándonos en los puntos principales del mismo. En la siguiente sección explicamos los criterios para la derivación automática de las funciones sintácticas. En la quinta sección mostramos los datos que corresponden a la evaluación. Finalmente, resumimos las conclusiones principales.

2 Características de los recursos sintácticos existentes

El corpus lingüístico etiquetado manualmente a nivel sintáctico siguiendo el análisis de la Gramática de Dependencias se denomina EPEC-DEP (BDT) y está constituido de 300.000 palabras del euskera estándar.

Con respecto a este recurso básico cabe destacar dos características:

- El trabajo previo llevado a cabo para la formalización del esquema de dependencias (Aranzabe 2008); (Aldezabal et al., 2009).
- La evaluación empírica de la calidad del etiquetado del corpus (Uria et al., 2009).

Ambas características redundan en la calidad y fiabilidad lingüística de nuestro punto de partida.

Por otro lado, disponemos del corpus EPEC etiquetado manualmente a nivel morfosintáctico siguiendo el formalismo de las gramáticas de restricciones (CG). Este corpus se compone de las mismas palabras que componen el Treebank EPEC-DEP (BDT), pero etiquetadas en este caso a nivel morfosintáctico y en formato CG (Karlsson et al., 1995) Para su etiquetado se parte del análisis automático de las gramáticas de restricciones y una vez analizado automáticamente los lingüistas han verificado y corregido en su caso el resultado de la desambiguación morfosintáctica automática. Estos análisis están pendientes de ser desambiguados a nivel sintáctico, es decir, si bien la categoría y el caso han sido resueltos, la desambiguación de las funciones sintácticas tales como por ejemplo la de sujeto u objeto están aún por resolver. Nuestro objetivo es por tanto asignar una sola función sintáctica a cada palabra partiendo del Treebank EPEC-DEP (BDT) construido manualmente.

3 Metodología para el estudio

El punto de partida lo constituye el conjunto de etiquetas fijadas para el análisis de dependencias (Aranzabe 2008); (Aldezabal et al., 2009) que han sido aplicadas manualmente por los lingüistas para la construcción de EPEC-DEP (BDT). Para reutilizar dicho esquema, se establecieron los siguientes pasos:

- Estudiar para cada relación de dependencia fijada en el esquema de anotación de dependencias las equivalencias con las funciones sintácticas siguiendo el formalismo CG.
- Una vez establecidas las equivalencias, definir los criterios por medio de los cuales se derivarán las funciones sintácticas siguiendo el formalismo CG.
- Implementar dichos criterios y estudiar los resultados de la aplicación de dicho criterios, para su posterior refinamiento si así procede.
- Evaluar la aplicación de dichos criterios por medio de dos lingüistas. Dos lingüistas se ocuparán de examinar por separado los resultados obtenidos automáticamente y determinarán la validez de los mismos.

En el siguiente apartado (4) explicaremos de modo somero el proceso relativo a la equiparación y definición de los criterios.

4 Criterios para la derivación automática de las funciones sintácticas

Los criterios para derivar automáticamente[2] las funciones sintácticas en estilo CG a partir del Treebank de dependencias se basan fundamentalmente en la especificación de las etiquetas de dependencias. A continuación se

[2] Basándonos en dichos criterios se implementó el programa en C++. Dicho programa examina un fichero de configuración y en virtud de los criterios lingüísticos mencionados, asigna la función sintáctica correspondiente.

muestra el esquema utilizado para describir las relaciones de dependencia correspondientes a los núcleos del sintagma nominal:

Etiqueta_dependencia (Caso del sintagma, núcleo, núcleo del sintagma, elemento portador del caso, función sintáctica)

Por ejemplo las relaciones de dependencia de las palabras que constituyen el sintagma nominal *Zure gorputza* en función de sujeto de la oración de la Figura 1 se etiquetan de la siguiente manera:

ncmod
(1. caso: genitivo,
2. núcleo del SN: *gorputza*,
3. modificador del SN: *zure*)

ncsubj
(1. caso: absolutivo,
2. núcleo de la oración: *zen*,
3. núcleo del SN: *gorputza*,
4. palabra que lleva el caso dentro del SN: *gorputza*,
5. Función: sujeto)

Teniendo en cuenta la relación de dependencia expresada por la etiqueta *ncsubj* (*non-clausal subject*) observamos que de dicha relación se puede derivar la función de sujeto (@SUBJ) en CG. De este modo y una vez determinada la equivalencia se establecen las condiciones que permiten derivar automáticamente una determinada función sintáctica para cada etiqueta de dependencias.

Por ejemplo, a partir de la etiqueta *ncsubj* se especifican las siguientes condiciones generales para derivar la función sintáctica de sujeto en CG (@SUBJ):

a. Asignar la función @SUBJ a la palabra del cuarto campo en la especificación de la etiqueta de dependencia.
b. Si la palabra del tercer campo no es la misma que la del cuarto, a esa palabra del tercer campo se le asignará la función @KM> (modificador del elemento portador del caso).
c. Si la palabra del cuarto y tercer campo son la misma, prevalecerá el criterio (a.).

Aplicando estos criterios al sintagma *Zure gorputza*, a la palabra *gorputza* se le asignará la función de sujeto (@SUBJ). En la Figura 3, esto se refleja por medio de la etiqueta *Correct* que asignamos automáticamente al análisis que contiene la función sintáctica obtenida mediante la aplicación de los criterios anteriormente explicados.

En relación al sintagma *mapa bat* (un mapa) la función de predicado (@PRED) desempeñada por *mapa* la derivaremos de la siguiente relación de dependencia:

ncpred
(1. caso: absolutivo,
2. núcleo de la oración: *zen*,
3. núcleo del SN: *mapa*,
4. palabra que lleva el caso dentro del SN: *bat*,
5. Función: predicado)

detmod
(1. caso: null,
2. núcleo del SN: *mapa*,
3. palabra que lleva el caso dentro del SN: *bat*)

En este caso partiendo de la etiqueta *ncpred* (*non-clausal predicate o sintagma en función de predicado*) establecemos los siguientes criterios generales:

a. Asignar la función @PRED a la palabra del cuarto campo o slot en la especificación de la etiqueta.
b. Asignar la función de verbo principal a la palabra del segundo campo (@+JADNAG).
d. Si la palabra del tercer campo no es la misma que la del cuarto, asignar a esa palabra la función @KM> (modificador del elemento portador del caso). Salvo que a dicha palabra se le haya aplicado anteriormente una regla para los elementos conjuntivos.
e. Si la palabra del tercer y cuarto campo son la misma, prevalecerá el criterio (a.).

Aplicando estos criterios al sintagma *mapa bat*, a la palabra *mapa* que se encuentra en el tercer campo o slot de la etiqueta de dependencias se le asignará la función de modificador del elemento portador del caso @KM>). Y a *bat* palabra que aparece en el cuarto campo se le asignará la función de predicado (@PRED). Ello viene reflejado en la Figura 3 por medio de la etiqueta *Correct* que asignamos automáticamente al análisis que contiene la función sintáctica obtenida mediante la aplicación de los criterios anteriormente explicados.

En la Figura 3 se muestra el resultado de la aplicación de los criterios para los sintagmas *Zure gorputza* y *mapa bat* en los cuales se han marcado las funciones sintácticas derivadas con la marca *Correct*.

```
"<$.>"<PUNT_PUNT>"
"<Zure>"<HAS_MAI>"
 Correct "zu"  PRON 2ª PERSON S GEN  @IZLG>
"<gorputza>"
 Correct  "gorputz" N C ABS S @SUBJ
   "gorputz" N C ABS S @OBJ
   "gorputz" N C ABS S @PRED
   "gorputz" N C @KM>
"<herrialde>"
"<bakoitza>"
"<non>"
"<kokatu>"
"<ez>"
"<nekien>"
"<mapa>"
 Correct "mapa"  N C @KM>
  "mapa"  N C ABS S @SUBJ
  "mapa"  N C ABS S @OBJ
 "mapa"  N C ABS S @PRED
"<bat>"
  "bat"  DET INDET  ABS S @SUBJ
  "bat"  DET INDET  ABS S @OBJ
  Correct"bat"  DET INDET  ABS S @PRED
"<zen>"
"<$.>"<PUNT_PUNT>"
```

Figura 3: Resultado de la derivación.

A continuación, explicaremos una serie de características básicas que se extraen del estudio comparativo.

En el análisis del sintagma nominal *mapa bat* (un mapa) se refleja una de las diferencias fundamentales entre el análisis de dependencias y el de las gramáticas de restricciones. Así, mientras que en las etiquetas de dependencias la función sintáctica principal recae en el elemento léxico (*mapa*), en la gramática de restricciones se asigna a la palabra portadora del caso en posición final del sintagma nominal (*bat*).

En relación al determinante *bat* (un) vemos que aparece en el análisis de dos etiquetas de relaciones de dependencias: *ncpred* y *detmod*.

Pero a la hora de derivar la función sintáctica de *bat* sólo tendremos en cuenta la etiqueta de *detmod* de la que derivaremos la función de modificador del sustantivo.

Las estructuras coordinadas son tratadas de manera distinta en ambos formalismos. En CG la conjunción se analiza con la etiqueta correspondiente al tipo de conexión que realiza, mientras que en GD es la conjunción la que se etiqueta con la función sintáctica correspondiente a la estructura coordinada. Así la relación de dependencia se expresa tomando como elemento gobernante la conjunción y los elementos coordinados como dependientes que se encuentran al mismo nivel. El análisis de las estructuras coordinadas resulta aún más complejo cuando además de las conjunciones los signos de puntuación, como por ejemplo la coma, funcionan como elementos de coordinación. Es éste por tanto uno de los fenómenos lingüísticos a tratar más profundamente.

Los criterios para derivar las funciones sintácticas constan de 10 reglas para las funciones sintácticas de los núcleos y 11 reglas para las funciones sintácticas de los dependientes. Hay a su vez un grupo de tres reglas que se encargan de las conjunciones y otra serie de categorías sintácticas.

Siguiendo este proceso a través del estudio del esquema general de cada etiqueta hemos establecido 41 criterios correspondientes a las principales etiquetas de dependencias, puesto que quedan fuera de este proceso de equiparación aquellas etiquetas denominadas como auxiliares. Estas etiquetas son las empleadas para etiquetar unidades multipalabra, posposiciones y partículas subordinantes independientes. En la Tabla 1 se muestran de modo simplificado las condiciones necesarias para derivar la función sintáctica correspondiente a los sintagmas, es decir, las etiquetas sintácticas en estilo CG que se pueden derivar de la correspondiente función sintáctica de dependencias (GD). Del mismo modo se han derivado las funciones sintácticas correspondientes a las oraciones.

Significado de la etiqueta	Etiqueta GD	Condiciones	Nº slot	Etiqueta CG
Sujeto	ncsubj	3 y 4 NO IGUAL; 3: @KM>	4	@SUBJ
Objeto	ncobj	3 y 4 NO IGUAL; 3: @KM>	4	@OBJ
Objeto indirecto	nczobj	3 y 4 NO IGUAL; 3: @KM>	4	@ZOBJ
Predicado	ncpred	2: @+JADNAG; 3 y 4 NO IGUAL; 3: @KM>	4	@PRED
Modificador	ncmod	2: CAT= V	4	@ADLG
Modificador	<ncmod	1: -	4	@<IA
Modificador	ncmod>	1: -	4	@IA>
Complemento del nombre	<ncmod	1: GEN	4	@<IZLG
Complemento del nombre	<ncmod	1: GEL	4	@<IZLG
Complemento del nombre	ncmod>	1: GEN	4	@IZLG>
Complemento del nombre	ncmod>	1: GEL	4	@IZLG>
Determinante	detmod>	---	3	@ID>
Determinante	<detmod	---	2	@<ID
Graduador	gradmod>	---	3	@GRAD>
Graduador	<gradmod	---	3	@<GRAD
Sujeto en aposición	aponcmod_subj	---	4	@SUBJ
Objeto en aposición	aponcmod_obj	---	4	@OBJ
Objeto indirecto en aposición	aponcmod_zobj	---	4	@ZOBJ
Adverbial en aposición	aponcmod_adlg	---	4	@ADLG
Complemento del nombre en aposición	aponcmod_izlg>	---	4	@IZLG>
Complemento del nombre en aposición	<aponcmod_izlg	---	4	@<IZLG

Tabla 1: Equivalencias sintácticas de los sintagmas.

A través del estudio comparativo hemos conseguido establecer las equivalencias a nivel sintáctico entre ambos formalismos. En el siguiente apartado mostraremos los resultados de la evaluación.

5 Evaluación

Para la evaluación se tomaron al azar 100 oraciones que cubrían todas las relaciones de dependencia. La evaluación consistió en examinar manualmente dichas oraciones. Y se observó que las reglas derivaban de forma correcta la etiqueta de CG en todos los casos.
Por tanto se consideró que no era preciso el examinar manualmente todo el corpus obtenido automáticamente. Antes de llevar a cabo dicha evaluación que se consideró como definitiva, se llevaron a cabo otra serie de evaluaciones que permitieron subsanar o completar la gramática.

El 25 % del corpus que ha quedado sin etiquetar automáticamente responde

fundamentalmente al hecho de que existen dos puntos de partida de análisis que son diferentes: en el análisis de CG los elementos multipalabra (ya sean posposiciones complejas, locuciones, partículas subordinantes, etc.) no son analizados como una sola unidad; por otro lado, el BDT que ha sido etiquetado manualmente por los lingüistas presenta estos elementos como una sola unidad, por tanto no hay correspondencia entre el número de tokens. Esta es la razón principal por la cual el proceso no se ha realizado totalmente de modo automático.

También cabría señalar otra serie de peculiaridades que presenta el corpus y que dificultan el proceso automático: títulos o encabezamientos, referencias bibliográficas, fórmulas matemáticas, estructuras parentizadas, vocativos, fechas entre corchetes, etc.

Como resultado de la aplicación de los criterios anteriormente señalados, hemos obtenido automáticamente las funciones sintácticas en el estilo de las gramáticas de restricciones en el 75 % de los casos partiendo del Treebank EPEC-DEP (BDT). Es decir, de las 304.751 palabras de las que consta el corpus, 228.982 han sido desambiguadas automáticamente de modo correcto. El 25 % restante (75.769) no se ha podido derivar automáticamente. Pero ello no significa que ambos formalismos no sean equiparables en todos los casos restantes. Así, cabe destacar que en la mayoría de los casos las diferencias radican a nivel de tratamiento de las unidades multipalabra, o de construcciones posposicionales o de ciertos elementos de

subordinación independientes. En todos ellos el denominador común es el de que nos encontramos con distintos estadios de análisis, en el análisis lingüístico llevado a cabo manualmente por los lingüistas se reconocen como una sola unidad las unidades multipalabra o se recogen como una unidad las construcciones posposicionales, por ejemplo.

En cambio, en el corpus analizado siguiendo el formalismo de las gramáticas de restricciones dichas estructuras aún no han sido procesadas en la mayoría de los casos, de ahí que no podamos equiparar propiamente dichos análisis.

6 Conclusión

Los recursos existentes se han reutilizado por medio de las reglas para derivar automáticamente las etiquetas en formato CG. Hemos establecido las bases metodológicas para la constitución del *gold standard* y hemos obtenido un 75% automáticamente. Por tanto, se trata de un trabajo en curso: queda por etiquetar manualmente el 25% del corpus que no se ha conseguido obtener automáticamente. Sobre este corpus un lingüista ha llevado a cabo un trabajo de evaluación del coste en número de horas. Se estima que serán necesarias 450 horas para llevar a cabo el etiquetado del 25% restante.

Por otro lado, si bien nuestro punto de partida ha sido el Treebank EPEC-DEP (BDT) observamos que los criterios establecidos también se podrían utilizar en sentido inverso, es decir, partiendo de un corpus etiquetado en estilo CG para obtener el etiquetado en dependencias. Con ello la reutilización se incrementaría. A su vez, la información formalizada para la derivación de etiquetas en uno u otro sentido, se podría reutilizar en técnicas de aprendizaje (*machine learning*), como por ejemplo, para entrenar un *parser* estadístico.

En un futuro se prevé el estudio detallado de las estructuras más complejas como la coordinación, aposición… con el objetivo de derivar automáticamente un mayor número de funciones sintácticas.

Agradecimientos

Este trabajo ha sido financiado por el Gobierno Vasco (IT344-10).

Bibliografía

Aduriz I., Arriola J. M., Artola X., Díaz de Ilarraza A., Gojenola K., y Maritxalar M. 1997. Morphosyntactic disambiguation for Basque based on the Constraint Grammar Formalism. *Proceedings of Recent Advances in NLP (RANLP97)*, páginas 282-288. Tzigov Chark, Bulgary.

Aduriz I. y Díaz de Ilarraza A. Morphosyntactic disambiguation and shallow parsing in Computational Processing of Basque. 2003. *Inquiries into the lexicon-syntax relations in Basque*. Bernarrd Oyharçabal (Ed.). University of the Basque Country.

Aduriz I., Arriola J. M., Artola X., Díaz de Ilarraza A., Gojenola K., Maritxalar M. y Urkia M. 2000. *Euskararako murriztapen-gramatika: mapaketak, erregela morfosintaktikoak eta sintaktikoak*. UPV/EHU/LSI/TR12-2000

Aduriz I., Aranzabe M. J., Arriola J. M., Atutxa A., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Oronoz M., Soroa A. y Urizar R. 2006. Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing. Andrew Wilson, Paul Rayson, and Dawn Archer. *Corpus Linguistics Around the World. Book series: Language and Computers.* Vol 56 (pag 1- 15). Rodopi Netherlands.

Aldezabal I., Aranzabe M.J., Díaz de Ilarraza A. y Fernández K. 2008. From Dependencies to Constituents in the Reference Corpus for the Processing of Basque. *Procesamiento del Lenguaje Natural*, nº 41 (2008), pp.147-154.

Aldezabal I., Aranzabe M. J., Arriola J. M. y Díaz de Ilarraza A. 2009. Syntactic annotation in the Reference Corpus for the Processing of Basque (EPEC): Theoretical and practical issues *Corpus Linguistics and Linguistic Theory 5-2,* 241-269. Mouton de Gruyter. Berlin-New York.

Aranzabe, M. J. 2008. Dependentzia-ereduan oinarritutako baliabide sintaktikoak: zuhaitz-bankua eta gramatika konputazionala. [Recursos sintácticos basados en la

Gramática de Dependencias: Treebank y Gramática Computacionsl]. PhD Thesis, Euskal Filologia Saila (UPV/EHU).

Carroll J., Briscoe T. y Sanfilippo A. 1998. Parser evaluation: A survey and a new proposal. International Conference on Language Resources and Evaluations, University of Granada (Spain).

Gelbukh A., Torres S. y Calvo H. 2005. Transforming a Constituency Treebank into a Dependency Treebank. *Procesamiento del Lenguaje Natural*, (35), 145-152.

Karlsson F., Voutilainen A., Heikkilä J. y Anttila A. 1995. *Constraint grammar: A language-independent system for parsing unrestricted text.* Berlin & NewYork: Mouton de Gruyter.

Mille, S., Burga, A.,Vidal,V. y Wanner, L. 2009. Towards a Rich Dependency Annotation of Spanish Corpora. In Proceedings of SEPLN, SanSebastian.

Nilsson, J. y Hall J. 2005. Reconstruction of the Swedish Treebank Talbanken. MSI report 05067, Växjö University: School of Mathematics and Systems Engineering.

Uria L., Estarrona A., Aldezabal I., Aranzabe M. J., Díaz de ILarraza A. y Iruskieta M. 2009. Evaluation of the Syntactic Annotation in EPEC, the Reference Corpus for the Processing of Basque Lecture Notes in Computer Science (LNCS) nº 5449, Alexander Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing*. pp 72-85. Mexico City, Mexico.

Voutilainen A., Purtonen T. K. y Muhonen K. 2012. Outsourcing Parsebanking: The FinnTreeBank Project. Diana Sousa, Krister Lindén, Wanjiku Nganga (Ed.), *Shall we Play the Festschrift Game? : Essays on the Occasion of Lauri Carlson's 60th Birthday*. pp 117-131. Springer Verlag.

Desarrollo de Recursos y Herramientas Lingüísticas

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 93-100 recibido 29-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Verb SCF extraction for Spanish with dependency parsing

Extracción de patrones de subcategorización de verbos en castellano con análisis de dependencias

Muntsa Padró
Universidade Federal do Rio Grande do Sul
Av. Bento Gonçalves, 9500
Porto Alegre -Brasil
muntsa.padro@inf.ufrgs.br

Núria Bel
Universitat Pompeu Fabra
Roc Boronat 138
08018 Barcelona - Spain
nuria.bel@upf.edu

Aina Garí
Universitat de Barcelona
Gran Via de les Corts Catalanes, 585
08007 Barcelona - Spain

Resumen: En este artículo presentamos los resultados de nuestros experimentos en producción automática de léxicos con información de patrones de subcategorización verbal para castellano. La investigación se llevó a cabo en el marco del proyecto PANACEA de adquisición automática de información léxica que redujera al máximo la intervención humana. En nuestros experimentos, se utilizó una cadena de diferentes herramientas que incluía 'crawling' de textos de un dominio particular, normalización y limpieza de los textos, segmentación, identificación de unidades, etiquetado categorial y análisis de dependencias antes de, finalmente, la extracción de los patrones de subcategorización. Los resultados obtenidos muestran una gran dependencia de la calidad de los analizadores de dependencias aunque, no obstante, están en línea con los resultados obtenidos en experimentos similares para otras lenguas.
Palabras clave: Adquisión automática de patrones de subcategorización, análisis de dependencias, adquisición léxica.

Abstract: In this paper we present the results of our experiments in automatic production of verb subcategorization frame lexica for Spanish. The work was carried out in the framework of the PANACEA project aiming at the automatic acquisition of lexical information reducing at maximum human intervention. In our experiments, a chain of different tools was used: domain focused web crawling, automatic cleaning, segmentation and tokenization, PoS tagging, dependency parsing and finally SCFs extraction. The obtained results show a high dependency on the quality of the results of the intervening components, in particular of the dependency parsing, which is the focus of this paper. Nevertheless, the results achieved are in line with the state-of-the-art for other languages in similar experiments.
Keywords: Automatic subcategorization frame acquisition, dependency parsing, lexical acquisition.

1 Introduction

Knowledge of Subcategorization Frames (SCF) implies the ability to distinguish, given a predicate in raw text and its co-occurring phrases, which of those phrases are arguments (obligatory or optional) and which adjuncts. Access to SCF knowledge is useful for parsing as well as for other NLP tasks such as Information Extraction (Surdeanu et al., 2003) and Machine Translation (Hajič et al., 2002). SCF induction is also important for other computational linguistic tasks such as automatic verb classification, selectional preference acquisition, and psycholinguistic experiments (Lapata et al., 2001; Schulte im Walde and Brew, 2002; McCarthy and Carroll, 2003; Sun et al., 2008a, 2008b).

In this paper we present the results of our experiments in automatic production of verb subcategorization frame lexica for Spanish with special focus on the use of statistical dependency parsing. The work was carried out in the framework of the PANACEA project (7FP-ICT-248064) aiming at the automatic

ISSN 1135-5948

acquisition of lexical information reducing at maximum human intervention. Therefore, fully automation of the process and human work reduction were the main criteria for choosing methods and assessing the results. In our experiments, a chain of different tools was used: domain focused web crawling, automatic cleaning, segmentation and tokenization, PoS tagging, dependency parsing and finally SCFs induction[1]. In the context of this project, we focused on maximizing precision. In order to contribute to the production of resources for working systems, we understood that it could be an asset to produce automatically a SCFs lexicon where good entries can be clearly separated from dubious ones. Then, human revision could concentrate on the dubious ones while still saving time and effort if those identified as reliable were actually good.

The obtained results show a high dependency on the quality of the results of the intervening components. Nevertheless, the results achieved are in line with the state-of-the-art for other languages in similar experiments.

2 Related work

The possibility to induce SCFs from raw corpus data has been investigated mostly for verbs and it is based on a first hypothesis generation step followed by a filtering step that tries to separate actual complement combination patterns from occasional combinations (see Korhonen, 2010, for a survey on different techniques for different languages). Current researched systems rely on the information supplied by an intermediate parser that identifies constituents and their grammatical function. Thus a first step collects sequences of constituents and their frequency, and a second step tries to select those combinations that are consistently found. Evaluation is made in terms of precision, i.e. only actual SCFs for a particular verb type, that is, a lemma, must be assigned, and coverage, i.e. all the possible SCFs for a particular verb type must be assigned. The main problem of current techniques has to do with maximizing both precision and coverage for each particular verb because (i) SCFs distribution is Zipfian and usual frequency filters fail to select infrequent patterns, and (ii) because the correlation between the conditional distribution of SCFs given a particular verb type and the unconditional distribution independent of specific verb types is very small.

Verb SCF acquisition for Spanish has already been addressed. Chrupala (2003) presented a system to learn subcategorization frames for 10 frequent verbs[2] of two classes, verbs of change and verbs of path, from a 370,000 word corpus by adopting the existing scheme of classification of Spanish SCFs from the SENSEM verb database developed in the VOLEM project (Fernández et al. 2002). The experiment searched chunked corpora and detected potential SCFs for 10 Spanish verbs. Semantic information of nouns, in particular the 'human' feature, is added in the chunking step in order to handle phenomena such as direct objects marked with the preposition 'a'. The SCFs hypothesis generation is based on matching the actual co-occurrences against a number of previously defined syntactic patterns associated with specific SCFs in the form of templates. A number of rules generate different variants of a number of initial, canonical templates.c For instance, a rule generates clitiziced variants of full NP SCFs.

As for the evaluation, 20 sentences for each one of 10 verbs were randomly selected and system results were compared with a manually corrected version of the SCFs selected. The Chrupala (2003) system achieved a precision of 0.56 in token SCF detection. The results were also evaluated for types: for each verb the number of detected SCFs was collected and compared with the manual reference with a filtering phase based on a relative frequency cut-off. Best results published were, at a cut-off of 0.07, 0.77 precision and 0.70 recall.

Esteve Ferrer (2004) carried out the SCF extraction experiment on a corpus of 50 million words, also PoS tagged and chunked. The task was to assign acquired SCFs to verb types after the two explained phases of hypothesis generation and posterior filtering. A predefined list of 11 possible SCFs, each made of plausible combinations of a maximum of two constituents, were considered. The predefined SCFs considered different prepositions that were grouped manually with semantic criteria. Hypothesis selection was performed with a Maximum Likelihood Estimate (MLE, Korhonen and Krymolowski 2002). Evaluation was carried out comparing with a manually

[1] All the PANACEA materials and tools are available at www.panacea-lr.eu

[2] Bajar, convertir, dejar, desatar, deshacer, llenar, preocupar, reducir, sorprender, decir.

constructed gold standard for a sample of 41 randomly chosen verbs that included frequent but also infrequent verbs. These experiments gave the following results at a frequency cut-off of 0.05: 0.71 precision, 0.61 recall.

The main novelties of our work with respect to these previous experiments for SCF extraction for Spanish verbs are: the use of a dependency-parsed corpus and no need to have a list of predefined templates or SCFs to match. A further innovative aspect investigated in our project has to do with the amount of expert language dependent knowledge involved in the used methods. Until recently, state of the art SCF acquisition systems used handcrafted rules to generate hypothesis (Chrupala, 2003) or to match natural language parser output to a set of pre-defined SCFs (Briscoe and Carroll, 1997; Korhonen, 2002; Preiss et al., 2007, Esteve Ferrer, 2004). More recent works, however, propose to use an inductive approach, in which the inventory of SCFs is also induced directly from parsed corpus data (O'Donovan et al., 2004; Cesley and Salmon-Alt, 2006; Ienco et al., 2008; Lenci et al., 2008; Kawahara and Kurohashi, 201). In Messiant (2008) inductive system, that we used as explained in section 3, candidate frames are identified by grammatical relation (GR) co-occurrences. The only given information is the label of GRs that are to be considered. Statistical filtering or empirically-tuned thresholds are again used to select frames for the final lexicon. This inductive approach has achieved respectable accuracy for different languages (0.60-0.70 F1-measure against a dictionary), do not involve predefined expert knowledge and is more portable than earlier methods. They are also highly scalable to large volumes of data, since the identification and selection of frames for producing the lexicon generally takes minimal time and resources.

The application of the inductive method is dependent, however, on the availability of a parser. The IULA Spanish Treebank (Marimon et al. 2012) allowed us to train two different statistical parsers: Malt (Nivre and Hall, 2005; Nivre et al., 2007) and one of the parsers in Mate-tools (Bohnet, 2010). These parsers were used to obtain the syntactic information to test the Messiant (2008) inductive method.

In the next sections we present the results of using the inductive approach for Spanish verbs. The only comparable exercise for Spanish is Altamirano and Alonso (2010). They developed a SCF extraction system based on SCF induction and frequency based selection. The system was tested using the SENSEM corpus (Castellón et al., 2006). The corpus contained 100 sentences for each of the 250 most frequent Spanish verbs. For the SCF induction experiment, sentences were manually annotated with GR information, what makes the experiment similar to ours. Note, however, that in our experiments automatic parsing introduced errors that affected the induction results, as we will discuss in section 5.1. The Altamirano and Alonso (2010) experiment evaluation was carried out by manually inspecting the results for the 20 most frequent verb senses. Results obtained were: 0.79 precision and 0.70 recall.

3 Methodology

For the SCF induction, we used a Messiant (2008) based SCF extractor as implemented in the *tcp_subcat_inductive* web service developed by University of Cambridge[3]. The input to the web service is the output of a parser either in RASP parser format or in the CoNLL format. The user can decide which GR labels are candidates to be arguments of a verb, and hence part of subcategorization frames, and which are not and should not be considered. Note that the user does not define specific combination patterns, as in earlier SCF acquisition approaches: if the user specifies DOBJ and XCOMP as GR labels of interest, but not MODIFIER, then the SCF inventory will consist of all observed combinations of DOBJ and XCOMP, while MODIFIER will never appear in any SCF.

The system outputs the observed frequency of combinations of the addressed GRs for each verb as potential SCFs, what allows filtering them by their frequency. An adequate filtering threshold is tuned heuristically, as we will see later. The concrete information we extracted in our experiments was:

- Subject and verb complements: Direct Object (DOBJ), Indirect Object (IO), predicative and object-predicative complements and prepositional phrase complements (PP): bounded preposition, direction and location PPs.
- For subject and complements, we also considered whether the complement is

[3]Available at http://registry.elda.org/services/304

realised by a noun phrase, or a clause phrase.
- For PP complements with bounded preposition we also extract the particular preposition.

All this information is extracted by the *tcp_subcat_inductive* tool using the adequate parametrization.

The experiments were carried out on two domain specific corpora: Environment (46.2M tokens) and Labour Legislation (53.9M tokens). The corpora were automatically crawled and cleaned (Bel et al. 2012). From these corpora, all sentences containing the target lemmas were extracted and parsed. For each corpus, 30 target verbs were selected and a gold-standard was manually annotated for evaluation purposes (up to 200 sentences were annotated for each verb). The gold-standard in the form of a lexicon of possible frames associated to each verb type was derived from the actual occurrences of target verb types. Because of the restriction of having a minimum number of occurrences, the final list of verb types differs for the two corpora.

The evaluation was made in terms of number of acquired SCFs that were indeed SCFs in the gold standard per verb type (precision) and SCFs acquired with respect to the number of SCFs for every verb type in the gold standard (recall).

4 Experiments

As already mentioned, the goal of the experiments was to assess the induction method for Spanish using an automatic chain of processing tools, in particular dependency parsing. In our experiments, we tried two different dependency parsers and we experimented with different filtering thresholds as well as with other filtering approaches (ensemble and pattern-based filtering) in order to raise precision and to guarantee a clear-cut between reliable SCF assignments and dubious ones that would still need human revision. In what follows we present the configuration of the different experiments performed.

4.1 Different parsers

The SCF extractor can be applied to any parsed corpus. Thus, we used two different parsers to produce the input of the SCF extractor. The parsers were: (i) Malt parser (Nivre and Hall, 2005; Nivre et al, 2007) optimized with MaltOptimizer (Ballesteros and Nivre, 2012). (ii) Mate graph-based re-scoring (completion model) parser (Bohnet and Kuhn, 2012; Bohnet and Nivre, 2012). Both parsers were trained with the IULA Treebank[4] (Marimon et al. 2012). The parsers in turn were applied to PoS tagged text obtained with FreeLing v3 tools (Padró and Stanilovsky, 2012). Both parses had a high performance in terms of Labelled Attachment Score (LAS), being Mate the parser with higher LAS (94,7% vs 93,2% for Malt, with a test set from the IULA Treebank). However the exact match score, i.e. every complement in the parsed sentence is correctly analysed, was around 50%. Note that SCF extraction identifies frequent GR combinations in whole sentences, and therefore is very much affected when the parser repeatedly delivers combinations of correct but also wrong GR. In order to sort out this problem we tried two different strategies: combining the results of two parsers and filtering known bad combinations.

4.2 Ensemble Strategy

Given that we had data from two different parsers, we tried to raise SCF extraction precision by selecting as good ones only those SCF-verb assignments that resulted from considering the data from the two parsers in the extraction phase. The hypothesis behind was that if a particular SCF is not output by two systems, using each a different parser, it is unlikely that the GR combination is correct.

4.3 Filter Strategy

In order to assess the frequency filtering with respect to precision, we cleaned parser results by applying hand-made filters for erasing known parser frequent errors, for instance SCFs with more than one subject or direct object, with both a by_agent and a direct object, and so on. This strategy was only applied to Malt parser results to assess the benefits of this strategy that, note, requires expert human knowledge.

5 Results

In this section we present the obtained results for the Labour Legislation (LAB) and the Environment (ENV) corpora.

[4] http://www.iula.upf.edu/recurs01_tbk_uk.htm

Table 1 and 2 show the summary of the obtained results with the experimental settings presented in the previous section for both corpora. Since we are especially interested in developing systems with high precision, in that table we present the best results maximizing F1 and maximizing precision. The table shows Precision, Recall and F1 values averaged for all verbs and the likelihood-based cut off to get these results. We also present results with the additional filters devised for known errors delivered by the Malt parser to assess their impact (Malt+F).

method	thresh	P	R	F1
Choosing best F1				
Malt	0.04	0.6923	0.5094	**0.5870**
Malt + F	0.04	0.8571	0.5094	**0.6391**
Mate	0.04	0.6848	0.5943	**0.6364**
Ensemble	0.04	0.7195	0.5566	**0.6277**
Choosing best P				
Malt	0.1	**0.8723**	0.3868	0.5359
Malt + F	0.09	**0.9167**	0.4151	0.5714
Mate	0.1	**0.8333**	0.4245	0.5625
Ensemble	0.09	**0.8800**	0.4151	0.5641

Table 1: Best results over LAB corpus

method	thresh	P	R	F1
Choosing best F1				
Malt	0.05	0.8421	0.5053	**0.6316**
Malt + F	0.04	0.9074	0.5158	**0.6577**
Mate	0.06	0.8947	0.5368	**0.6711**
Ensemble	0.05	0.8548	0.5579	**0.6752**
Choosing best P				
Malt	0.1	**0.9545**	0.4421	0.6043
Malt + F	0.07	**0.9778**	0.4632	0.6286
Mate	0.1	**0.9375**	0.4737	0.6294
Ensemble	0.08	**0.9787**	0.4842	0.6479

Table 2: Best results over ENV corpus

Additionally, Figure 1 shows details for all the results using Malt over the LAB corpus. In this figure we can see how increasing the filtering threshold leads to a better precision but to a loss in recall. Note that the frequency cut-off maximizing precision separates those assignments that are almost certain and would not need further revision. Table 3 shows the best results in terms of F1 obtained (Mate parser and ENV corpus) related to the number of extracted SCFs. Note that our gold-standard has 32 possible SCFs.

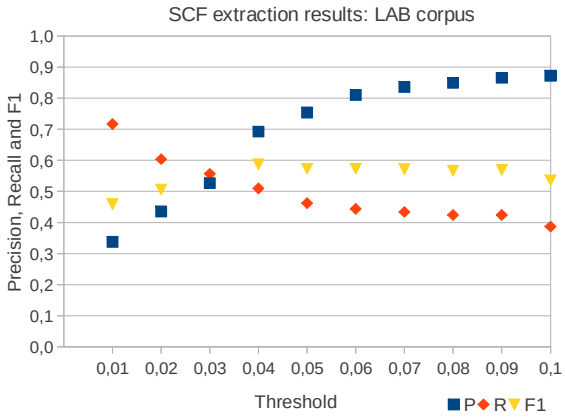

Figure 1: SCF results with Malt and LAB corpus

Thres.	# SCF	P
0.01	24	0.5462
0.02	22	0.6162
0.03	18	0.6867
0.04	16	0.7794
0.05	14	0.8548
0.06	11	0.9245
0.07	10	0.9583
0.08	10	0.9787
0.09	10	0.9778
0.1	10	0.9778

Table 3: Number of SCF extracted with Mate and ENV corpus, and precision results

5.1 Discussion

Figure 2 shows graphically the results of every experimental set up. Malt and Mate parser results influenced the performance of the extractor. The results show that all strategies can achieve good precision scores, but with a dramatic cost in recall, as it was expected: infrequent SCFs are left out.

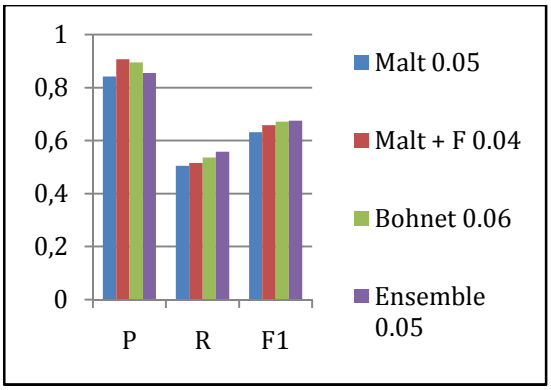

Figure 2: Comparison of Results for best F1 dataset in terms of Precision, Recall and F1 for ENV corpus.

The low recall for all systems is also due to the poor performance of both parsers for identifying some particular dependencies. The clearest example is the difficulties both parsers have in identifying indirect objects (IO). For example, Mate parser annotates IOs with 68% Precision and 52% Recall. Malt parser obtains poorer results. Those figures make the parser output hardly reliable for this low frequent complement (Padró et al., 2013).

The parser limitation to correctly detect IOs is also the reason for the differences between the two corpora results. The evaluation gave quite surprisingly different results for the two corpora, ENV corpus delivering overall better results. Manual inspection demonstrated that the difference came from the particular verbs chosen for the gold standard. In the ENV gold standard only 9 SCFs include IO, while in the LAB gold standard there were 18 SCFs. The parsers rarely deliver parses with this type of complements because they systematically assign a wrong label, therefore the SCF extractor never proposes candidates with IO although they are present at the gold standard. Therefore, in the LAB test set the results show a lower recall because accidentally there were more IOs to be found.

It is clear that using Mate parser to annotate the corpus, the system obtains just a slightly lower precision than the combined Malt and hand-made filter strategy (Malt+F), but better recall and F1 even although with a higher frequency cut-off. This means that using Mate parser we got competitive results without the need of developing hand-made rules, thus, resulting in a more general approach.

The ensemble strategy delivered poorer results in terms of precision gain, but it had better recall scores. This is partially due to the fact that better precision results are obtained with a lower threshold and therefore, more candidates are taken into account. We also found interesting that also for the same threshold, in some cases, an improvement on recall is observed. This is due to the fact that when combining SCFs extracted with both parsers the frequencies associated to each SCF change, making possible that some SCFs that were filtered with a given threshold for one system, are not filtered after the ensemble. The case of "presentar" (‘to present’) in ENV corpus when using a threshold of 0.03 is an example. In the gold standard this verb has three assigned SCF: transitive verb with noun phrase (both direct object and subject are NP), intransitive verb (subject as NP) and ditransitive verb (subject and DOBJ are NP, and a further IO). Ditransitive SCF has a frequency in the gold standard of 0.02, and it is not learned by any of the systems (as said, IOs are badly tagged by the parsers, so it is very hard to learn SCFs that contain them). Transitive and intransitive are acquired by both Malt and Mate parsers, but with Malt parser data the extractor assigns to the intransitive frame a frequency of 0.02 and is thus not selected with a threshold of 0.03. On the other hand, with Mate parser data the system do extract the intransitive SCF, but with a frequency of 0.11 (closer to the gold standard frequency, which is 0.16). Thus, after the ensemble, the intransitive frame receives a frequency of 0.07 and is thus selected with 0.03 threshold. Note that we performed the ensemble before the filtering and apply the thresholds to the obtained results, in order to be able to capture these changes in frequency.

Comparing our strategy with previous experiments, we see that, although it is impossible to compare scores, the most noticeable fact is that the use of a dependency parser leads to competitive results in terms of precision, although with a poorer recall. However, this strategy requires less previous specific knowledge and manual work.

Furthermore, the number of SCFs in our gold-standards is bigger than those of previous work (32 different SCFs vs. 23 and 11 of Chrupala, 2003 and Esteve, 2004 respectively). This is also a further factor for assessing the low recall scores we got. In fact we are learning a similar number of SCFs than previous work, as shown in Table 3, but since our gold standard is more fine-grained, the resulting recall is lower.

6 Conclusions and future work

In this work we tested a SCF acquisition method for Spanish verbs. The used system extracts the SCFs automatically from dependency-parsed corpora building the SCF inventory at the same time as the lexicon.

We have seen that parser errors severely affect the SCF extraction results. Even though we are using state of the art parsers with very competitive performance, the systematic errors they produce for some infrequent complements make it impossible the identification of SCFs that contain such complements and thus causes

low recall. Nevertheless, we have obtained a system with good precision, though the recall needs still to be improved.

In order to improve the results, some future work will be necessary to improve parser output. One possible line is to filter out unreliably parsed sentences before running the SCF acquisition system, for example, performing the ensemble of the two parsers at sentence level instead of applying it to the output of the SCF extractor. Nevertheless, we do not expect that to solve the problem of undetected complements, such as IOs. Specific improvements in parsing will, therefore, be needed. In that line, we would encourage the parsing community to start considering other ways of evaluating parser accuracy apart from LAS. We have seen in our work that parsers with very high LAS may fail to label very important but infrequent complements that are needed for subsequent tasks.

7 Acknowledgments

This work was funded by the European Project PANACEA (FP7-ICT-2010- 248064). We specially thank Laura Rimell, Prokopis Prokopidis and Vassilis Papavasiliou, PANACEA partners, for their support. Our gratitude also to Miguel Ballesteros, Héctor Martínez and Bernd Bohnet for their kind support for MALT-optimizer and MATE tools.

8 References

Altamirano, R, and Alonso, L. (2010) IRASubcat, a highly customizable, language independent tool for the acquisition of verbal subcategorization information from corpus. Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas.

Ballesteros, M. and Nivre, J. (2012). MaltOptimizer: A System for MaltParser Optimization. In Proceedings of the International Conference on Language Resources and Evaluation, LREC'12.

Bel, N., Papavasiliou, V., Prokopidis, P., Toral, A., Arranz, V. (2012). "Mining and Exploiting Domain-Specific Corpora in the PANACEA Platform", in The 5th Workshop on Building and Using Comparable Corpora. LREC'12.

Bohnet, B. (2010). Top Accuracy and Fast Dependency Parsing is not a Contradiction. Proceedingf of 23rd International Conference on Computational Linguistics (COLING 2010).

Bohnet, B. and Kuhn, J (2012). The best of both worlds: a graph-based completion model for transition-based parsers. In Proceedings of the EACL 2012.

Bohnet, B. and Nivre, J. (2012). A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In Proceedings of the EMNLP-CoNLL 2012.

Briscoe, E. J. and Carroll, J. (1997). Automatic extraction of subcategorization from corpora. In Proceedings of the 5th ACL Conference on Applied Natural Language Processing,

Castellón, I., A. Fernández, G. Vázquez, L. Alonso, J.A. Capilla (2006). The Sensem Corpus: a Corpus Annotated at the Syntactic and Semantic Level. In Proceedings of the International Conference on Language Resources and Evaluation, LREC'06.

Chesley, P. and Salmon-Alt, S. (2006) Automatic extraction of subcategorization frames for French. In Proceedings of the International Conference on Language Resources and Evaluation, LREC'06.

Chrupala, G. (2003). Acquiring Verb Subcategorization from Spanish Corpora. DEA Thesis, University of Barcelona.

Esteve-Ferrer, E. (2004). Towards a semantic classification of Spanish verbs based on subcategorisation information. In Proceedings of the ACL 2004 Workshop on Student Research.

Fernández, A., P. Saint-Dizier, G. Vázquez, F. Benamara, M. Kamel (2002). "The VOLEM Project: a Framework for the Construction of Advanced Multilingual Lexicons", in Proceedings of the Language Engineering Conference.

Hajič, J; M. Čmejrek; B. Dorr; Y. Ding; J. Eisner; D. Gildea; T. Koo; K. Parton; G. Penn; D. Radev; and O. Rambow (2002). Natural language generation in the context of machine translation. Technical report. Center for Language and Speech Processing,

Johns Hopkins University, Baltimore. Summer Workshop Final Report.

Ienco, D.; S. Villata and C. Bosco (2008). Automatic extraction of subcategorization frames for Italian. In Proceedings of the International Conference on Language Resources and Evaluation, LREC'08.

Kawahara, D and Kurohashi, S. (2010): Acquiring Reliable Predicate-argument Structures from Raw Corpora for Case Frame Compilation, In Proceedings of the International Conference on Language Resources and Evaluation, LREC'10.

Korhonen, A. (2002). Subcategorization acquisition. Ph.D. thesis, University of Cambridge Computer Laboratory.

Korhonen, A. (2010). Automatic Lexical Classification - Bridging Research and Practice. In Philoshophical Transactions of the Royal Society. 368: 3621-3632.

Korhonen A. and Y. Krymolowski (2002). On the Robustness of Entropy-Based Similarity Measures in Evaluation of Subcategorization Acquisition Systems. In Proceedings of the Sixth CoNLL.

Lapata, M. F. Keller, and S. Schulte im Walde (2001). Verb frame frequency as a predictor of verb bias. Journal of Psycholinguistic Research, 30(4):419-435.

Lenci R, Mcgillivray B, Montemagni S, Pirrelli V. (2008) Unsupervised acquisition of verb subcategorization frames from shallow-parsed corpora. In Proceedings of the International Conference on Language Resources and Evaluation, LREC'08.

Marimon, M., Fisas, B., Bel, N., Arias, B., V ázquez, S., Vivaldi, J., Torner, S., Villegas, M. and Lorente, M. (2012). The IULA Treebank. In Proceedings of the International Conference on Language Resources and Evaluation, LREC'12.

McCarthy, D. and J. Carroll (2003). Disambiguating nouns, verbs and adjectives using automatically acquired selectional preferences. Computational Linguistics 29:4.

Messiant, C. (2008). A subcategorization acquisition system for french verbs. In Proceedings of the ACL2008 (Student Research Workshop).

Nivre, J and Hall, J. (2005). MaltParser: A language-independent system for data-driven dependency parsing. In Proceedings of the 4th Workshop on Treebanks and Linguistic Theories (TLT).

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S. and Marsi, E. (2007). Maltparser: A language-independent system for data-driven dependency parsing. Natural Language Engineering, 13:95–135.

O'Donovan R, Burke M, Cahill A, van Genabith J, Way A. (2004) Large-scale induction and evaluation of lexical resources from the penn-ii treebank. In Proceedings of ACL 2004.

Padró, LL. and Stanilovsky, E. (2012); FreeLing 3.0: Towards Wider Multilinguality. In Proceedings of the International Conference on Language Resources and Evaluation, LREC'12.

Padró, M., Ballesteros, M., Martínez, H. and Bohnet, B. (2013). Finding dependency parsing limits over a large Spanish corpus. In Proceedings of IJCNLP 2013.

Preiss, J., Briscoe, E. J. and A. Korhonen. (2007). A System for Large-scale Acquisition of Verbal, Nominal and Adjectival Subcategorization Frames from Corpora. In Proceedings of ACL 2007.

Surdeanu, M.; Harabagiu, S.; Williams, J. and Aarseth, S. (2003). Using predicate-argument structures for information extraction. In Proceedings of ACL 2003.

Schulte im Walde, S. and C. Brew (2002). Inducing German semantic verb classes from purely syntactic subcategorisation information. In Proceedings of ACL 2002.

Sun, L.; A. Korhonen and Y.Krymolowski. (2008a), Automatic Classification of English Verbs Using Rich Syntactic Features. In Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008).

Sun, L.; A. Korhonen and Y. Krymolowski. (2008b). Verb Class Discovery from Rich Syntactic Data. Ninth International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2008).

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 101-108 recibido 30-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Two Approaches to Generate Questions in Basque*

Dos aproximaciones para generar preguntas en euskera

Itziar Aldabe y **Itziar Gonzalez-Dios** y **Iñigo Lopez-Gazpio,**
Ion Madrazo y **Montse Maritxalar**
IXA NLP Group, University of the Basque Country UPV/EHU
Manuel Lardizabal Pasealekua 1 20018 Donostia
{itziar.aldabe, itziar.gonzalezd, ilopez077, jmadrazo003, montse.maritxalar}@ehu.es

Resumen: En este artículo se presenta un generador de preguntas basado en chunks y otro generador basado en dependencias sintácticas. Ambos generan preguntas en euskera a nivel de frase y utilizan el rasgo de animado/inanimado de los nombres, las entidades nombradas y los roles semánticos de los verbos, así como la morfología de los sintagmas nominales. Se describen dos experimentos de generación de preguntas basadas en textos didácticos, en los que una lingüista analiza la gramaticalidad y lo apropiado de las preguntas generadas a partir de frases simples, así como sus correspondientes pronombres interrogativos.
Palabras clave: generación de preguntas, recursos didácticos, rasgos semánticos, morfosintaxis

Abstract: This article presents a chunker-based question generator (QG) and a QG system based on syntactic dependencies. Both systems generate questions in Basque at the sentence level and make use of the animate/inanimate feature of the nouns, named entities, semantic roles of the verbs, and the morphology of the noun phrases. Two experiments to generate questions were carried out based on educational texts. Then, a linguist analysed the grammaticality and appropriateness of the questions generated from single sentences, as well as their interrogative pronoun.
Keywords: question generation, educational resources, semantic features, morphosyntax

1 Motivation

A new community of interdisciplinary researchers[1] have found a common interest in generating questions.[2] The first workshop on the question generation shared task and evaluation challenge (QGSTEC-2008) began the discussion on the fundamental aspects of question generation (QG) and set the stage for future developments in this emerging area. QG is defined (Rus and Graesser, 2009) as the task of automatically generating questions from some form of input, for which the input could vary from raw text to in-depth semantic representation.

Most of the current QG systems are mainly focused on the generation of questions based on single sentences (Rus and Graesser, 2009; Boyer and Piwek, 2010; Graesser et al., 2011). The generation task contains three steps (Rus and Graesser, 2009): content selection, question type selection, and question construction. The content identification and question type selection (i.e. interrogative pronoun) are usually carried out based on various linguistic information. This information is obtained by means of several natural language processing (NLP) tools: syntactic analysers, named entity recognisers, correference resolution systems and semantic role labellers. In contrast, the question construction is usually based on some transformation rules and patterns.

There is few research work on the generation of questions from paragraphs. An

* Acknowledgments: we thank the linguist Ainara Estarrona for her help in the definition of the wh-word lists related to the semantic roles. This research was partially funded by the IRAKURRI project (SPE12UN091), the Ber2Tek project (IE12-333), and the SKaTeR project (TIN2012-38584-C06-02)

[1] Researchers from various disciplines such as cognitive science, computational linguistics, computer science, discourse processing, educational technologies and language generation.

[2] http://www.questiongeneration.org/

ISSN 1135-5948

approach is presented in Mannem, Prasad, and Joshi (2010) where the generation is based on semantic roles of predicates. Agarwal, Shah, and Mannem (2011) also present a system which generates questions based on more than one sentence. For that task, they use discourse connectives. Other researchers address the task of generating the questions from a more pedagogical or psychological point of view. For instance, Mostow and Chen (2009) present a system based on a situation model which is based on characters' mental states. More recently, Olney, Graesser, and Person (2012) generate questions from concept maps based on psychological theories.

Aldabe, Maritxalar, and Soraluze (2011) have probed the viability of the QG task for Basque language. They use a numerical entity recogniser and classifier to detect numerical entities and generate questions about them. Previous to the creation of the questions, the system automatically detects the clauses to be used for the generation. However, the present work deals with the automatic generation of Basque questions using single sentences as the source text for the generation.

This article presents two approaches to generate questions, a chunker-based generation and a generation based on syntactic dependencies. Both approaches created direct questions regarding the noun phrases of the sentences of the corpora. In general, we foresee a better performance when using syntactic dependencies. However, we expect that a chunker-based generation can also be suitable if we limit the source input to single sentences. As the final aim of the system is to be used in the education domain, authors evaluate both approaches whith texts prepared to work on science and technology at secondary school and texts prepared for a language learning scenario. The evaluation focused on how well each approach transforms a sentence into its corresponding interrogative form.

The paper is structured as follows. Section 2 presents the main features used for the generation process. Section 3 describes the implemented systems. Section 4 explains the results of the experiments. Finally, conclusions and future work are commented on section 5.

2 Question Generation

This work presents two question generation systems for Basque. The generation is based on the morphological information of the noun phrases. The systems also use semantic features during the generation process.

2.1 QG based on Noun Phrases

Basque is a Pre-Indo-European language and differs considerably in grammar from the languages spoken in surrounding regions. It is, indeed, an agglutinative head-final isolated language. The case system is ergative-absolutive. The inflections of determination, number and case appear only after the last element in the noun phrase. This last element can be the noun, but also typically an adjective or a determiner. Basque nouns belong to a single declension and its 18 case markers are invariant. Functions, normally fulfilled by prepositions, are realised by case suffixes inside wordforms.

In this work we intend to automatically generate questions about all the noun phrases appearing in each sentence. To that end, the detection of the case markers of the noun phrases is the starting point for the generation process. We report two QG systems in order to compare a chunker-based approach and a dependency-based approach.

We choose 5 case markers as the starting point for the experiments: absolutive (ABS), ergative (ERG), inessive (INE), allative (ALL) and ablative (ABL). As explained in section 4.1, all of them cover almost the 90% of all the noun phrases found in corpora when generating questions in our scenario. Absolutive and ergative cases accumulate the highest percentage of noun phrases in the corpus, as they are related to the subject and direct object syntactic functions. The inessive case is used in noun phrases with different adverbial functions (temporal, location, etc). And the ablative and allative cases give us the chance to work with the animate/inanimate features. The mentioned 5 cases need different wh-words (interrogative pronouns) depending on the features of the head in the noun phrase.

2.2 Semantic Features for QG

We have explored the animate/inanimate feature of the nouns, the use of named entities (person, location and organisation) and the semantic roles to deal with the generation of

questions.

- Animate/inanimate feature: the QG generators use the work done by Díaz de Ilarraza, Mayor, and Sarasola (2002), where semantic features of common nouns are extracted semi-automatically from a monolingual dictionary. Both systems consider the animate/inanimate feature of 15,000 nouns.
- Named entities: both QG systems include a named entity recogniser and classifier named *Eihera* (Alegria et al., 2003). They use this tool to identify person, place and organisation entities.
- Semantic roles: The QG generators take into account a corpus manually tagged at the predicate level with verb senses, argument structure and semantic roles (Aldezabal et al., 2010). This corpus is based on the work done in Aldezabal (2004), which includes an in-depth study of 100 verbs for Basque. Based on the occurrences of the 100 verbs, we have worked with the following roles from VerbNet (Kipper et al., 2006): actor, attribute, agent, beneficiary, cause, destination, direction, experiencer, extent, instrument, location, manner, patient, predicate, product, recipient, source, theme, temporal and topic. The mandatory roles of patterns with a probability higher than 75% are considered to be candidates.

3 QG-Malti and QG-Ixati

The article reports two question generation systems for Basque, QG-Malti and QG-Ixati. QG-Malti is a QG system based on *Maltixa* (Bengoetxea and Gojenola, 2010), a dependency parser for Basque. QG-Ixati is a QG system which uses *Ixati* (Aduriz et al., 2004), a chunker for Basque. Previous to the selection of the noun phrase (the content selection step), both systems perform a morphosyntactic analysis of the source texts.

As proposed in Rus and Graesser (2009), both QG systems can be described as a three-step process: content selection, question type selection and question construction. The main constraint on the present study is that the systems only select sentences which contain a single finite verb. Before the content selection process, QG-Malti also splits coordinate sentences into single sentences.

The goal of both approaches is to generate questions at sentence level. QG-Malti discards the sentences which have discourse elements whose function is to connect the sentence with other elements outside the sentence, but it rejects them only in case the discourse elements are not at the beginning of the sentence. QG-Ixati, however, can not discard this kind of sentences as the analyser Ixati does not detect this kind of discourse relations.

3.1 Target Selection

As mentioned, both systems generate questions related to all the noun phrases that occur in the sentences of the source text[3]. The generation process uses morphosyntactic features of the output of the corresponding analyser to select **the candidate target**. In the case of the QG-Ixati, the candidate target is the whole noun phrase (chunk). However, in the case of QG-Malti the candidate target is the word whose morphological analysis has the target case marker. And then, the dependency structure of the analysis is used to construct the corresponding noun phrase.

When there is more than one occurrence for the same case marker inside the same sentence, only one of those occurrences is used to generate a question. Based on the fact that in Basque the relevant information of a sentence is close to the verb, QG-Malti selects as the candidate target (word) the one which is closest to the verb. And, if there are two candidates at the same distance to the verb, it selects the one located on the left to the verb. The reason for this criterion is that in Basque the informationally relevant phrase of a sentence precedes immediately the verb.

QG-Ixati, however, establishes a preference criterion based on various semantic features of the candidate targets (noun phrases). It gives a higher priority to the animate/inanimate feature and named entity tag than to the semantic roles. The priority is obtained as follows:

1. If the head of the noun phrase is a named entity (person, place or organisation) or its animate/inanimate feature is known, the QG system establishes a weight of 2 for the noun phrase.
2. If the noun phrase fulfills one of the

[3]We use 5 declension cases in the experiments of the present work.

	Animate	Person	Inanimate	Place	Organisation	No semantic feature
ABS	*Nor*		*Zer*			*Nor/Zer*
ERG	*Nork*		*Zerk*			*Nork/Zerk*
INE	*Norengan*		*Non*			*Norengan/Non/Noiz*
ALL	*Norengandik*		*Nondik*			*Norengandik/Nondik*
ABL	*Norengana*		*Nora*			*Norengana/Nora*

Table 1: Question type based on named entities, animate/inanimate and case markers. Nor (Who-ABS); Zer (What-ABS); Nork (Who-ERG); Zerk (What-ERG); Norengan (To whom); Non (Where); Noiz (When); Norengandik (From whom); Nondik (From where); Norengana (To whom); Nora (To where)

mandatory roles of a particular verb subcategorization pattern, the QG system establishes a weight of 1 for the given noun phrase.

The system chooses the noun phrase with the highest priority. In the cases that the system still assigns the same weight to different noun phrases, the selected candidate is the one which is closest to the verb. And in case of still being a tie, the system chooses the phrase located on the left to the verb.

3.2 Question Type Selection

QG-Malti and QG-Ixati follow the same criteria when selecting the question type to be generated. The selection of the question type is based on the linguistic information of the corresponding candidate target. For each case marker and linguistic feature (animate/inanimate, named entity, semantic role and morphology), an expert in the field established the most probable question type (wh-word) based on linguistic studies, as well as on her experience.

Table 1 shows the question types selected by the QG systems related to the named entity, animate/inanimate feature and case marker of the candidate target. For example, if the head of the noun phrase is identified as a person named entity and its corresponding case marker is the absolutive, the *NOR* (Who-ABS[4]) wh-word is selected.

The question type is also selected based on the semantic role of the candidate target. In total, 11 different roles have been linked to targets with the absolutive case, 7 to the ergative, 8 to the inessive, 5 to the ablative, and 5 to the allative. Depending on the semantic role of the candidate target, the QG system establishes its corresponding wh-word. For each verb, mainly only one question type is linked to each role. But, there are some exceptions, for example, the verb *compare* can have an animate or inanimate *patient* that correspond to the *NOR* (Who-ABS) and *ZER* (What-ABS) question types respectively.

3.3 Question Construction

In this phase, each QG system applies its own strategy based on the information given by the corresponding analyser. QG-Malti constructs the questions using the dependency relation structure analysed in the source sentence. QG-Ixati uses the information of the chunks detected during the morphosyntactic analysis of the source sentence.

The question building is based on simple transformation rules defined in the system. The first element of the constructed question is the wh-word. Following the wh-word, the main verb is established. Then, the rest of the elements (dependency structures or chunks) that are to the left of the verb in the source sentence are added to the question. Finally, the elements that appeared on the right of the source sentence's verb are appended to the generated question.

During the development of the systems we realised that some discourse connectives (e.g. the connective *gainera* [5]) caused some noise to the generated questions. In most of the cases where the connective was at the beginning of the source sentence, such a noise could be avoided if the connective was deleted when constructing the question. That is why we decided to delete from the source sentences all the discourse connectives which appear at the beginning of the sentence.

[4]The ABS mark refers to the fact that the wh-word takes the absolutive case marker.

[5]Basque word for *in addition*

4 Evaluation

For the experiments, we chose texts about science and technology for secondary school learners and a specialised corpus in language learning because one of the final aims is to use QG systems into the education domain. In this work, as a first step, the evaluation focused on how well each QG system transforms a sentence into its corresponding interrogative form.

We focused on the evaluation of the syntactic correctness and fluency of the generated questions. To do so, a human judge followed the same classification as the one proposed in Boyer and Piwek (2010). We also studied the quality of the question types determining whether the generated wh-words asked about the source sentence. Finally, the expert also established whether the question was appropriate in relation to the source sentence.

4.1 Datasets

The science and technology (ST) dataset is composed of 5 texts about science and technology. One expert who works on the generation of learning materials defined the 5 texts as adequate for secondary school learners (Aldabe and Maritxalar, 2010). The main topics of these texts were: Continent; the Earth; Bats; the Arctic; and Computers respectively. All the texts have a similar length. In total, the dataset contains 176 sentences, being the average length of a sentence 13 words.

The language learning (LL) dataset focuses on a specialised corpus for Basque language learning, which is a collection of learning-oriented Basque written texts. The corpus is classified into different language levels[6] in accordance with the Common European Framework of Reference for Languages (Little, 2011). In the present work, the intermediate level of the corpus is the basis to generate the questions. The corpus is composed of near 80,000 sentences (over one million words), and the average length of a sentence is 13 words.

The ST dataset contains 646 noun phrases and the LL dataset has 200,000 noun phrases. Looking at the 5 case markers that are the starting point of the systems to generate the questions, almost 90% of the noun phrases are covered with the mentioned target case markers in both datasets. In the ST dataset, 55% of the noun phrases have the absolutive case marker, the 12% of the phrases have the ergative case marker, the 16% of phrases are inessive, the 3% of noun phrases have the allative case and the 2% of them the ablative case. In the LL dataset, the 60% have the absolutive case marker, the 11% of the phrases have the ergative case marker, and the 16% of phrases are inessive. Regarding the allative and ablative cases the percentage is near the 3%. The rest of case markers are under the 4% in both datasets.

4.2 Experiments

For each dataset, experiments with both QG systems were performed. The questions generated by QG-Malti and QG-Ixati were manually evaluated at different levels by one linguist.

As regards the question-types, a linguist judged whether the generated wh-words asked about the source sentence (yes/no). For that, the source sentence (input for the QG system) and the candidate target (answer to the generated question) were provided.

When checking the grammaticality, the linguist evaluated the syntactic correctness and fluency of the generated questions. For that, only the generated questions were provided. The questions were classified and differentiated among: i) correct questions; ii) questions which need minor changes (punctuation, capitalization, spelling or dialectical variants); iii) questions with major changes that are unnatural for native speakers even they are grammatically correct; and iv) incorrect questions due to the grammar, including oral speech style.

Finally, the judge established if the generated questions were appropriate (yes/no). For that, in addition to each question, the corresponding answer was also shown. When evaluating the appropriateness of the questions only correct questions and questions which nedded minor changes were considered.

4.2.1 ST dataset experiment

The experiment with the ST dataset reflects an educational scenario where the creation of updated material using texts from the web is crucial for the motivation of learners and

[6] Although the language level of a text can be a controversial aspect because it is difficult to define, in our source corpus, expert teachers classified the texts into specific levels.

	ST-common			ST-divergent		
	Wh-word	**Grammar**	**Appropr.**	**Wh-word**	**Grammar**	**Appropr.**
QG-Malti	76%	54%	46%	72.6%	59.7%	41.9%
QG-Ixati	88%	66%	64%	87.1%	48.4%	48.4%

Table 2: Results for the ST common and divergent inputs

teachers. Both systems generated questions for all the candidate targets of the 5 texts. Based on the 5 case markers, QG-Malti and QG-Ixati generated 112 and 81 questions respectively.

	Wh-word	**Grammar**	**Appropr.**
QG-Malti	75.0%	57.1%	44.6%
QG-Ixati	87.6%	59.3%	58.0%

Table 3: Percentage of correct questions of the QG systems for the ST dataset

Table 3 presents the evaluation results as regards wh-words, grammaticality and appropriateness in the ST dataset. The grammar column groups questions marked as correct and questions which need minor changes. In general, QG-Ixati obtains better results than QG-Malti, but, QG-Malti generates more questions. Thus, QG-Malti generates 64 grammatically correct questions while QG-Ixati generates 48.

The generation processes of QG-Malti and QG-Ixati differ mainly due to the analysers and the target selection criteria. However, both systems have in common some instances. We refer to common instances to those which have the same candidate target with the same case marker. Out of the 112 and 81 generated questions both systems have in common 50 instances. Thus, apart from the these common instances, QG-Malti selects 62 sentences to generate the questions, while QG-Ixati chooses other 31 different ones. Table 2 presents the manual evaluation results based on this distinction. As regards the common instances (ST-common column), QG-Ixati obtains better results in terms of wh-words (88%), grammaticality (66%) and appropriateness (64%). The comparison of the divergent samples (ST-divergent column) with the common instances of each system shows different results. On the one hand, it is remarkable the improvement of the grammaticality of QG-Malti (59.7%) compared to its common instances (54%). On the other hand, QG-Ixati obtains worse results in terms of grammaticality (48.4%) and appropriateness (48.4%), compared to the common instances (66% and 64% respectively).

Thus, even the overall results are better for QG-Ixati, the number of grammatically correct questions of the divergent dataset is higher in the case of QG-Malti. These results must be analysed deeply as we foresee that the target case markers and the used analysers can have an influence on the results.

4.2.2 LL dataset experiment

The aim of the LL dataset experiment is to analyse the influence of the case marker of the noun phrase chosen as the answer to the generated question. This is why the sample contains 20 questions per case marker for each QG system selected at random[7]. In this experiment, a total of 100 generated questions for each system are evaluated.

Table 4 shows the evaluation results per case marker. In general, both systems obtain grammatically better questions when the generation is based on noun phrases with absolutive or inessive case markers. QG-Ixati obtains better overall results compared to QG-Malti. It is noticeable the difference on the grammaticality of the absolutive (QG-Ixati, 85% and QG-Malti, 60%) and ergative (QG-Ixati, 65% and QG-Malti, 45%) case markers. In contrast, QG-Malti performs better in terms of grammaticality and appropriateness of the allative and ablative case markers, and the wh-word of the ergative.

Although the source sentences are the same for both systems, the systems sometimes differ in the source candidate targets for the generation process. Out of the 100 questions, both systems have in common 47 questions. Table 5 presents the results of the 100 questions (LL-overall column) and the 47 common questions (LL-common column) in terms of wh-word, grammaticality and appropriateness.

The grammatically is better for QG-Malti

[7]The source sentences were the same for both QG systems.

	QG-Malti			QG-Ixati		
	Wh-word	**Grammar**	**Appropr.**	**Wh-word**	**Grammar**	**Appropr.**
ABS	70%	60%	55%	85%	85%	60%
ERG	95%	45%	40%	80%	65%	60%
INE	60%	85%	60%	70%	85%	60%
ALL	70%	45%	35%	85%	35%	30%
ABL	50%	50%	45%	55%	40%	35%

Table 4: Percentages per case markers (20 questions per case marker)

	LL-overall			LL-common		
	Wh-word	**Grammar**	**Appropr.**	**Wh-word**	**Grammar**	**Appropr.**
QG-Malti	69%	57%	47%	70.2%	63.8%	57.4%
QG-Ixati	75%	62%	50%	68.1%	59.6%	48.9%

Table 5: Results for the LL overall and LL common inputs

when looking at the common instances (from 57% to 63.8%) and it is lower for QG-Ixati (from 62% to 59.6%). Looking at the case markers of the 47 questions, just 3 out of the 47 questions correspond to the absolutive noun phrases and this is the main reason for getting worst results when using QG-Ixati.

4.3 Preliminary Error Analysis

The analysis of the results as well as the subsequent meetings with the expert allowed us to carry out a preliminary error analysis of the systems.

As regards the grammatical correctness of the questions, we have classified the erroneous questions in different categories: (i) questions which are grammatically correct but unnatural as regards the speakers; (ii) questions which contain orthographic errors; (iii) questions which are incorrectly generated in terms of morphology; (iv) questions which refer to oral speech; (v) problems with punctuation marks; and (vi) questions with an incorrect word order.

One of the reasons to generate ungrammatical questions is due to the type of the source input. In the analysis of the results we detected: i) some source input that correspond to subordinate clauses; ii) some source input that correspond to relative clauses and iii) some typos or spelling errors at word level. When analysing the results without taking into account the mentioned questions, the grammaticality and appropriateness measures of both QG systems improve 6 points for the ST dataset, and more than 10 points for the the LL dataset. In contrast, the number of correct wh-words hardly varies.

5 Conclusions and Future Work

Our QG systems created questions in order to ask about noun phrases at sentence level. With that end, a chunker-based QG system, QG-Ixati, and a QG system based on dependency structures, QG-Malti, have been implemented. Both systems deal with Basque language and make use of the animate/inanimate feature of the nouns, named entities (person, location and place), the semantic roles of the verbs, as well as the morphology of the noun phrases.

The results of the experiments show that QG-Malti generates a higher number of questions in a real scenario (ST dataset), however its general performance is slightly worse than QG-Ixati. The results for the LL dataset show a noticeable difference in grammaticality between both systems when generating questions about noun phrases with the absolutive and ergative case markers.

Future work will focus on the improvement of the systems. Once the roles are detected automatically, the semantic role approach would cover more verbs. Thus, we plan to focus on the analysis and integration of new Basque NLP tools or knowledge representations in order to generate questions that require deeper understanding. In addition, in the case of QG-Malti we want to improve the system using the information about syntactic dependencies, to discard ungrammatical source input for the generator, and to improve the results of the identification of the question type.

Bibliography

Aduriz, Itziar, María Jesús Aranzabe, Joxe Mari Arriola, Arantza Díaz de Ilarraza, Koldo Gojenola, Maite Oronoz, and Larraitz Uria. 2004. A cascaded syntactic analyser for Basque. *Computational Linguistics and Intelligent Text Processing*, pages 124–134.

Agarwal, Manish, Rakshit Shah, and Prashanth Mannem. 2011. Automatic question generation using discourse cues. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9. Association for Computational Linguistics.

Aldabe, Itziar and Montse Maritxalar. 2010. Automatic distractor generation for domain specific texts. In *Proceedings of the 7th International Conference on NLP, IceTAL 2010*. Springer, pages 27–38.

Aldabe, Itziar, Montse Maritxalar, and Ander Soraluze. 2011. Question generation based on numerical entities in Basque. In *Proceedings of AAAI Symposium on Question Generation*, pages 2–8.

Aldezabal, Izaskun. 2004. *Aditazpikategorizazioaren Azterketa Sintaxi Partzialetik Sintaxi Osorako bidean. 100 aditzen azterketa, Levin-en (1993) lana oinarri hartuta eta metodo automatikoak baliatuz.* Ph.D. thesis, Euskal Filologia Saila. UPV/EHU.

Aldezabal, Izaskun, María Jesús Aranzabe, Arantza Díaz de Ilarraza, Ainara Estarrona, and Larraitz Uria. 2010. EusPropBank: Integrating semantic information in the Basque dependency treebank. In *Computational Linguistics and Intelligent Text Processing*. Springer, pages 60–73.

Alegria, Iñaki, Nerea Ezeiza, Izaskun Fernandez, and Ruben Urizar. 2003. Named Entity Recognition and Classification for texts in Basque. In *II Jornadas de Tratamiento y Recuperación de Información, JOTRI, Madrid.*

Bengoetxea, Kepa and Koldo Gojenola. 2010. Application of different techniques to dependency parsing of Basque. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 31–39. Association for Computational Linguistics.

Boyer, Kristy Elizabeth and Paul Piwek, editors. 2010. *Proceedings of QG2010: The Third Workshop on Question Generation*. Pittsburgh: questiongeneration.org.

Díaz de Ilarraza, Arantza, Aingeru Mayor, and Kepa Sarasola. 2002. Semiautomatic labelling of semantic features. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Graesser, Arthur, James Lester, Jack Mostow, Rashmi Prasad, and Svetlana Stoyanchev. 2011. Question generation papers from the AAAI fall symposium. Technical report, FS-11-04.

Kipper, Karin, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. In *In Proceedings of 5th international conference on Language Resources and Evaluation*.

Little, David. 2011. The common european framework of reference for languages: A research agenda. *Language Teaching*, 44(03):381–393.

Mannem, Prashanth, Rashmi Prasad, and Aravind Joshi. 2010. Question generation from paragraphs at UPenn: QGSTEC system description. In *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 84–91.

Mostow, Jack and Wei Chen. 2009. Generating instruction automatically for the reading strategy of self-questioning. In *Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*, pages 465–472. IOS Press.

Olney, Andrew M, Arthur C Graesser, and Natalie K Person. 2012. Question generation from concept maps. *Dialogue & Discourse*, 3(2):75–99.

Rus, Vasile and Arthur C. Graesser, editors. 2009. *The Question Generation Shared Task and Evaluation Challenge*.

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 109-116 recibido 19-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Prueba de Concepto de Expansión de Consultas basada en Ontologías de Dominio Financiero

Proof of Concept of Ontology-based Query Expansion on Financial Domain

Julián Moreno Schneider
Grupo de Bases de Datos Avanzadas
Departamento de Informática
Universidad Carlos III de Madrid
Avda. Universidad, 30, 28911
Leganés, Madrid, España
jmschnei@inf.uc3m.es

Thierry Declerck
Language Technology Lab
German Research Center for Artificial Intelligence, DFKI GmbH
Stuhlsatzenhausweg, 3, 66123
Saarbrücken, Alemania
declerck@dfki.de

José Luis Martínez Fernández
DAEDALUS – Data, Decisions and Language S.A.
Avda. de la Albufera 321, 28031
Madrid, Spain
jmartinez@daedalus.es

Paloma Martínez Fernández
Grupo de Bases de Datos Avanzadas
Departamento de Informática
Universidad Carlos III de Madrid
Avda. Universidad, 30, 28911
Leganés, Madrid, España
pmf@inf.uc3m.es

Resumen: Este trabajo presenta el uso de una ontología en el dominio financiero para la expansión de consultas con el fin de mejorar los resultados de un sistema de recuperación de información (RI) financiera. Este sistema está compuesto por una ontología y un índice de Lucene que permite recuperación de conceptos identificados mediante procesamiento de lenguaje natural. Se ha llevado a cabo una evaluación con un conjunto limitado de consultas y los resultados indican que la ambigüedad sigue siendo un problema al expandir la consulta. En ocasiones, la elección de las entidades adecuadas a la hora de expandir las consultas (filtrando por sector, empresa, etc.) permite resolver esa ambigüedad.
Palabras clave: Búsqueda Semántica, Recuperación de Información, Ontología, Expansión de Consulta

Abstract: This paper explains the application of ontologies in financial domains to a query expansion process. The final goal is to improve financial information retrieval effectiveness. The system is composed of an ontology and a Lucene index that stores and retrieves natural language concepts. An initial evaluation with a limited number of queries has been performed. Obtained results show that ambiguity remains a problem when expanding a query. The filtering of entities in the expansion process by selecting only companies or references to markets helps in the reduction of ambiguity.
Keywords: Semantic Search, Information Retrieval, Ontology, Query Expansion

1 Introducción

La cantidad de información semántica existente en formato electrónico es cada vez mayor. Cualquier consulta que se realice para acceder a estos vastos volúmenes de información semántica proveerá grandes cantidades de resultados, imposibles de procesar manualmente. Por otro lado, si dicha consulta no se ha planteado adecuadamente, la respuesta puede no ser la esperada. Esto nos lleva a pensar que la utilización de técnicas externas que ayuden al usuario a concretar o enriquecer la consulta realizada pueden ayudar a gestionar grandes cantidades de información.

ISSN 1135-5948

En este sentido, la utilización de elementos que aporten información semántica a la consulta pueden ayudar a completar su formulación para obtener mayor cantidad de información a la vez que más diversa. Incluso puede llegar a resolver problemas de ambigüedad Martínez-Fernández et al. (2008).

El dominio contemplado en esta investigación es el financiero, concretamente noticias sobre economía e información de mercados bursátiles. Este dominio presenta la particularidad de que es altamente cambiante, lo que complica la recuperación de información semántica y requiere además que la información recuperada sea actual y precisa si se va a utilizar en procesos de toma de decisiones.

Como ejemplo intuitivo podríamos mencionar a una persona que quiere invertir en bolsa y ha oído que Inditex es una buena opción. Si el usuario busca el nombre de esta empresa en cualquier sistema de RI obtendrá documentos en los que se mencione la propia empresa. Aunque esta puede ser información interesante para el usuario, está claro que no es suficiente. El interés que tiene el usuario en esta empresa podría extenderse a sus filiales, empresas asociadas, información relacionada con sus directivos o con empresas del mismo sector.

El usuario no tiene todos esos datos antes de realizar la consulta, por lo que ayudarle a enriquecer su consulta puede ser una buena forma de mejorar la calidad del proceso de recuperación.

2 Trabajo Relacionado

La expansión de consultas es un tema que ha sido ampliamente estudiado en diversos dominios como describen Bhogal, Macfarlane y Smith (2007). Carstens (2011) describe tres modos de realizar la expansión de una consulta (EC): manual, automático e interactivo. La EC manual es aquella que realiza el usuario directamente; en la EC interactiva es el sistema propone una expansión que el usuario debe confirmar, mientras que en la automática el sistema no necesita confirmación. Este último tipo de expansión es el que vamos a describir en mayor profundidad.

Carpineto et al. (2012) realizan un estudio en profundidad de la expansión automática de consultas. Mencionan aplicaciones como la búsqueda de respuestas (en inglés, *question answering*) o la recuperación de información en entornos multilingües y describen las técnicas empleadas en la expansión, desde la utilización de ontologías hasta el análisis de *logs* de consultas.

Dentro de la investigación de expansión automática, se debe distinguir entre trabajos que realizan la expansión de la consulta con información obtenida de la colección de documentos y aquellos que expanden la consulta con información externa.

Enmarcado en el primer tipo, Dragoni, Pereira y Tettamanzi (2012) describen un sistema de RI basado en ontologías, pero sin hacer una expansión de la consulta. En su lugar plantean la extracción de los conceptos de la ontología presentes en la consulta y en los documentos y calculan la concordancia en función de estos conceptos. Wollersheim y Rahayu (2005) presentan un trabajo similar

En lo que se refiere a trabajos independientes de la colección, la propuesta de Díaz-Galiano, Martín-Valdivia y Ureña-López (2009) utiliza una ontología para expandir consultas en el dominio biomédico y trabaja sobre un sistema multimodal. La ontología ayuda a expandir la consulta con términos relacionados que aparecen en otros documentos médicos. Farhoodi et al. (2009) describen un trabajo similar, en el que la expansión se realiza utilizando una ontología basada en Wikipedia. Intentan focalizar las necesidades de información del usuario haciendo desambiguación a la vez que expanden la consulta. Utilizan un sistema de vectores para expandir la consulta si algún término de esta se encuentra en la ontología. Utilizan Google como sistema de RI para evaluar la expansión.

Otros trabajos que realizan expansión de la consulta son: Schweighofer y Geist (2007) trabajan en el dominio legal y utilizan una ontología para adecuar las consultas a la terminología legal. Emplean la ontología para añadir términos semánticamente similares; Wu, Wu y Fu (2007) procesan términos de una consulta en función de un razonador lógico y de la frecuencia de aparición de los términos, para posteriormente expandirlos contra una ontología; y, por su parte, Liu y Li (2011) hacen expansión de consultas basándose en una ontología de tres capas.

Mandala, Tokunaga y Tanaka (2000) realizan una expansión de consulta basada en tesauros, utilizando una combinación de la relevancia de cada término de la consulta en

cada tesauro del que disponen para determinar si el término se expandirá o no.

El trabajo más cercano al presentado en este artículo es el de Dey et al (2005). En él se utilizan ontologías para mejorar las expresiones de búsqueda, seleccionando primero los términos de la consulta que son adecuados para ser expandidos y posteriormente realizando la expansión utilizando la ontología.

El trabajo de Fu, Jones y Abdelmoty (2005) realiza expansión de la información geográfica de la consulta en función de una ontología específica de dominio y una ontología geográfica. La expansión de la consulta en la recuperación de imágenes se analiza y se describe en el trabajo (Gulati y Sharma, 2010). Las técnicas de expansión de consulta son similares pero la consulta posteriormente se aplica a recuperación de imágenes en lugar de textos.

El sistema (Tuominen et al, 2009) realiza una expansión de consulta textual basada en las ontologías publicadas en el servicio ONKI Ontology Service[1]. Lo destacable de este sistema no es su funcionamiento sino que es un *widget* que se puede integrar en cualquier sistema, como páginas web o sistemas de recuperación. Ha sido evaluado con ontologías específicas de dominio y generales.

Para finalizar, en CLEF[2] también se han hecho muchos experimentos sobre expansión de consultas. En concreto, (Martínez-Fernández et al, 2008) concluye que si no se resuelve la ambigüedad de algún modo, la expansión no funciona (introduce más ruido) y empeora los resultados de precisión y cobertura.

En resumen, los trabajos actuales han utilizado ampliamente ontologías para realizar expansión de consultas y han demostrado ser una herramienta adecuada para ello. Por eso, este trabajo persigue explotar esta técnica para aplicarla al dominio financiero.

3 Propuesta

Este trabajo plantea el uso de una ontología para mejorar los resultados de la búsqueda en un dominio particular.

La Figura 1 muestra la arquitectura completa de la propuesta que consta de dos partes: una parte de expansión de la consulta y otra de acceso a una ontología.

[1] http://onki.fi/
[2] http://www.clef-initiative.eu//

La búsqueda basada en ontología se describe cómo la obtención de información semánticamente relacionada con una consulta planteada. Considerando un funcionamiento típico de la ontología, una búsqueda debería tomar como punto de partida un concepto (individuo o clase) de la propia ontología. En este trabajo se ha utilizado una versión más amplia y se ha implementado una búsqueda textual sobre la ontología, permitiendo la obtención de conceptos (individuos y clases) de la ontología en función de una búsqueda expresada en lenguaje natural.

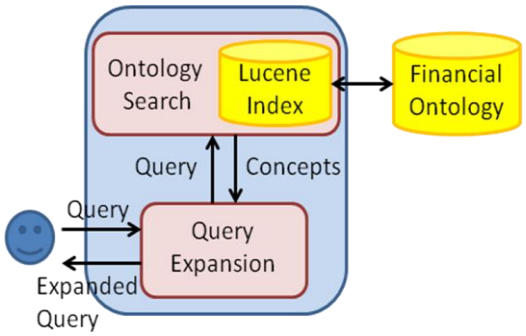

Figura 1. Arquitectura de la propuesta

Aunque no se cubre en este trabajo, esta estructura favorece la recuperación multilingüe, tal como explican Declerck y Gromann (2012). Esto se debe, entre otras cosas, a que si la ontología es multilingüe, almacena relaciones entre conceptos de distintos idiomas sin ambigüedad. Por ejemplo, *salt* en inglés estará relacionada con el concepto *sal (condimento)* en castellano y no con la forma imperativa del verbo *salir*.

3.1 Descripción de la Ontología de Doble Capa

La ontología desarrollada en este trabajo se centra en el dominio financiero y para ser más concretos, en mercados bursátiles y en las empresas que operan en estos mercados. El formato de esta ontología está basado en el utilizado por Liu y Fi (2011) aunque sin dividir el contenido en varias ontologías. Únicamente disponemos de una ontología en la que podemos identificar dos 'capas': la primera que relaciona todas las entidades presentes en un mercado bursátil y la segunda que asigna metadatos a cada una de estas entidades.

La primera de las capas está compuesta por cuatro clases diferentes:

- Mercados: representación semántica de los mercados bursátiles. Contiene empresas que operan en ellos.
- Sectores: elemento perteneciente a la clasificación de los diferentes ámbitos en los que las empresas pueden operar.
- Empresas: empresa que opera en mercados bursátiles y en sectores. Está relacionada con las personas que trabajan en ella (junta directiva, presidentes, etc.).
- Personas: personas físicas que forman parte o representan a las empresas.

Además de las propias clases, las relaciones semánticas entre ellas (representadas por propiedades de objeto) también se engloban en la primera capa. Estas relaciones son las que ofrecen una completitud semántica a las entidades. La lista completa de las relaciones existentes se muestra en la Tabla 1. El dominio y el rango pueden tomar los siguientes valores: Mercado (M), Empresa (E), Sector (S) o Persona (P).

La segunda capa es la que aporta información complementaria a cada una de las entidades. Esto se realiza a través de la asignación de anotaciones a las entidades. Estas anotaciones son de dos tipos (nombre y descripción) y pueden aparecer en cualquier idioma. En nuestro caso nos hemos limitado a español e inglés. Como ejemplo se muestran las anotaciones del concepto *Indra* en formato OWL en la Tabla 1.

Propiedad	**Dominio**	**Rango**
hasCeo	E	P
isParticipatingIn	E	M
hasActivityIn	E	S
hasCompany	S	E
hasCompanyIn	M	E
hasPresident	E	P
isCEOIn	P	E
isPresidentOf	P	E
isSubSectorOf	S	S
isSuperSectorOf	S	S

Tabla 1: Propiedades de Objetos de la Ontología

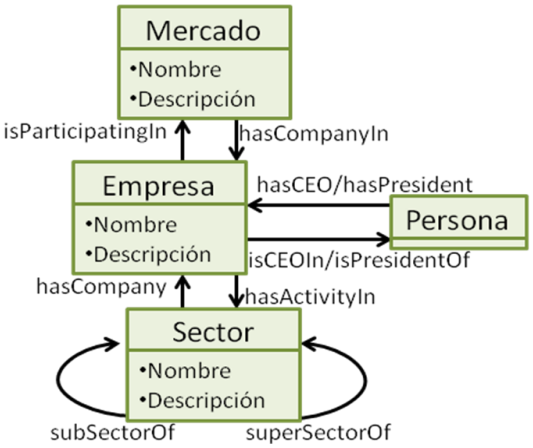

Figura 2. Diagrama de la Ontología

Empresa
Nombre (esp.): INDRA SISTEMAS, S.A.,SERIE A
Nombre (ing.): INDRA SISTEMAS, S.A.,SERIE A
Descripción (español): Indra es la multinacional de tecnologías de la Información número 1 en España y una de las principales de Europa y Latinoamérica. Es la segunda compañía europea por capitalización bursátil de su sector y es también la segunda empresa española que más invierte en I+D. En 2008 sus ventas alcanzaron los 2.380 M , de los que un tercio procedieron del mercado internacional. Cuenta con más de 29.000 profesionales y con clientes en más de 100 países. Indra es líder en soluciones y servicios de alto valor añadido para los sectores de Seguridad y Defensa, Transporte y Tráfico, energía e Industria, Servicios Financieros, Sanidad y Administraciones Públicas, Telecom y Media.

Sector
Nombre (esp.): Tecnología y telecomunicaciones
Nombre (ing.): Technology and telecommunication
Descripción (español): Este sector engloba aquellas actividades relacionadas con las telecomunicaciones tales como la telefonía (tanto básica como móvil), y el diseño, instalación, gestión y mantenimiento de redes e infraestructura de comunicaciones. Además, se incluyen todas aquellas actividades de electrónica y software así como las empresas dedicadas a la fabricación y distribución de hardware tecnológico y equipamiento.

Tabla 2: Ejemplo de metadatos asociados a empresa y sector.

3.2 Construcción de la Ontología

La ontología ha sido creada automáticamente a partir de un *web crawler* sobre la página de la Bolsa de Madrid[3].

Mediante este *crawler* se obtuvo tanto información de las empresas como de los sectores. En total se obtuvieron 41 sectores que están agrupados en 6 sectores generales que a su vez se dividen en subsectores. En cuanto a las empresas, se han obtenido 34 empresas que cotizan en el mercado bursátil de Madrid.

La información que se obtuvo de las empresas está compuesta por: nombre, descripción, dirección, últimas cotizaciones, información bursátil, etc. Para la aplicación que se ha desarrollado, solo se ha utilizado el nombre y la descripción.

Con respecto a los sectores, que también fueron obtenidos con un *web crawler* sobre la página de la Bolsa de Madrid, se obtuvo el nombre de cada sector y su descripción (ambos en español y en inglés). Además, también se almacenó la jerarquía de sectores, para poder determinar aquellos que pertenecen a otros. En el caso de los sectores, toda la información obtenida (en español) se utiliza en la ontología.

Una vez que se dispone de toda esta información, es necesario poblar la ontología. Los sectores tienen distintas realizaciones lingüísticas en las descripciones de las empresas. Por ejemplo, en el caso de *Bankia* el identificador del sector es *"Serv.Financieros e Inmob. - Bancos/Cajas Ahorro"*, mientras que el nombre del sector que le corresponde es *"Servicios financieros e inmobiliarias"*. Por este motivo se ha aplicado una normalización para hacer corresponder la etiqueta de sector asociada a la empresa con su correspondiente sector en la clasificación.

La distancia utilizada para comparar los nombres es la distancia de Jaro-Winckler (Wincler, 1990). Se valoró la utilización de la distancia de Levenshtein (Levenshtein, 1965), pero dado que los sectores en las empresas aparecían con abreviaturas, se descartó esa posibilidad.

3.3 Búsqueda de Lenguaje Natural

El sistema de búsqueda en lenguaje natural se basa en la utilización conjunta de la ontología y un índice de Lucene[4] creado sobre la misma. Este índice tiene tres campos: el nombre de los conceptos (individuos), su descripción y la URL que los identifica. La consulta textual se lanza sobre el nombre del concepto y su descripción y se recuperan las URLs de los conceptos como identificador semántico. El peso que se otorga al consultar cada campo del índice es el mismo. Tanto para la indexación de las descripciones como para la consulta se utiliza el analizador Snowball[5] de Lucene. Además, el peso que se otorga a todos los campos del índice es el mismo.

El resultado obtenido de consultar este índice son los identificadores únicos (URLs) de los conceptos almacenados en la ontología.

3.4 Enriquecimiento de Consulta

El enriquecimiento de la consulta se realiza con los nombres de los conceptos que devuelve la búsqueda en lenguaje natural sobre la ontología.

La expansión de la consulta se ha abordado desde dos puntos de vista diferentes: lanzando la consulta completa del usuario a la búsqueda de lenguaje natural y el análisis semántico de la consulta para expandir únicamente las entidades.

El análisis semántico de la consulta se ha realizado utilizando Textalytics[6], un conjunto de APIs de procesamiento lingüístico en distintos idiomas, incluyendo español e inglés. Esta herramienta ha permitido el reconocimiento de las entidades de la consulta, obteniendo además etiquetas semánticas que permiten distinguir el tipo de entidad, es decir, si se trata de una persona, una organización, un lugar, etc.

En ambos casos, la consulta planteada por el usuario se mantiene invariable mientras que la información extraída de la ontología se añade directamente al final de la propia consulta. La información obtenida puede ser de cinco tipos diferentes: (i) nombre de la empresa (+Nom); (ii) texto completo de la descripción de la empresa; (iii) nombre de la empresa y texto completo de la descripción de la empresa; (iv) entidades extraídas de la descripción de la empresa;(+Ents.Descr.) y (v) nombre de la empresa y entidades extraídas de la descripción de la empresa.

[3] http://www.bolsamadrid.es

[4] http://lucene.apache.org/core/

[5] http://lucene.apache.org/core/old_versioned_docs/versions/3_5_0/api/all/org/apache/lucene/analysis/snowball/SnowballAnalyzer.html

[6] http://www.textalytics.com

En una segunda aproximación, se ha utilizado la ontología para extraer información relacionada (sectores en los que operan las empresas) con los conceptos recuperados mediante la búsqueda textual. Esto añade dos tipos nuevos de información a añadir: (vi) nombre del sector en el que operan las empresas (+Sect.); (vii) nombre de la empresa y del sector; y permite añadir el sector en los tipos (iii) y (v) previamente definidos.

4 Evaluación preliminar

Este sistema está trabajando en el dominio financiero sobre una ontología poblada con información en español del mercado bursátil de Madrid (IBEX35).

La evaluación del sistema se realiza de manera incremental, evaluando en primera instancia la búsqueda en la ontología para posteriormente evaluar la expansión de la consulta. La finalidad de dividir la evaluación radica en comprobar la influencia que tiene en la expansión de la consulta los posibles fallos producidos por la eficacia de la búsqueda en la ontología.

4.1 Búsqueda en la Ontología

Para la evaluación de la búsqueda en la ontología se ha diseñado una prueba de Cranfield. Para cada una de las consultas planteadas se ha definido un *gold standard* (definido por los expertos que han colaborado en la evaluación) de conceptos que deben recuperarse, por lo que la evaluación de esta búsqueda se hará mediante la precisión y el *recall* o cobertura que ofrece la propia recuperación. En la Figura 3 se muestra la precisión y cobertura de la búsqueda con dos números diferentes de conceptos recuperados (10 y 1) para cada consulta. Al calcular la media se observa que la precisión aumenta (de 36% a 50%) y que la cobertura baja (de 45,5% a 35,4%).

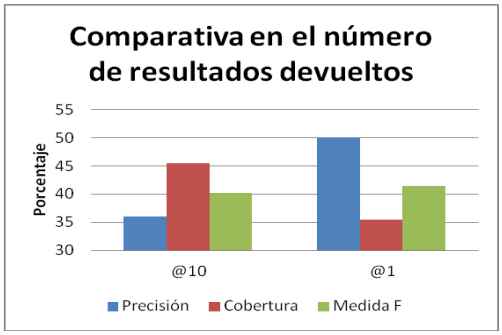

Figura 3: Comparativa de la Recuperación en función del número de Conceptos devueltos.

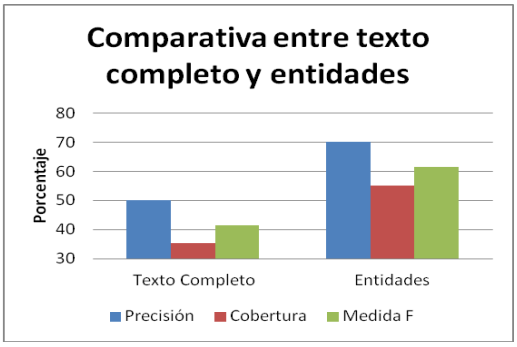

Figura 4: Comparativa numérica de la búsqueda en ontología con texto completo o entidades.

La segunda evaluación que se consideró es la comparación entre la recuperación cuando se lanza el texto completo contra la ontología y cuando se lanzan únicamente las entidades (Figura 4).

Los resultados de precisión y cobertura son muy parecidos en ambos casos, siendo mejores en el caso en el que solo se utilizan las entidades para consultar la búsqueda en la ontología: 70% precisión y 55% cobertura (buscando el texto completo tenemos 50% y 34,4%).

4.2 Expansión de Consulta

Para realizar esta prueba se han utilizado 20 consultas textuales del dominio financiero que se han sido generadas por un licenciado en Economía y en Administración de Empresas. Algunos ejemplos de consulta son: acciones de Telefónica, ¿Cuál ha sido la última cotización de Bankia?, empresas eléctricas que suben, liquidez de Repsol.

Cada una de estas consultas ha sido enriquecida utilizando la búsqueda en ontología, y posteriormente el resultado ha sido evaluado manualmente por dos expertos diferentes, que han valorado la consulta de tres maneras:

- Completamente Expandida (CE): la consulta se ha expandido y toda la información que se ha añadido es relevante para la búsqueda que queremos hacer.
- Parcialmente Expandida (PE): la consulta se ha expandido, pero lo que se ha añadido no es suficiente o tiene una parte correcta y otra no.

- Sin expansión (SE): no se ha expandido o la información que se ha añadido no es relevante para lo que buscamos.

Los expertos han presentado valoraciones dispares a la hora de valorar los resultados, lo que introduce una variable de subjetividad en el proceso. Para calcular el porcentaje de acuerdo que existe entre ellos se ha procedido a calcular el coeficiente Kappa (Cohen, 1960). Este coeficiente se ha calculado en los siete casos diferentes de expansión que se quieren probar (Figura 5).

Según los coeficientes obtenidos, existe una gran disparidad en el acuerdo entre los evaluadores. Las únicas informaciones añadidas a la consulta que son útiles son el nombre de la empresa (k=0,69), el sector en el que opera la empresa (k=0,7) y las entidades presentes en la descripción de la empresa (k=0,75).

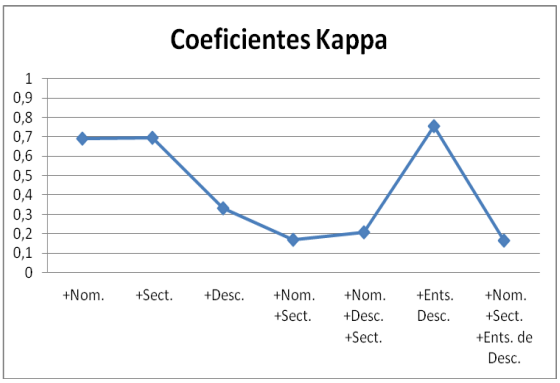

Figura 5: Coeficiente Kappa para evaluar el porcentaje de acuerdo entre los evaluadores.

En la Figura 6 se muestra el porcentaje (media de ambos expertos) de consultas que se han considerado de cada tipo para cada información añadida.

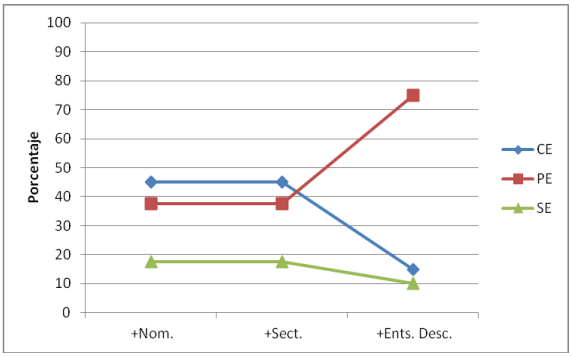

Figura 6: Porcentajes de consultas de cada tipo al realizar las expansiones.

Al añadir el nombre o el sector de la empresa, el porcentaje de consultas que son evaluadas como CE es mucho mayor (45%) que en el caso de añadir las entidades de la descripción (15%). Por el contrario, el porcentaje de consultas consideradas como PE es mucho mayor en el último caso (75%) mientras que en los anteriores era mucho más bajo (37,5%). Estas cifras llevan a que el número de consultas valoradas como SE es menor en el último caso (10% por el 17,5% de los dos primeros casos).

5 *Conclusiones y Líneas Futuras*

Se ha mostrado una prueba de concepto de expansión de consulta usando una ontología pero los resultados no permiten asegurar la utilidad del enfoque, para ello sería necesario escalar el tamaño de la ontología y aplicarlo a un problema de recuperación de información financiera.

La expansión de consulta presenta unos porcentajes bajos en el caso de consultas sin expansión: un 17,5% para +Nom y +Sect. y un 10% en +Ents.Descr.

En los casos de expansión +Nom y +Sect. se añaden pocos términos, mientras que en el caso +Ents.Descr se añaden muchos términos (porque la descripción de una empresa hace referencia a muchas entidades). En ambas situaciones hay ambigüedad, introducida por la recuperación de conceptos que no son relevantes para la consulta, y que en una primera aproximación se ha eliminado limitando los resultados al primero que devuelve el índice de Lucene. Sin embargo, este es un enfoque que habría que mejorar introduciendo algún algoritmo que haga uso de las relaciones semánticas de la ontología para resolver la ambigüedad y añadir a la consulta aquellos términos relacionados entre sí.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el proyecto Trendminer (EU FP7-ICT287863), el proyecto Monnet (EU FP7-ICT 247176) y MA2VICMR (S2009/TIC-1542).

Bibliografía

Bhogal, J., A. Macfarlane y P. Smith. 2007. A review of ontology based query expansion. En *Information Processing and Management*, 43, 4, pp: 866-886.

Carpineto, C. y G. Romano. 2012. A survey of automatic query expansion in information retrieval. *ACM, Comput. Surv. 44, 1, Article 1 (January 2012), 50 pages, DOI = 10.1145/2071389.2071390*

Carstens, C. 2011. Ontology Based Query Expansion-Retrieval Support for the Domain of Educational Research.

Cohen, J. 1960. A coefficient of agreement for nominal scale. En *Educat. Psychol. Measure, 20*, pp: 37-46.

Declerck, T. y D. Gromann. 2012. Towards the Generation of Semantically Enriched Multilingual Components of Ontology Labels. En: *Proceedings of the 3rd International Workshop on the Multilingual Semantic Web (MSW3) volume 936, CEUR Workshop Proceedings, pp: 11-23.*

Dey, L., S. Singh, R. Rai y S. Gupta. 2005. Ontology Aided Query Expansion for Retrieving Relevant Texts. En *Advances in Web Intelligence*, pp: 988-991.

Díaz-Galiano, M. C., M. T. Martín-Valdivia y L. A. Ureña-López. 2009. Query expansion with a medical ontology to improve a multimodal information retrieval system. En *Computers in Biology and Medicine, 39, 4* pp: 396-403.

Dragoni, M., C. C. Pereira y A. G. B. Tettamanzi. 2012. A conceptual representation of documents and queries for information retrieval systems by using light ontologies. En *Expert Systems with Applications 39, 12*, pp: 10376-10388.

Farhoodi, M., M. Mahmoudi, A. M. Zare Bidoki, A. Yari y M. Azadnia. 2009. Query Expansion Using Persian Ontology Derived from Wikipedia. En *World Applied Sciences Journal 7 (4)*, pp: 410-417.

Fu, G., C. B. Jones y A. I. Abdelmoty. 2005. Ontology-based Spatial Query Expansion inInformation Retrieval. En *OTM Conferences 2005*, pp: 1466-1482.

Gulati, P. y A. K. Sharma. 2010. Ontology Driven Query Expansion for Better Image. Retrieval. En *International Journal of Computer Applications 5(10)*, pp: 33–37.

Levenshtein, V.I. Binary codes capable of correcting deletions, insertions, and reversals. 1966. *Doklady Akademii Nauk SSSR*, 163(4):845-848, 1965 (Ruso). Traducción a inglés en Soviet Physics Doklady, 10(8):707-710.

Liu, L. y F. Li. 2011. 3-layer Ontology Based Query Expansion for Searching. En *ISNN 2011, Part III, LNCS 6677,* pp: 621–628.

Mandala, R., T. Tokunaga y H. Tanaka. 2000. Query expansion using heterogeneous thesauri. En *Information Processing and Management 36*, pp: 361-378.

Martínez-Fernández, J. L., A. M. García-Serrano, J. Villena-Román y P. Martínez. 2008. Expanding Queries Through Word Sense Disambiguation, Evaluation of Multilingual and Multi-modal Information Retrieval; LNCS Vol. 4730, pp: 613-616.

Schweighofer, E. y A. Geist. 2007. Legal Query Expansion using Ontologies and Relevance Feedback. En *LOAIT 2007,* pp: 149-160.

Tuominen, J., T. Kauppinen, K. Viljanen, y E. Hyvonen. 2009. Ontology-Based Query Expansion Widget for Information Retrieval. En *Proceedings of the 5th Workshop on Scripting and Development for the Semantic Web (SFSW 2009), 6th European Semantic Web Conference (ESWC 2009).*

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods (American Statistical Association)*: 354–359.

Wollersheim, D. y W. J. Rahayu. 2005. Ontology Based Query Expansion Framework for Use in Medical Information Systems., En *IJWIS 2005, 1, (2)*, pp: 101-115.

Wu, F., G. Wu y X. Fu. 2007. Design and Implementation of Ontology-Based Query Expansion for Information Retrieval. En *IFIP - The International Federation for Information Processing 254*, pp: 293-298.

Aprendizaje Automático en PLN

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 119-126 recibido 18-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Exploring Automatic Feature Selection for Transition-Based Dependency Parsing

Explorando la Selección Automática de Características para Analizadores Basados en Transiciones.

Miguel Ballesteros
Natural Language Processing Group,
Universitat Pompeu Fabra, Spain
miguel.ballesteros@upf.edu

Resumen: En este artículo se investigan técnicas automáticas para encontrar un modelo óptimo de características en el caso de un analizador de dependencias basado en transiciones. Mostramos un estudio comparativo entre algoritmos de búsqueda, sistemas de validación y reglas de decisión demostrando al mismo tiempo que usando nuestros métodos es posible conseguir modelos complejos que proporcionan mejores resultados que los modelos que siguen configuraciones por defecto.
Palabras clave: Análisis de dependencias, MaltOptimizer, MaltParser

Abstract: In this paper we investigate automatic techniques for finding an optimal feature model in the case of transition-based dependency parsing. We show a comparative study making a distinction between search algorithms, validation and decision rules demonstrating at the same time that using our methods it is possible to come up with quite complex feature specifications which are able to provide better results than default feature models.
Keywords: Dependency parsing, MaltOptimizer, MaltParser

1. Introduction

The choice of features to build data-driven NLP applications is something that needs to be done to produce good and competitive results. Besides application parameters, feature selection is the central way of tuning a system and it is not an easy task. It is difficult for researchers without specialized knowledge and it is also complicated for experienced researchers because it normally requires a search in a large space of possible cases. This is time consuming and demands deep knowledge of the task and the parsing algorithms involved.

Automatic feature selection is a process commonly used in machine learning, where the features that perform better are selected automatically for a single task. Since the inclusion of *MaltOptimizer*, it is not a matter of task expertise anymore (Ballesteros and Nivre, 2012), howewer for other tasks and parsing packages it still requires a lot of user action to produce a model capable of providing results that are comparable to the state of the art. We believe that this fact is still an issue in nowadays dependency parsing and Natural Language Processing (Smith, 2011), and this is why we took it as an inspiration.

In this paper we introduce and compare some automatic feature selection techniques, that are based on the ones implemented in MaltOptimizer. These techniques find an optimal feature set for a transition-based dependency parser: MaltParser (Nivre et al., 2007). Since MaltParser is based on support vector machines (henceforth, SVM[1]), there is no way to handle previously the weight of the features, because finding the appropriate weights for different features is exactly what a SVM does.

Therefore, we show firstly how it is possible to produce an optimal feature set for a transition-based dependency parser. Secondly, we compare a couple of algorithms and different criteria when selecting features, and we show how and why we get different results. Finally, we show some conclusions and ideas for further work.

[1] In our experiments we selected LIBLINEAR (Fan et al., 2008), to the detriment of LIBSVM (Chang and Lin, 2001), as training engine due to the inclusion of MaltOptimizer algorithms which are restricted to LIBLINEAR.

ISSN 1135-5948

2. Automatic Feature Selection

Automatic feature selection techniques have become a need in many natural language processing applications, but at this writing there is still not a significant amount of publications in the NLP community facing this problem (Smith, 2011). The objectives of an automatic feature selection technique are the following: (i) avoid overfitting and improve the performance, (ii) produce faster and more effective models and (iii) get more information from the annotated data.

There are two main approaches of finding an optimal feature set, others, as the ones that we show in the present paper, can be derived from these two:

Forward Selection. The process normally starts with zero features, and it adds them one by one, keeping them if they provide improvements. Besides the mixed backward-forward selection of MaltOptimizer (Ballesteros and Nivre, 2012), we can find an example on automatic feature selection carried out in this way for transition-based parsing (Nilsson and Nugues, 2010).

Backward Selection. The processes normally start with a big set of features, which is the case in transition-based parsing (or all the features, if possible) and remove them one by one, at each step removing the one that produces a feature set that performs better in a significant way. In this case, the concept of significant is normally more relaxed because a feature set with less features is less sensitive to overfitting and probably more efficient. As we already mentioned, MaltOptimizer also provides a backward selection of features but it is a merge between backward and forward.

We can find some relevant work that solve and study the problem of feature selection in a similar way as in the present paper in research areas different than NLP. For instance, the work done by Das and Kempe (2011), in which they demonstrated that "greedy algorithms perform well even when the features are highly correlated", which is something that is inherent to transition-based dependency parsing. Similarly, Pahikkala et al. (2010) show how to speed up a forward feature selection by applying a greedy search, which motivated the experiments shown in the present paper.

In the case that occupies our study, which is transition-based dependency parsing, we can have in principle an infinite set of possible features, but it is possible to isolate a rather small pool (to be handled automatically) of potentially useful features for each window.[2] We are normally able to tune the part–of–speech window, morphology, such as gender or number and features based on the partially built dependency tree. We can also be able to provide very useful conjunction features, which means that two features are considered as single feature by the parser. All of this can be done normally within two data structures, the buffer and the stack, but in some cases, depending on the parsing algorithm, we can also have a third (or even fourth) data structure that can be included in the feature specification.

3. MaltParser Feature Language

A transition-based parser uses two data structures (as mentioned above, there could be some auxiliary data structures, but there are at least two), a *buffer* and a *stack*. The buffer provides the words that are going to be used during the parsing process and the stack stores the words that are producing arcs from/to them. MaltParser implements four families of transition-based parsers and all of them use features over the stack and the buffer, which basically means that the parsing algorithm would take into account the annotated info (or partially built trees) of the words that are in the first positions of the stack and the buffer. However it is possible to define features in any position of the data structures, and this is why the search may be very extensive. Figure 1 shows the transition system of one of the parsing algorithm families (Nivre's), and Figure 2 shows how a transition-based parser works for a given sentence following it.

MaltParser uses the CoNLL data format and it is therefore possible to generate features over the columns annotated in the CoNLL files. It is possible to define features over the stack and buffer slots, containing information about part-of-speech (fine-grained: POSTAG, and coarse-grained: CPOSTAG), simple word (FORM), stemmed version of the word (LEMMA), a list of morphosyntactic features (FEAT) and dependency struc-

[2]The concept 'window' refers to the different columns (POSTAG, CPOSTAG, LEMMA, FEATS, DEPREL) that a CoNLL file has. See `http://ilk.uvt.nl/conll/#dataformat` for more information.

$$\textsc{Shift}: \langle \Sigma, i|B, H, D\rangle \Rightarrow \langle \Sigma|i, B, H, D\rangle$$
$$\textsc{Reduce}: \langle \Sigma|i, B, H, D\rangle \Rightarrow \langle \Sigma, B, H, D\rangle$$
$$\textsc{Left-Arc}\ (r): \langle \Sigma|i, j|B, H, D\rangle \Rightarrow \langle \Sigma, j|B, H[i \rightarrow j], D[i \rightarrow \mathrm{r}]\}\rangle \quad \text{if h(i)} \neq 0.$$
$$\textsc{Right-Arc}\ (r): \langle \Sigma|i, j|B, H, D\rangle \Rightarrow \langle \Sigma|i|j, B, H[j \rightarrow i], D[j \rightarrow \mathrm{r}]\}\rangle \quad \text{if h(j)} = 0.$$

Figure 1: Transition System for Nivre's algorithms (Nivre et al., 2007). B refers to the buffer and Σ to the stack. H and D conform the partially built dependency structure referring to heads and dependency labels.

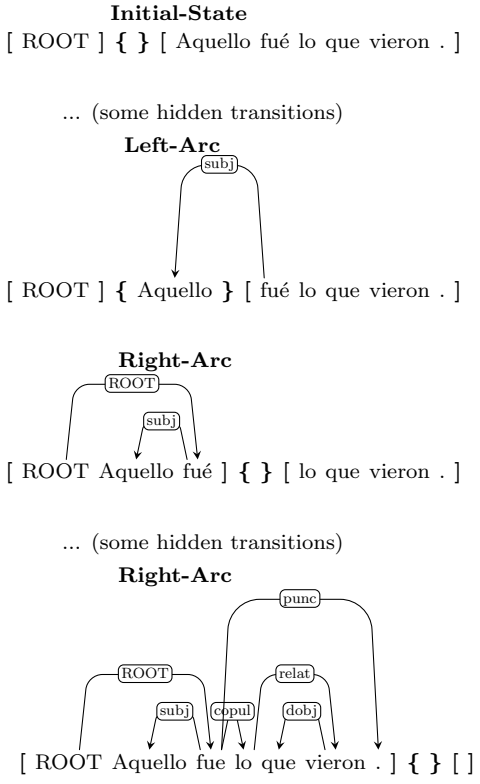

Figure 2: Parsing example for a sentence written in Spanish: *Aquello fue lo que vieron [That is what they saw].* The buffer is the structure on the right, and the stack is on the left.

tures that are being produced in the parsing process (DEPREL).

4. Feature Selection Algorithms

In this Section we describe the two implemented approaches that we are willing to test. Both approaches carry out the steps implemented in MaltOptimizer but they perform the search differently. The algorithms basically provide a backward and forward search of features by performing the following steps: (i) modify POSTAG and FORM features , (ii) modify DEPREL and POSTAG-DEPREL merge features over the partially built dependency structure, (iii) try with CPOSTAG, FEATS and LEMMA features if possible and (iv) add conjunctions of POSTAG and FORM features.

As mentioned by Ballesteros and Nivre (2012) the algorithm steps are not the same for all parsing algorithms; as shown in Section 3, the algorithms make use of different data structures, but the steps and the data structures are more or less equivalent.

As we mentioned in Section 2, these methods start with ***backward*** selection experiments removing features from the default feature model with the intention of testing whether they are useful. After that, they try with ***forward*** selection experiments, by testing features one by one and in combination. In this phase, a threshold of 0.05% LAS[3] (Labeled Attachment Score) is used to determine whether a feature is useful or not.

In this Section we describe the two implemented approaches that follow the steps presented above.

4.1. Relaxed Greedy Algorithm

The Relaxed Greedy approach traverses all the steps presented above adding one feature at a time and keeping the feature set that produces the best outcome. This Relaxed Greedy algorithm tries with all the backward and forward operations for all the steps shown at the beginning of Section 4, and it does not prune the search at all. Therefore, it can be understood as an exhaustive feature search that adds two, three or even more features at a time. We could think that an exhaustive feature search prevents getting stuck

[3] LAS = Percentage of scoring tokens for which the system has predicted the correct labeled attachments.

in local optima, which is something that intuitively could happen to the Greedy algorithm, presented in next subsection.

This algorithm implies running a high number of experiments because it just adds and tries with a big set of experiments, keeping the best feature model after each attempt. We could therefore expect that this algorithm overfits the performance in some cases providing a model with lower training error but higher test error.

Summing up, we have two different hypotheses for this algorithm: (1) it would not get stuck in local optima, (2) it could overfit the performance.

4.2. Greedy Algorithm

The Greedy algorithm is the one implemented and included in the MaltOptimizer distribution (Ballesteros and Nivre, 2012), it minimizes the number of experiments according to linguistic expert knowledge and experience (Nivre and Hall, 2010). It also follows the steps shown at the beginning of Section 4. However, in spite of trying with all the big set of possible features for each step as it is done in the Relaxed Greedy algorithm, it does the following:

1. It prunes the search when a backward feature selection of features provides improvements for a specific window, because it does not try with any forward selection experiments.
2. It prunes the search when a forward selection of features is not successful for a specific window, because it does not try with more forward selection experiments.

Therefore, it drastically reduces the number of iterations for backward and forward operations comparing with Relaxed Greedy.

For this algorithm we also have two different hypotheses: (1) it could intuitively get stuck in a local optima because it reduces the number of experiments and it could prune the search very early expecting that the search may not produce good results and (2) we could also expect that it underfits the performance due to (1). However, it is also worth remarking that the steps of this algorithm were developed with deep proven experience.

In the following Section we show an in-depth comparison between the Greedy and the Relaxed Greedy algorithms taking some experimental results into account, we therefore show which algorithm is the most accurate in order to get an optimal feature set.

5. Experiments and Results

As we mentioned in Sections 4.1 and 4.2, we expect that the Greedy and the Relaxed Greedy algorithms could underfit and overfit the performance respectively, we took these two facts as hypotheses. Therefore, we try to extract conclusions by running and comparing the outcomes of both algorithms. We train models using data sets for 6 different languages that were included in the CoNLL-X Shared Task (Buchholz and Marsi, 2006) (Arabic, Dutch, Slovene, Spanish, Swedish and Turkish), and we also test the models produced over the separate test-sets not used during the optimization.

With the intention of having a starting point of comparison we run the first phases of MaltOptimizer in order to set some parameters (phase 1) and select the parsing algorithm (phase 2) that performs the best over each data set.

In the rest of this section we firstly show the results of each of the algorithms implemented during the training phase in which we get the optimal feature set, afterwards, we show the test results when we test the outcome model with an unseen test set not used during the optimization (Section 5.1). Finally, and considering the results of the first two experiments, we perform a 5-fold cross validation strategy to demonstrate its usefulness (Section 5.2). We also show three different kind of sentence selection strategies for the folds.

It is worth mentioning that we always compare our results with the results given by default feature models to ensure the usefulness of our methods.

5.1. Results and Comparisons between Greedy and Relaxed Greedy

Table 1 shows the results of the Greedy algorithm and the Relaxed Greedy algorithm for a selection of languages. Note that these results are obtained using 80% of the training set for training and 20% as development test set, which were obtained using the entire training set and a separate held-out test set for evaluation. Therefore, these are the

results obtained during the optimization process.

Language	DefaultFM	Greedy	Relaxed Greedy
Arabic	63.84	65.56 (+1.72)	**66.00 (+2.16)**
Dutch	78.02	**82.63 (+4.61)**	82.49 (+4.47)
Slovene	68.40	71.71 (+3.31)	**72.43 (+4.03)**
Spanish	76.64	79.38 (+2.74)	**79.62 (+2.98)**
Swedish	83.50	84.09 (+0.59)	**84.20 (+0.70)**
Turkish	58.29	66.92 (+8.63)	**67.19 (+8.90)**

Table 1: Labeled attachment score with comparison to default feature model (MaltParser in its default settings without feature selection) and the greedy approach during the optimization process.

We can observe how Relaxed Greedy seems to beat the results of Greedy, with the exception of Dutch. Nevertheless, the differences are not very remarkable. Relaxed Greedy always carries out more than 100 different experiments, and Greedy between 40 and 50, depending on the pruning decisions and results.

This fact means, that the decisions taken during the development of the Greedy algorithm seem to be the correct ones. This fact is also evidenced in the Figure 5.1 in which we show the results of the Greedy and the Relaxed Greedy algorithms for the Slovene example.[4] We can see how the Greedy algorithm achieves an optimal accuracy faster than Relaxed Greedy, but in some cases it seems that it gets stuck (in optimization time) in local optima because the Relaxed Greedy approach beats these results finding eventually a more accurate feature configuration. And finally, it is also interesting to remark that the Greedy algorithm rarely produce results that are worse than the baseline or default feature model in none of its steps, however Relaxed Greedy does.

In order to find out the second hypothesis, whether the algorithms overfit or underfit the performance, and also whether our methods are really useful or not, we tested the obtained feature model with the real testing data set used in the CoNLL-X Shared Task, the Table 2 shows the results obtained.

In this case, most of the differences are indeed statistically significant comparing with the default models results.[5] According to McNemar's test we got significant improvements for Dutch, Slovene, Spanish and Turkish while the ones obtained for Arabic and Swedish are not better enough. Moreover, for the languages in which we have statistically significant improvements, these ones are for $p<0.01$ and for $p<0.05$ and the Z value varies from 2.687 in the case of Spanish to 12.452 in the case of Turkish. Taking into account the size of the testing data sets these results are quite remarkable.

Language	DefaultFM	Greedy	Relaxed Greedy
Arabic	64.93	**66.01**	65.71
Dutch	72.63	**77.23**	76.89
Slovene	69.66	**73.68**	73.26
Spanish	78.68	**80.00**	79.84
Swedish	83.50	83.81	**83.85**
Turkish	56.32	64.11	**64.31**

Table 2: Labeled attachment score with comparison to default feature models and the greedy approach for a selection of languages using the optimized models.

Comparing the Greedy algorithm and the Relaxed Greedy algorithm we can conclude that the Greedy one (which is much faster[6]) is more capable of providing a competitive feature model for the real case (in which the user would need to parse sentences that are not included neither in the test set nor in the training set) because the Relaxed Greedy models seem to be overfitted to the test set in most of the cases. Running the McNemar's test most of the differences are not statistically significant neither for $p<0.01$ nor for $p<0.05$, but for the Slovene treebank there is a statistically significant difference for $p<0.05$ with a Z value of 2.171, taking into account that the test sets are small - 5000 tokens- this is an interesting result. In summary, the outcomes given over most of languages nevertheless strongly suggests that the simple Greedy algorithm is more accurate and it does not underfit the performance.

These results led us to think that we should consider more conservative criteria for accepting improvements during feature selection. Therefore, in the following section we show a more informative approach, a K-Fold cross validation experiment for the Greedy al-

[4] In the Figure we simulate that the Greedy algorithm is waiting for the Relaxed Greedy before taking another step.

[5] We run the statistically significant tests by using MaltEval (Nilsson and Nivre, 2008)

[6] Every step of the algorithm requires to train a parsing model and test with the separate held-out test set, and depending on the size of the training treebank it could take a while

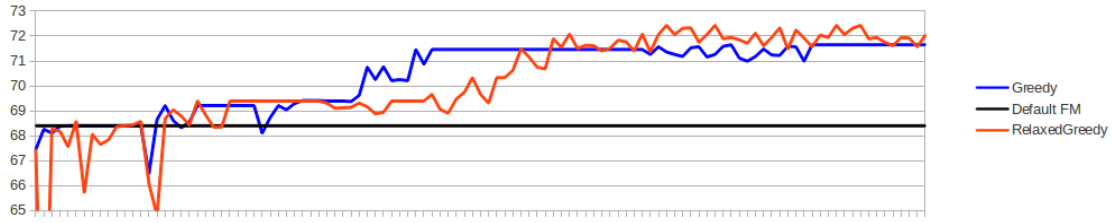

Figure 3: Results obtained by Greedy and Relaxed Greedy in every step of the algorithms for the Slovene treebank. The X axis shows the number of steps in the optimization process and the Y axis shows the performance achieved.

gorithm, because it is the one that provides better results in the present experiment and it is the only one that provides a statistically significant difference compared to Relaxed Greedy.

5.2. 5-Fold Cross Experiment

We decided to carry out a 5-fold cross validation experiment to be included in the validation step of the Greedy algorithm due to the results obtained in the real case with Greedy and Relaxed Greedy and taking into account that one of the best ways of estimating the generalization performance of a model trained on a subset of features is to use cross-validation, as shown in (John, Kohavi, and Pfleger, 1994).

We divided the corpus in 5 folds in order to have similar number of sentences in the folds as we had in the previous experiments, when we divided the corpus in 80% for training and 20% for testing.

It is well known that there are various ways of extracting the folds from the training corpora. For the present experiment and in order to get a more complex and interesting comparison we try two different approaches: (i) Extracting the sentences in an iterative way, by doing a simple split, firstly the sentences for fold 1, then sentences for fold 2 and so on and (ii) a pseudo randomize selection of sentences which provides more heterogeneous folds. We could come up with the following hypotheses: we could expect that the simple split selection of sentences will underfit the performance and we could also expect that the pseudo randomize selection will provide better results.

We also decided to implement three different criteria in order to decide whether a feature set is worth to be included in the final feature model: (i) considering that the average LAS over all folds must beat the result of the best feature model so far, (ii) considering that the majority of folds (in this case 3 of 5) must beat the result of the best feature model so far, and (iii) considering that all the folds must beat the result of the best feature model so far. Therefore, we could come up with the following hypotheses regardless or whether we use the simple split selection of sentences or the pseudo-randomize selection of sentences:

1. We could expect that (i) and (ii) will provide similar results, and it seems that both of them will neither underfit nor overfit the performance.
2. We could also expect that (iii) is going to underfit the performance in most of the cases.

In the following Subsections we show a set of experiments in which we discuss whether our hypotheses are corroborated or falsified.

5.2.1. Simple Split Selection of Sentences

The simple split selection of sentences only provides improvements for Slovene and Turkish for the *average* and the *majority* criteria, producing 70.24 LAS in the case of Slovene and 66.00 LAS in the case of Turkish. It seems that this selection of sentences is not very representative of the data set and this fact misleads the results when considering 5-fold cross validation.

The *average* and the *majority* criteria even come up with the same feature set and in the real case (training with the whole training set and testing with the test set) they got 73.52 LAS in the case of Slovene, and 64.45 LAS in the case of Turkish. These results compared with the ones that we got applying the simple Greedy step wise approach

are better for Turkish (+0.3) and worse for Slovene (-0.2). These differences are not statistically significant according to McNemar's test, neither for $p<0.01$ nor for $p<0.05$.

5.2.2. Pseudo Randomize Selection of Sentences

We believe that the pseudo randomize selection of sentences is more representative of the real case. Our methods provide the results of Table 3, which also shows the results of the Greedy algorithm without making use of the 5-fold cross validation. Moreover, Figure 5.2.2 shows the results of the 5 folds, with *average* criterion, pseudo randomize selection of sentences and the Slovene corpus, we can see how all the folds produce high results if we compare with the simple split selection.

Language	DefaultFM	Greedy	Average	Majority	All
Arabic	63.84	65.56	66.44	**66.62**	65.33
Dutch	78.02	**82.63**	82.32	82.29	81.42
Slovene	68.40	71.71	**72.00**	**72.00**	69.46
Spanish	76.64	**79.38**	79.29	79.29	76.64
Swedish	83.50	**84.09**	83.50	83.50	83.50
Turkish	58.29	66.92	67.11	67.01	**67.37**

Table 3: Labeled attachment score with comparison to default feature model and the greedy approach for a selection of languages from the CoNLL-X shared task (Buchholz and Marsi, 2006), reporting the results of the 5-fold cross validation.

As we can see the results of the 5 fold cross validation strategy are more informative, intuitively, we can rely more in the feature models obtained during the process because they have been tested over 5 different folds and represent the real case in a more accurate way. In order to demonstrate this fact, we set up Table 4 which shows the results of the obtained feature model when we test them with the test set of the CoNLL-X Shared Task.

As observed in Table 4, the 5-fold cross validation produces higher results for Arabic, Spanish and Turkish, while the simple Greedy algorithm produces better results in the other 3. The *All* criterion seems to be very restrictive because it leads to underfitting, however, *average* and *majority* produce similar and robust results.

Nevertheless, the differences for Slovene are statistically significant according to McNemar's test in favor for the Greedy algorithm for $p<0.01$ and for $p<0.05$. But, the differences for the Turkish algorithm are statistically significant in favor of the 5-fold cross experiment (for the three cases) running McNemar's test only for $p<0.05$ and a Z value of 2.296. The rest of the differences are not significant.

Language	DefaultFM	Greedy	Average	Majority	All
Arabic	64.93	66.01	66.21	**66.27**	65.61
Dutch	72.63	**77.23**	76.97	76.39	75.73
Slovene	69.66	**73.68**	73.32	73.32	71.64
Spanish	78.68	80.00	**80.46**	**80.46**	78.68
Swedish	83.50	**83.81**	83.59	83.59	83.59
Turkish	56.32	64.11	64.85	**65.01**	64.99

Table 4: Labeled attachment score with comparison to default feature model and the greedy approach for a selection of languages from the CoNLL-X shared task (Buchholz and Marsi, 2006), reporting the results of the 5-fold cross validation, making use of the training and test set of the CoNLL-X Shared Task

We can conclude that the Greedy algorithm by itself can provide results as good as an approach that follows more informative criteria, in this case, K-Fold cross validation. Nonetheless, it seems worth to carry out both experiments because in some cases we can find statistically significant improvements when we check over 5 different divisions of the corpus (or folds) and vice versa.

It is also worth noting that comparing the results of the simple split selection of sentences (shown in Section 5.2.1) for Slovene and Turkish (which are the ones that provide improvements) with the corresponding outputs produced by the pseudo randomize selection of sentences by running McNemmar's test. We get a statistically significant different in favor of the pseudo randomize selection for $p<0.05$. Therefore, we can also conclude that the results produced by the pseudo randomize selection are not overfitted and the ones produced by the simple split selection of sentences are underfitted by a misleading selection of sentences.

6. Conclusions

We have demonstrated that different criteria for automatic feature selection in the case of transition based dependency parsing can be accomplished successfully and produce variant and strong results in the final performance. Moreover, both search algorithms presented produce consistent improvements over the default settings using different validation procedures. According to our results

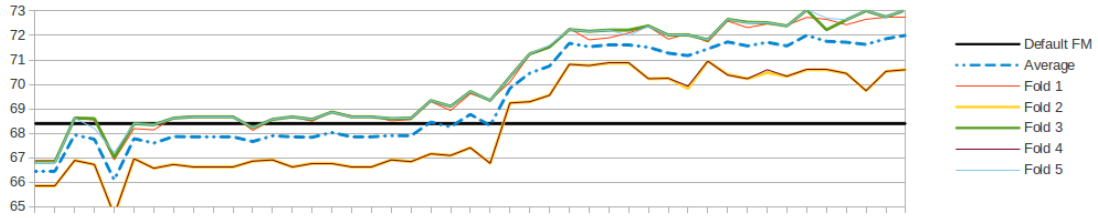

Figure 4: Results obtained by the 5 fold cross experiments with pseudo randomize selection of sentences in every step of the algorithm for the Slovene treebank.

and taking into account that all of our results are inherently based on the same Greedy algorithm, we believe that it is better to follow proven experience and linguistic expertise in this kind of experiments.

It is worth mentioning that we tried to alter the order between the different steps shown in Section 4, but we did not get any improvement nor any significant differences between the different feature sets and the different algorithms. A specific order for a treebank was useful and better, but it was not the same for a different treebank. Therefore, we plan to carry out an in-depth comparison following different experiment orders, not simply altering the order between the steps but making a cross experimental comparison.

Acknowledgments

Thanks to Joakim Nivre, who guided me in the development of MaltOptimizer and the algorithms that are explained in the present paper.

References

Ballesteros, Miguel and Joakim Nivre. 2012. MaltOptimizer: A System for MaltParser Optimization. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*.

Buchholz, Sabine and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. pages 149–164.

Chang, Chih-Chung and Chih-Jen Lin, 2001. *LIBSVM: A Library for Support Vector Machines*. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Das, Abhimanyu and David Kempe. 2011. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 1057–1064, New York, NY, USA, June. ACM.

Fan, R.-E., K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

John, George H., Ron Kohavi, and Karl Pfleger. 1994. Irrelevant Features and the Subset Selection Problem. In *International Conference on Machine Learning*, pages 121–129.

Nilsson, Jens and Joakim Nivre. 2008. Malteval: an evaluation and visualization tool for dependency parsing. In *(LREC'08)*, Marrakech, Morocco, may.

Nilsson, Peter and Pierre Nugues. 2010. Automatic discovery of feature sets for dependency parsing. In *COLING*, pages 824–832.

Nivre, Joakim and Johan Hall. 2010. A quick guide to maltparser optimization. Technical report.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryiğit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13:95–135.

Pahikkala, Tapio, Antti Airola, and Tapio Salakoski. 2010. Speeding up greedy forward selection for regularized least-squares. In *ICMLA*, pages 325–330.

Smith, Noah A. 2011. *Linguistic Structure Prediction*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool.

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 127-134 recibido 02-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Una aproximación supervisada para la minería de opiniones sobre tuits en español en base a conocimiento lingüístico

A supervised approach to opinion mining on Spanish tweets based on linguistic knowledge

David Vilares, Miguel A. Alonso y **Carlos Gómez-Rodríguez**
Departamento de Computación, Universidade da Coruña
Campus de Elviña, 15011 A Coruña
{david.vilares, miguel.alonso, carlos.gomez}@udc.es

Resumen: En este artículo se describe un sistema para la clasificación de la polaridad de tuits escritos en español. Se adopta una aproximación híbrida, que combina conocimiento lingüístico obtenido mediante PLN con técnicas de aprendizaje automático. Como paso previo, se realiza una primera etapa de preprocesado para tratar ciertas características del uso del lenguaje en Twitter. A continuación se extrae información morfológica, sintáctica y semántica, para utilizarla posteriormente como entrada a un clasificador supervisado. La evaluación de la propuesta se lleva a cabo sobre el corpus TASS 2012, anotado para realizar tareas de clasificación con cuatro y seis categorías. Los resultados experimentales muestran un buen rendimiento para ambos casos, lo que valida la utilidad práctica de la propuesta.
Palabras clave: Análisis del sentimiento, Minería de opiniones, Análisis sintáctico de dependencias, Aprendizaje automático, Twitter

Abstract: This article describes a system that classifies the polarity of Spanish tweets. We adopt a hybrid approach, which combines linguistic knowledge acquired by means of NLP with machine learning techniques. We carry out a preprocessing of the tweets as an initial step to address some characteristics of the language used in Twitter. Then, we apply part-of-speech tagging, dependency parsing and extraction of semantic knowledge, and we employ all that information as features for a supervised classifier. We have evaluated our proposal with the TASS 2012 corpus, which is annotated to undertake classification tasks with four and six categories. Experimental results are good in both cases and confirm the practical utility of the approach.
Keywords: Sentiment Analysis, Opinion Mining, Dependency Parsing, Machine Learning, Twitter

1. Introducción

Con la explosión de la Web 2.0 muchos usuarios emplean los medios sociales para compartir sus opiniones y experiencias acerca de productos, servicios o personas. Esta situación ha despertado un gran interés a nivel empresarial, ya que se ve en estos recursos un mecanismo para conocer, de manera eficaz y global, el punto de vista de consumidores sobre una gran variedad de temas. Sin embargo, el análisis manual de este tipo de medios no es una solución viable, dado el flujo ingente de opiniones que en ellos se expresan. A este respecto, la minería de opiniones (MO), conocida también como análisis del sentimiento; es una reciente área de investigación centrada en tareas como determinar automáticamente si en un texto se opina o no, o si la *polaridad* o *sentimiento* que se expresa en él es positiva, negativa o mixta. También es útil de cara a la extracción automática de características, lo que permite conocer la percepción que se tiene sobre aspectos concretos de un tema (p. ej. *"La película Y tiene un gran final"*).

En este contexto, una de los medios sociales más populares es Twitter. En esta red de microblogging, los usuarios expresan sus opiniones en mensajes de hasta 140 caracteres, especialmente sobre temas de actualidad, lo que constituye una importante fuente de información desde el punto de vista de la inteligencia de negocio.

En este artículo presentamos una aproximación que combina conocimiento morfológico, sintáctico y semántico con técnicas de

ISSN 1135-5948

aprendizaje automático para tratar de clasificar la opinión de tuits escritos en español. Para evaluar nuestra propuesta se ha empleado el corpus TASS 2012, donde se distinguen hasta seis categorías distintas.

El resto del artículo se organiza como sigue. En la sección 2 se revisa brevemente el estado del arte en lo referido a la clasificación de la polaridad, centrándonos en estudios relacionados con Twitter. En la sección 3 detallamos nuestra aproximación. En la sección 4 se ilustran los resultados experimentales. Por último, en la sección 5 se presentan las conclusiones y las líneas de trabajo futuras.

2. Estado del arte

Desde que la MO ha sido asumida como un nuevo reto por parte del área de PLN, una de las tareas en las que más esfuerzos se han depositado es la clasificación de la polaridad. Dos han sido los principales enfoques propuestos para resolver este problema: el semántico y el supervisado. El primero (Turney, 2002), se apoya en diccionarios semánticos, conocidos también como lexicones de opiniones, donde a cada palabra que denota opinión se le asigna una orientación semántica (OS). El segundo enfoque (Pang, Lee, y Vaithyanathan, 2002), propone una solución basada en aprendizaje automático (AA), asumiendo la tarea como un problema genérico de clasificación.

La MO se ha centrado tradicionalmente en el análisis de textos extensos. Sin embargo, en los últimos años se ha prestado especial atención a la detección del sentimiento en microtextos, sobre todo en lengua inglesa, dado el éxito de las redes de microblogging. Por ejemplo, Thelwall et al. (2010) proponen el algoritmo SentiStrength para el análisis de textos cortos en la red social MySpace. En (Thelwall, Buckley, y Paltoglou, 2011) se estudia cómo los eventos sociales influyen en el número de opiniones expresadas en Twitter. En (Thelwall, Buckley, y Paltoglou, 2012) se extiende y adapta SentiStrength a otros medios de comunicación web de mensajes cortos. Por otro lado, Bakliwal et al. (2012) presentan un método de puntuación del sentimiento no supervisado y comparan su propuesta con una aproximación supervisada, alcanzando una precisión similar. Zhang et al. (2011) proponen un método supervisado para el análisis de tuits que se entrena a partir de los datos proporcionados por un analizador semántico base no supervisado.

Un problema del análisis del sentimiento sobre textos cortos es el coste que presenta la creación de corpora y de recursos específicos. Para el castellano, destaca el corpus TASS 2012, presentado en el Taller de Análisis del Sentimiento en la SEPLN (Villena-Román et al., 2013), y que utilizamos en este artículo. Se trata de una colección de tuits en español escritos por diversas personalidades públicas. Está constituido por un conjunto de entrenamiento y uno de test que contienen 7,219 y 60,798 tuits respectivamente. Cada uno de ellos está anotado con una de la siguientes categorías: *muy positivo* (P+), *positivo* (P), *neutro/mixto* (NEU), *negativo* (N), *muy negativo* (N+) o *sin opinión* (NONE). En relación con la detección del sentimiento, el taller propuso dos actividades: una de clasificación en seis categorías (con las seis clases mencionadas) y otra de clasificación en cuatro (las clases P+ y N+ se incluyen en las clases P y N respectivamente). La etiquetación del conjunto de test se caracteriza por haber sido obtenida mediante un *pooling* de los resultados de los sistemas que participaron en el taller; seguida de una revisión manual para los casos conflictivos.

Existen diversas propuestas que han evaluado este corpus, la mayoría de ellas enmarcadas dentro de un enfoque basado en AA. Por ejemplo, Saralegi Urizar y San Vicente Roncal (2012) plantean una solución supervisada que emplea conocimiento lingüístico para obtener los atributos de entrada al clasificador. Realizan tareas de lematización y etiquetación, y consideran aspectos relevantes en un entorno de MO, como son los emoticonos o la negación. Batista y Ribeiro (2013) plantean una aproximación supervisada. Sin embargo, en lugar de entrenar un clasificador que distinga n polaridades, construyen *n-1* clasificadores binarios. Finalmente combinan los resultados de los modelos entrenados para eliminar ambigüedades en la clasificación de ciertos tuits. Trilla y Alías (2012) proponen un esquema de clasificación de texto basado en el Multinomial Naive Bayes. Martínez Cámara et al. (2012) también plantean una solución de AA supervisado, en concreto aplicando Support Vector Machines (SVM). Como entrada al clasificador usan una bolsa de unigramas que representa el mensaje. Además incluyen otras características como emotico-

nos, o el número de palabras positivas y negativas presentes en un tuit. En la misma líneas, Fernández Anta et al. (2012) comparan el rendimiento de varios clasificadores supervisados sobre el corpus TASS 2012. Por otra parte, Martín-Wanton y Carrillo de Albornoz (2012) proponen un método basado diccionarios afectivos y WordNet, mientras que Castellano González, Cigarrán Recuero, y García Serrano (2012) abordan el problema del análisis del sentimiento desde una perspectiva de recuperación de información. Por último, Moreno-Ortiz y Pérez-Hernández (2012) emplean Sentitext, un sistema léxico no supervisado para el análisis del sentimiento de textos en castellano, para clasificar la polaridad.

3. Sistema híbrido de clasificación de polaridades

En este artículo se propone una aproximación que combina conocimiento lingüístico con técnicas supervisadas para tratar de resolver los problemas que ambos enfoques presentan separadamente, y que se ven acentuados en un medio como Twitter. Las aproximaciones semánticas se caracterizan por emplear un diccionario de OS genérico. Este enfoque ha demostrado ser útil en distintos ámbitos, pero su rendimiento en Twitter disminuye ya que en este medio existe una elevada frecuencia de abreviaturas, emoticonos o expresiones que denotan opinión, pero cuyas OS no se encuentran en un lexicón genérico, lo que en términos de rendimiento se traduce en un bajo *recall* (Zhang et al., 2011). Respecto a las aproximaciones supervisadas, el principal problema reside en su dependencia del dominio y en el coste de crear conjuntos de entrenamiento. Estos métodos representan el texto como una bolsa de palabras, aprendiendo satisfactoriamente la percepción de un término para un ámbito en concreto. Sin embargo, su rendimiento cae drásticamente al clasificar textos de un campo distinto (Taboada et al., 2011).

Nuestra propuesta parte de un sistema base, para a continuación determinar cómo la información morfosintáctica sirve de ayuda a un clasificador supervisado. Además, se propone un método automático para maximizar el rendimiento en un dominio en concreto, atendiendo a criterios semánticos. Como clasificador, se ha optado por una SMO, una implementación de SVM presentada en (Platt, 1999), y que se incorpora en el software de algoritmos de aprendizaje automático WEKA (Hall et al., 2009). La elección de este clasificador se debe al buen rendimiento que se obtuvo con él sobre el corpus de entrenamiento y a los buenos resultados que otros estudios han conseguido sobre el mismo corpus (Saralegi Urizar y San Vicente Roncal, 2012).

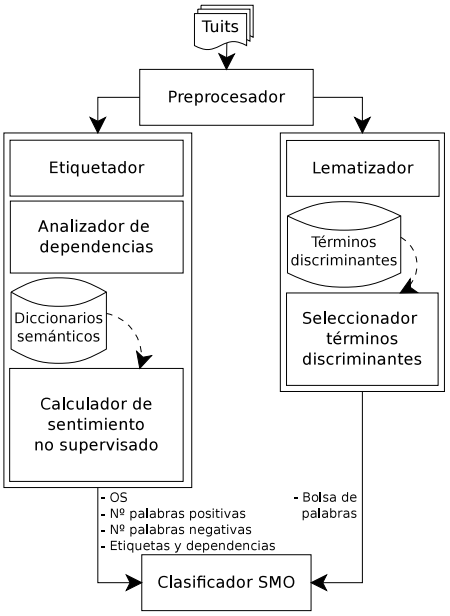

Figura 1: Arquitectura general del sistema

La figura 1 describe la arquitectura general del sistema, cuyo funcionamiento pasamos a describir en los siguientes subapartados.

3.1. Preprocesado

Como paso previo, todos los tuits fueron sometidos al siguiente preprocesado *ad-hoc*, para tratar el uso particular que se hace del lenguaje en Twitter:

- *Tratamiento de emoticonos*: Existe una gran variedad de símbolos que se emplean para reflejar un estado de ánimo. Para identificarlos se utilizó la colección recogida en (Agarwal et al., 2011). Cada emoticono se sustituye por una de estas cinco etiquetas: muy positivo (EMP), positivo (EP), neutro (ENEU), negativo (EN) y muy negativo (EMP).
- *Normalización de URL's*: Las direcciones web presentes en un tuit son sustituidas por la cadena *"URL"*.
- *Corrección de abreviaturas más frecuentes*: Se sustituyen algunos de los vocablos no gramaticales más habituales (p.

ej. *"q"*, *"xq"*,...) por su forma reconocida.

- *Normalización de risas*: Las expresiones típicas que permiten reflejar este fenómeno vía escrita (p. ej. *"jajjja"*, *"JJJEJEJE"*,...), son normalizadas como *jxjx* donde $x \in \{a, e, i, o, u\}$.

- *Tratamiento de elementos específicos de Twitter ("@"y "#")*: Las menciones al usuario se mantienen eliminando la *"@"* y capitalizando la primera letra (p. ej. *"@usuario"* pasa a ser *"Usuario"*).[1] Respecto a los hashtags (p. ej. *"#sepln"*), si aparece al principio o al final del tuit se elimina el mismo. En caso contrario se suprime solamente la *"#"* (p. ej. *"#sepln"* pasa a ser *"sepln"*).

3.2. Propuesta base

Como sistema base se utilizó una aproximación semántica presentada en (Vilares, Alonso, y Gómez-Rodríguez, 2013). Se realizan tareas de segmentación, tokenización y etiquetación morfológica para luego obtener el *árbol de dependencias* de cada oración mediante algoritmos de análisis sintáctico de dependencias, utilizando para ello MaltParser (Nivre et al., 2007). Este tipo de análisis establece vínculos padre/dependiente entre pares de palabras. A cada uno de esos vínculos se les denomina *dependencia* y son anotados con la función sintáctica que relaciona los dos términos. En la figura 2 se ilustra un ejemplo de este tipo de análisis.

El árbol de dependencias se emplea entonces para realizar el análisis del sentimiento sobre los tuits, tratando sintácticamente tres de las construcciones más significativas en el ámbito de la MO: la intensificación, las oraciones subordinadas adversativas y la negación. Para determinar la OS de las palabras que denotan opinión, se utiliza el SO-DictionariesV1.11Spa, una colección de diccionarios semánticos genéricos presentados en (Brooke, Tofiloski, y Taboada, 2009). Como resultado del análisis semántico se obtiene la OS global del texto. Este valor, junto con el número total de palabras positivas y negativas presentes en un tuit, se emplearán como atributos para entrenar un clasificador.

[1]La letra mayúscula se fuerza para que se trate como un nombre propio en la etapa de análisis morfológico.

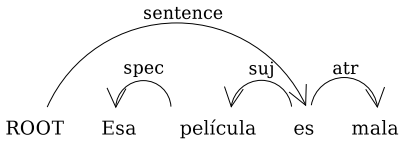

Figura 2: Ejemplo de un árbol de dependencias.

3.3. Información morfosintáctica (IMS)

La utilización de etiquetas morfológicas como un elemento que por sí mismo sirva de ayuda en la clasificación de la polaridad de un texto es una cuestión que ha sido discutida en varias ocasiones. Pak y Paroubek (2010) observaron que la distribución de etiquetas en textos de microblogging es distinta según se trate de un mensaje positivo, negativo o informativo. En la misma línea, Spencer y Uchyigit (2012) y Saralegi Urizar y San Vicente Roncal (2012) destacan una mayor presencia de ciertas categorías gramaticales en textos subjetivos respecto a los objetivos, como es el caso de los adjetivos o las interjecciones. En el corpus TASS 2012 también se observó que algunas etiquetas son más habituales dependiendo de la categoría del tuit. En la tabla 1 mostramos la frecuencia de aparición para algunas de ellas. Para tratar este fenómeno, el número total de apariciones de cada etiqueta en un tuit se incorporará como atributo de entrada a nuestro clasificador.

Cat.	a	n	v	i	f
P+	0,060	0,256	0,111	0,004	0,215
P	0,056	0,266	0,119	0,002	0,198
NEU	0,057	0,254	0,133	0,001	0,163
N	0,050	0,263	0,132	0,001	0,161
N+	0,060	0,266	0,118	0,001	0,154
NONE	0,048	0,299	0,090	0,001	0,220

Tabla 1: Frecuencia de etiquetas en el conjunto de entrenamiento: adjetivos (a), nombres (n), verbos (v), interjecciones (i) y signos de puntuación (f)

Del mismo modo, creemos que ciertos tipos de funciones sintácticas pueden ser más habituales en un tuit en función de su polaridad. En la tabla 2 se muestra la frecuencia de algunas dependencias en los tuits del corpus TASS 2012. Destaca el empleo de dependencias como la de complemento agente (cag), que sugiere que, en las opiniones negativas, la forma pasiva es más habitual. Siguiendo

la misma estrategia que para el caso de las etiquetas, el número de apariciones de cada tipo de dependencia en un tuit se incluirá como atributos de entrada al clasificador.

Cat.	ci	atr	cc	cag
P+	0,008	0,105	0,042	0,004
P	0,010	0,010	0,051	0,000
NEU	0,010	0,141	0,053	0,001
N	0,009	0,000	0,055	0,150
N+	0,007	0,000	0,049	0,145
NONE	0,179	0,008	0,003	0,001

Tabla 2: Frecuencia de etiquetas sintácticas en el conjunto de entrenamiento: complemento indirecto (*ci*), atributo (*atr*), complemento circunstancial (*cc*) y complemento agente (*cag*)

3.4. Adaptación al dominio (AD)

La utilización de conocimiento semántico genérico permite obtener un buen rendimiento base. Sin embargo, estos lexicones presentan varios problemas. Uno de los inconvenientes es su baja cobertura, limitada a los términos incluidos en ellos. Otro es el referido a las OS asignadas a cada término. Éstas, al adaptarse a un contexto general, pueden ser incorrectas para ciertos dominios y contextos, lo que repercute negativamente en el rendimiento. Por ejemplo, la palabra *"asesino"* se percibe intuitivamente como negativa, pero dentro de un dominio de películas, probablemente ese término sirva para describir la temática de la misma, pero no su calidad.

También deben considerarse las peculiaridades del medio en el que se se opina. Los mensajes publicados en Twitter incluyen, con frecuencia, elementos que expresan subjetividad u opinión, pero cuya OS no se refleja en un lexicón de opiniones. Un ejemplo es el uso de emoticonos (*":)"*, *":("*,...) o de interjecciones (*"jaja"*, *"jeje"*,...). Del mismo modo, la utilización de elementos como la etiqueta *Follow Friday* (*"FF"*) o la difusión del mensaje de otro usuario (*"RT"*), llevan asociada con frecuencia una carga subjetiva.

Nuestra solución consiste en desarrollar un mecanismo que permita enriquecer y adaptar automáticamente el conocimiento semántico a las características del dominio. Para ello, fueron dos las tareas que se abordaron: *selección de los términos más discriminantes* y *modificación de los diccionarios genéricos*.

3.4.1. Selección de los términos más discriminantes

El objetivo es crear una bolsa de palabras formada por términos que permitan distinguir bien entre las distintas categorías; de forma que cada uno de esos vocablos se incorporen como entradas del clasificador. Sobre el esquema de pesado utilizado para estos atributos, se ha optado por una ocurrencia binaria, dado que es poco habitual que una misma palabra se repita dentro de un tuit.

Para conocer los elementos que pueden constituir dicha bolsa de palabras, se ha utilizado el soporte de selección de atributos de WEKA y el conjunto de entrenamiento del corpus TASS 2012. Como paso previo se preprocesaron y lematizaron los tuits. Los términos resultantes se han clasificado ordenadamente en función de la ganancia de información que proporcionan respecto a la clase. Para que dicha selección fuese más robusta, se llevó a cabo una validación cruzada de 10 iteraciones. La lista de vocablos resultantes, para las actividades de clasificación en cuatro y seis clases, está constituida por más de 14.000 elementos. Sin embargo, se ha comprobado empíricamente que las mejores configuraciones, en términos de rendimiento, se obtienen con un número reducido de palabras (entre 100 y 400). En la tabla 3 se muestran algunos elementos característicos de esta red de microblogging y la posición que ocupan en la clasificación.

Término	Posición (4 clases)	Posición (6 clases)
EP (emoticono)	1	1
URL	4	4
FF	30	47
jaja	101	11.964
EN (emoticono)	70	649

Tabla 3: Relevancia de términos para la detección del sentimiento en el corpus TASS 2012

3.4.2. Modificación de los diccionarios genéricos

Como en el apartado anterior, se han extraído un listado de los elementos más discriminantes, del corpus de entrenamiento, aunque en este caso solo se consideraron los tuits con polaridad positiva (P, P+) o negativa (N, N+). De esta manera se conocen los elementos polares más representativos. A continuación se

asigna, automáticamente, una OS adaptada (OSA) a cada uno de ellos; en función de sus apariciones en textos positivos y negativos, y se aplica el siguiente conjunto de reglas:

1. Si la OSA es inferior a 0,5 en valor absoluto entonces el término correspondiente no modifica el lexicón de opiniones.
2. Si el vocablo ya existe en los diccionarios semánticos genéricos y sus respectivas OSA y OS genérica tienen el mismo signo de polaridad, entonces el valor de esta última prevalece.
3. Si las dos reglas anteriores no se cumplen, el término correspondiente pasa a formar parte del diccionario con su OSA.

Con este método se extrajeron más de 10.000 nuevos términos polares, que fueron ordenados, como en el caso anterior, en función de la ganancia de información que proporcionan respecto a la clase. Además, los diccionarios semánticos también se adaptaron para resolver el problema del principio de cortesía (Brown y Levinson, 1987). Existe una tendencia humana a suavizar las críticas negativas mediante la utilización de eufemismos o negaciones de términos positivos. Para compensarlo, algunos sistemas léxicos (Taboada et al., 2011) potencian la OS de los elementos negativos. En nuestra propuesta también se determinó, empíricamente, que incrementar en un 100 % la OS de este tipo de términos, mejora ligeramente el rendimiento.

El enriquecimiento y adaptación de los diccionarios semánticos genéricos logró un incremento notable cuando fue considerado aisladamente. Sin embargo, su efecto se disipó casi por completo al aplicarlo conjuntamente junto con la selección de los términos más relevantes. De todos modos, se comprobó empíricamente que, incluir entre el 50 % y el 70 % del total de los términos polares negativos extraídos, tiene un ligero beneficio en el rendimiento; lo que refuerza la hipótesis de la tendencia positiva en el lenguaje humano.

4. Resultados experimentales

Para la evaluación de nuestra propuesta hemos utilizado el conjunto de test del corpus TASS 2012. Se han realizado experimentos para las dos tareas de clasificación de polaridad propuestas en el taller: cuatro categorías (P, NEU, N y NONE) y seis categorías (P+, P, NEU, N, N+ y NONE). La tablas 4 y 5 ilustran los resultados para ambas tareas, desglosados según las distintas versiones desarrolladas. En ambos casos, la propuesta inicial consigue un buen rendimiento. La incorporación de la información morfosintáctica mejora la detección de la polaridad para textos positivos y negativos. Ello refuerza la idea de que los usuarios tienden a emplear ciertas etiquetas y patrones sintácticos según el tipo de opinión a transmitir. La *accuracy* obtenida en la versión final sugiere que, aunque con los lexicones de opiniones genéricos y la morfosintaxis se alcanza un buen rendimiento, es necesario incorporar conocimiento semántico especifico del dominio para optimizar la propuesta.

Medida	**Base**	**+ims**	**+ad**
F_{p}	0,631	0,680	0,745
F_{neu}	0,000	0,000	0,054
F_{n}	0,566	0,603	0,671
F_{none}	0,574	0,564	0,620
Accuracy	0,587	0,615	0,676

Tabla 4: Resultados obtenidos sobre la evaluación del conjunto de test (4 clases)

Medida	**Base**	**+ims**	**+ad**
F_{p+}	0,609	0,637	0,705
F_{p}	0,000	0,040	0,307
F_{neu}	0,000	0,009	0,089
F_{n}	0,452	0,478	0,512
F_{n+}	0,000	0,120	0,441
F_{none}	0,575	0,605	0,648
Accuracy	0,523	0,546	0,600

Tabla 5: Resultados obtenidos sobre la evaluación del conjunto de test (6 clases)

En ambas tareas, la clasificación de los tuits neutros alcanza un bajo rendimiento, lo que creemos que es debido a dos factores. El primero está relacionado con una característica intrínseca de este tipo de críticas: la mezcla de ideas a favor y en contra dificulta la clasificación de estos textos, que pueden ser confundidos con opiniones positivas o negativas; más aún cuando se trata de tuits, donde un usuario no dispone de espacio para desarrollar su argumento. El segundo es referido al criterio de clasificación establecido en el corpus TASS 2012, donde el límite entre un tuit neutro y uno sin opinión, o con polaridad positiva o negativa, es difuso. Ello

afecta negativamente a las tareas de clasificación, dada la falta de un criterio objetivo que permita diferenciar las distintas polaridades. En una línea similar, este problema ya ha sido comentado por otros autores que han trabajado sobre el mismo corpus (Saralegi Urizar y San Vicente Roncal, 2012). La tabla 6 compara la *accuracy* de nuestra aproximación con las de los participantes del TASS 2012. Algunos de ellos enviaron varias propuestas al taller, aunque aquí solo se indica aquella con la que obtuvieron un mayor rendimiento. En (Villena-Román et al., 2013) se encuentran en detalle los resultados para cada una de ellas.

Propuesta	***Acc.* 4 cat**	***Acc.* 6 cat**
ELHUYAR	0,711	0,653
L2F-INESC	0,691	0,622
Nuestra propuesta	0,676	0,600
LA SALLE-URL	0,619	0,570
SINAI-UJAEN	0,606	0,549
LSI UNED	0,590	0,538
LSI UNED2	0,501	0,404
IMDEA	0,459	0,360
UMA	0,351	0,167

Tabla 6: Comparativa con los participantes del TASS 2012

5. Conclusiones y trabajo futuro

Este artículo presenta una propuesta que emplea conocimiento lingüístico para entrenar un clasificador que detecte el sentimiento de tuits escritos en español. Los resultados experimentales muestran un buen rendimiento y sugieren que la estructura morfosintáctica de los textos es útil para detectar la polaridad.

De cara al futuro hay varios aspectos que nos gustaría explorar. El preprocesamiento actual tuits es bastante simple. Nos gustaría determinar cómo una normalización gramatical de los tuits podría ayudar a clasificar la polaridad. A este respecto, en (Oliva et al., 2013) se propone un sistema de normalización de SMS que podría servir para enriquecer nuestro preprocesado. También creemos que puede ser de utilidad integrar en nuestra aproximación la propuesta de (Batista y Ribeiro, 2013), donde se propone entrenar varios clasificadores binarios y combinar los resultados obtenidos, explotando así las diferencias entre textos que expresan sentimientos distintos. La forma de emplear las dependencias sintácticas como entrada al clasificador también es un aspecto en el que nos gustaría profundizar. Actualmente solamente se utiliza como atributos el número total de cada tipo de dependencia en un tuit. Nos gustaría explorar cómo utilizar las tripletas padre/dependencia/dependiente puede ayudar en tareas de clasificación de polaridad.

Agradecimientos

Trabajo parcialmente financiado por el Ministerio de Economía y Competitividad y FEDER (TIN2010-18552-C03-02) y por la Xunta de Galicia (CN2012/008, CN 2012/319).

Bibliografía

Agarwal, A., B. Xie, I. Vovsha, O. Rambow, y R. Passonneau. 2011. Sentiment analysis of Twitter data. En *Proceedings of the Workshop on Languages in Social Media*, LSM '11, páginas 30–38, Stroudsburg, PA, USA. ACL.

Bakliwal, A., P. Arora, S. Madhappan, N. Kapre, M. Singh, y V. Varma. 2012. Mining sentiments from tweets. En *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, páginas 11–18, Jeju, Korea. ACL.

Batista, F. y R. Ribeiro. 2013. The L2F Strategy for Sentiment Analysis and Topic Classification. *Procesamiento de Lenguaje Natural*, 50:77–84.

Brooke, J., M. Tofiloski, y M. Taboada. 2009. Cross-Linguistic Sentiment Analysis: From English to Spanish. En *Proceedings of the International Conference RANLP-2009*, páginas 50–54, Borovets, Bulgaria. ACL.

Brown, P. y S Levinson. 1987. *Politeness, Some universals in language use*. Cambridge, Cambridge University Press.

Castellano González, A., J. Cigarrán Recuero, y A. García Serrano. 2012. Using IR techniques for topic-based sentiment analysis through divergence models. En *TASS 2012 Working Notes*, Castellón, Spain.

Fernández Anta, A., P. Morere, L. Núñez Chiroque, y A. Santos. 2012. Techniques for Sentiment Analysis and Topic Detection of Spanish Tweets: Preliminary Report. En *TASS 2012 Working Notes*, Castellón, Spain.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, y I.H. Witten. 2009. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, Noviembre.

Martín-Wanton, T. y J. Carrillo de Albornoz. 2012. Sistema para la Clasificación de la Polaridad y Seguimiento de Temas. En *TASS 2012 Working Notes*, Castellón, Spain.

Martínez Cámara, E., M. T. Martín Valdivia, M. A. García Cumbreras, y L. A. Ureña López. 2012. SINAI at TASS 2012. *Procesamiento de Lenguaje Natural*, 50:53–60.

Moreno-Ortiz, A. y C. Pérez-Hernández. 2012. Lexicon-Based Sentiment Analysis of Twitter Messages in Spanish. En *TASS 2012 Working Notes*, Castellón, Spain.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, y Erwin Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Oliva, J., J. I. Serrano, M. D. Del Castillo, y A. Igesias. 2013. A SMS normalization system integrating multiple grammatical resources. *Natural Language Engineering*, 19:121–141, 0.

Pak, A. y P. Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. En *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, Mayo. European Language Resources Association (ELRA).

Pang, B., L. Lee, y S. Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. En *Proceedings of EMNLP*, páginas 79–86.

Platt, J. C. 1999. Advances in kernel methods. MIT Press, Cambridge, MA, USA, capítulo Fast training of support vector machines using sequential minimal optimization, páginas 185–208.

Saralegi Urizar, X. y I. San Vicente Roncal. 2012. Detecting Sentiments in Spanish Tweets. En *TASS 2012 Working Notes*, Castellón, Spain.

Spencer, J. y G. Uchyigit. 2012. Sentimentor: Sentiment Analysis on Twitter Data. En *The 1st International Workshop on Sentiment Discovery from Affective Data*, Bristol, United Kingdom.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, y M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Thelwall, M., K. Buckley, y G. Paltoglou. 2011. Sentiment in Twitter events. *J. Am. Soc. Inf. Sci. Technol.*, 62(2):406–418.

Thelwall, M., K. Buckley, y G. Paltoglou. 2012. Sentiment strength detection for the social web. *J. Am. Soc. Inf. Sci. Technol.*, 63(1):163–173.

Thelwall, M., K. Buckley, G. Paltoglou, D. Cai, y A. Kappas. 2010. Sentiment in short strength detection informal text. *J. Am. Soc. Inf. Sci. Technol.*, 61(12):2544–2558.

Trilla, A. y F. Alías. 2012. Sentiment Analysis of Twitter messages based on Multinomial Naive Bayes. En *TASS 2012 Working Notes*, Castellón, Spain.

Turney, P. D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. En *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, páginas 417–424, Stroudsburg, PA, USA. ACL.

Vilares, D., M. A. Alonso, y C. Gómez-Rodríguez. 2013. Clasificación de polaridad en textos con opiniones en español mediante análisis sintáctico de dependencias. *Procesamiento de Lenguaje Natural*, 50:13–20.

Villena-Román, J., S. Lana-Serrano, J. C. González Cristóbal, y E. Martínez-Cámara. 2013. TASS - Worshop on Sentiment Analysis at SEPLN. *Procesamiento de Lenguaje Natural*, 50:37–44.

Zhang, Lei, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, y Bing Liu. 2011. Combining lexicon-based and learning-based methods for Twitter sentiment analysis. Informe Técnico HPL-2011-89, HP Laboratories, Palo Alto, CA.

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 135-142 recibido 25-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Adapting Text Simplification Decisions to Different Text Genres and Target Users

Adaptación de algoritmos de toma de decisiones de simplificación de textos a diferentes corpus y audiencias

Sanja Štajner
University of Wolverhampton, UK
sanjastajner@wlv.ac.uk

Horacio Saggion
Universitat Pompeu Fabra, Spain
horacio.saggion@upf.edu

Resumen: Hemos analizado las alineaciones a nivel de oración de dos corpus paralelos de textos originales y sus simplificaciones creados con diferentes objetivos. Hemos clasificado las alineaciones que se observan y diseñado un algoritmo de clasificación capaz de predecir si las oraciones de un texto serán eliminadas, segmentadas, o transformadas durante el proceso de simplificación. Hemos realizado una evaluación cruzada en cada uno de los corpus así como una evaluación en la cual se entrena en algoritmo en un corpus y se lo evalúa en el otro.
Palabras clave: Simplificación de textos, clasificación de oraciones, adaptación de métodos

Abstract: We investigate sentence deletion and split decisions in Spanish text simplification for two different corpora aimed at different groups of users. We analyse sentence transformations in two parallel corpora of original and manually simplified texts for two different types of users and then conduct two classification experiments: classifying between those sentences to be *deleted* and those to be *kept*; and classifying between sentences to be *split* and those to be left *unsplit*. Both experiments were first run on each of the two corpora separately and then run by using one corpus for the training and the other for testing. The results indicated that both sentence decision systems could be successfully trained on one corpus and then used for a different text genre in a text simplification system aimed at a different target population.
Keywords: Text simplification, sentence classification, method adaptation

1 Introduction

Since the late nineties several initiatives which proposed guidelines for producing plain, easy-to-read and more accessible documents have emerged, e.g. "The Plain Language Action and Information Network (PLAIN)"[1], "Make it Simple, European Guidelines for the Production of Easy-to-Read Information for people with Learning Disability" (Freyhoff et al., 1998), "Am I making myself clear? Mencap's guidelines for accessible writing"[2], and "Web content accessibility guidelines"[3]. All these initiatives increased the interest in the use of natural language processing in the development of assistive technologies and automatic text simplification, as it is clear that manual simplification cannot match the rate of production of texts, particularly of newswire texts which are being constantly generated.

The first systems aimed at automatic text simplification were rule-based, e.g. (Chandrasekar, 1994; Devlin, 1999; Devlin and Unthank, 2006). Syntactic simplification modules usually consisted of a set of rules which are recursively applied to each sentence as long as it is possible. Lexical simplification modules were traditionally based on substitution of difficult infrequent words with their simpler synonyms.

With the emergence of Simple English Wikipedia[4] the approaches to automatic text simplification became more data-driven. Biran et al. (2011) and Yatskar et al. (2010), apply an unsupervised method for learning pairs of complex and simple synonyms from a corpus of texts from the original Wikipedia

[1] http://www.plainlanguage.gov/

[2] http://november5th.net/resources/Mencap/Making-Myself-Clear.pdf

[3] http://www.w3.org/TR/WCAG20/

[4] http://simple.wikipedia.org

ISSN 1135-5948

and Simple English Wikipedia. Coster and Kauchak (2011a; 2011b) address the problem of text simplification as an English-to-English translation problem. They use the standard machine translation tools trained on the parallel corpus of aligned sentences from original and Simple English Wikipedia, to build an automatic text simplification system. Although the results show that the machine translation approach to text simplification works well for English, the same approach cannot be applied to other languages, as Simple Wikipedia does not exist for many languages (Spanish among them). Another limitation is that, although it imposes the use of Basic English vocabulary, shorter sentences and simpler grammar, Simple English Wikipedia does not follow easy-to-read guidelines for writing for people with cognitive disabilities. Therefore, it may not represent a good training material for text simplification for this target audience.

The compilation of a parallel corpus of original and manually simplified texts for specific target audiences (e.g. people with learning or language disabilities) is both time-consuming and expensive (involving special training for human annotators and adaptation of easy-to-read guidelines for a specific language and target population). Therefore, it would be important to investigate whether the simplification systems (or some of their components) developed for one specific target population and text genre could also be used for text simplification aimed at other target populations and different text types – a problem never addressed before. This paper fills that gap, exploring whether sentence deletion and split decisions learned from a parallel corpus of news texts compiled for the needs of a specific user group could be used for different user groups and text genres. As shown in this paper, the decisions learned can be transferred to a new corpus if an appropriate learning algorithm is used.

The reminder of the paper is organised as follows: Section 2 presents the most relevant previous work on the topic of sentence decisions in text simplification; Section 3 describes the corpora used in this study and presents the results of the initial analysis of detected sentence transformations in both corpora; Section 4 introduces the features and the settings for the two classification experiments; Section 5 presents and discusses the results of the classification experiments; and Section 6 draws attention to the main findings of the presented study and offers possible directions for future work.

2 Related Work

Various studies have described necessary transformations to be included in an automatic text simplification system for the English language. They analysed the parallel corpora of original and manually simplified texts aimed at different target audiences: (1) for children (Bautista et al., 2011), using Encyclopedia Britannica and Britannica Elemental (Barzilay and Elhadad, 2003); (2) for language learners (Petersen and Ostendorf, 2007), using original and abridged texts from Literacyworks[5]; and (3) for audiences with various reading difficulties (Biran, Brody, and Elhadad, 2011; Yatskar et al., 2010; Coster and Kauchak, 2011a; Coster and Kauchak, 2011b), using original and Simple English Wikipedia.

Petersen and Ostendorf (2007) reported that 30% of sentences were completely eliminated, while 19% of sentences were split into two or more sentences by the human editors while simplifying texts for language learners in English. Caseli et al. (2009) showed sentence splitting to be the second most frequent simplification operation, present in 34% of the original sentences (straight after lexical substitution present in 46% of the sentences), while only 0.28% of sentences were completely eliminated, during the manual simplification of text for people with low literacy levels in Brazilian Portuguese. Štajner et al. (2013) performed a similar analysis on a small corpus of original and manually simplified texts (37 text pairs) in Spanish, aimed at people with cognitive disabilities. They reported sentence deletion and sentence splitting as being almost equally present simplification operations (21% and 23% of original sentences, respectively).

Motivated by those previous studies, this article: (1) analyses the types of applied manual transformations in Spanish text simplification aimed at two different target populations: people with intellectual disabilities (Down's syndrom), and people with autism spectrum disorders (ASD); and (2) proposes

[5]http://literacynet.org/cnnsf/index_cnnsf.html

the algorithms for classification of original sentences into those which should be *deleted*, *split*, and left *largely unchanged*.

More importantly, this study goes one step further by testing whether the sentence classification system built on one specific text genre and aimed at one specific target population can successfully be applied in other text genres and for different target populations for which parallel corpora of original and manually simplified texts may not exist. To the best of our knowledge, this is the first study addressing the problem of method adaptation in text simplification.

3 Corpora

The main corpus (Corpus A henceforth) used in the experiments contains 195 original and manually simplified news articles in Spanish (a total of 1118 orignal sentences), provided by the Spanish news agency Servimedia[6] and compiled under the Simplext project[7] (Saggion et al., 2011). Simplifications have been applied by trained human editors, familiar with the particular needs of a person with cognitive disabilities (Down's syndrom) and following a series of easy-to-read guidelines. The corresponding pairs of original and simplified texts were first sentence aligned using an automatic alignment tool (Bott and Saggion, 2011) and then manually post-edited in order to provide 100% accurate sentence alignment.

The second corpus (Corpus B henceforth) is significantly smaller and comprises 25 original and manually simplified texts (a total of 320 original sentences) of different genres: literature, news, health, general culture and instructions. It was compiled under the FIRST project[8] (Orasan, Evans, and Dornescu, 2013). Texts were manually simplified by five experts who have experience of working with people with autism, having in mind the particular needs of this target population. The corresponding pairs of original and simplified texts were sentence aligned manually, thus ensuring alignment to be 100% accurate.

3.1 Sentence Transformations

By automatically processing the aligned sentences in Corpus A it was found that: (1) the original sentence was neither split nor deleted ("1-1" alignment) in 566 cases; (2) the original sentence was split into two or more sentences ("1-n" alignment) in 358 cases; and (3) the original sentence was completely deleted ("1-0" alignment) in 186 cases. The same analysis of the aligned sentences in Corpus B (total of 305 sentences) revealed that: (1) the original sentence was neither split nor deleted ("1-1" alignment) in 192 cases; (2) the original sentence was split into two or more sentences ("1-n" alignment) in 70 cases; and (3) the original sentence was completely deleted ("1-0" alignment) in 43 cases (Table 3.1).

Transformation		Corpus A	Corpus B
"1-0"	deleted	186 (17%)	43 (14%)
"1-n"	split	358 (32%)	70 (23%)
"1-1"	same	275 (25%)	178 (58%)
	reduced	291 (26%)	14 (5%)
Total ("1-x")		1110 (100%)	305 (100%)

Table 1: Corpus analysis

More detailed analysis of "1-1" aligned sentences, revealed that in many cases original sentences were significantly longer than their simplified versions, thus indicating that certain parts of the original sentences were omitted during the simplification process, as in the following example of original (1) and its corresponding simplified sentence (2):

1. *"El Premio de la Cinematografía y de las Artes Audiovisuales está destinado a recompensar la aportación más sobresaliente en el ámbito cinematográfico español puesta de manifiesto a través de una obra hecha pública durante 2009, o de una labor profesional desarrollada durante ese mismo año."*

2. *"El Premio Nacional de Cine se da a la mejor película o trabajo del año 2009."*

Therefore, the "1-1" aligned sentences were further divided into two groups: *same* – those sentences which were only slightly modified (the difference between number of words in the original and simplified sentence is less than ten words); and *reduced* – those sentences whose lengths were significantly reduced during the simplification (the difference between number of words in the original and simplified sentence is ten or more words). Unlike Corpus A, which contains a

[6] http://www.servimedia.es/
[7] http://www.simplext.es/
[8] http://first-asd.eu/

significant number of *reduced* sentences, Corpus B contains only 14 cases of these sentences (Table 3.1). These sentences were thus excluded from Corpus B in all classification experiments.

Analysis of sentence transformations in both corpora revealed an additional, frequently occurring type of transformation – *enlarged* sentences (simplified sentence is at least ten words longer than its original). All of those were the result of adding a definition of a complex term, as in the following example of original (1) and its corresponding simplified sentence (2):

1. *"He visitado cientos de mundos, he sido dama victoriana, rey medieval y bucanero."*
2. *"Al leer novelas he visitado cientos de mundos, he sido una dama de la época victoriana (época transcurrida entre 1837 y 1901), un rey medieval (de la época transcurrida entre el siglo V y el siglo XV) y un bucanero (un pirata que en los siglos XVII y XVIII robaba las posesiones españolas de ultramar)."*

These *enlarged* sentences did not significantly differ from the *same* sentences in terms of the features used in this paper. Therefore, they were counted as occurences of the *same* sentences and treated as such in all classification experiments.

3.2 Additional Types of Sentence Transformations

While the aforementioned sentence transformations were expected to be found in the corpora, it was surprising to discover that in several cases (four in Corpus A and six in Corpus B) two original sentences were merged into one simplified sentence ("2-1" alignment), as in the following pair of two original sentences (1) and their corresponding simplified sentence (2):

1. *"El ámbito médico más afectado por las agresiones de pacientes, es, en virtud del observatorio creado por los colegios de facultativos, el de Atención Primaria, donde se contabilizaron en 2010 el 65% de los atentados a profesionales sanitarios. Y el grupo de edad más castigado, el que va desde los 46 a los 55 años."*
2. *"Los médicos que sufren más ataques son los de alrededor de 50 años y los que trabajan en centros médicos pequeños."*

In addition to the very frequent type of *enlarged* sentences, in several cases, even whole sentences were added as a definition. Especially interesting are the cases in which the addition of a definition (in a separate sentence) occurred simultaneously with sentence splitting as in the following case of original sentence (1) and its corresponding simplified paragraph (2) in Corpus B:

1. *"Este nombre se da a una mezcla gaseosa, líquida y sólida de hidrocarburos, que se ha encontrado en depósitos de rocas sedimentarias, en diferentes proporciones y en distintos lugares de la Tierra."*
2. *"El petróleo es una mezcla: Gaseosa, líquida y sólida de hidrocarburos. Los hidrocarburos son una mezcla de hidrógeno y carbono. El petróleo se ha encontrado en depósitos de rocas sedimentarias (en capas de rocas), en diferentes cantidades y en diferentes lugares de la Tierra."*

These *merged* and *added* sentences were not used in any of the classification experiments presented in this paper.

4 Experimental Settings

The corpora were parsed with state-of-the-art Connexor's Machinese parser[9] and the features (Table 3.2) were automatically extracted using the parser's output. Each sentence is represented as vector of 24 features inspired by the works of Štajner et al. (2013), Gasperin et al. (2009), Petersen and Ostendorf (2007), and Drndarevic and Saggion (2012). Features 1-19 and 21-22 count the number of occurrences of the feature in the sentence (e.g. feature 1 counts how many verbs the sentence has while feature 10 counts the number of determiners in the sentence). Feature 20 represents the position of the sentence in the text.

All classification experiments were conducted in Weka Experimenter (Witten and Frank, 2005), employing four different classification algorithms: Naive Bayes (John and Langley, 1995); SMO (Weka implementation of Support Vector Machines) with normalisation and using poly kernels (Keerthi et al.,

[9] www.connexor.eu

#	Code	Feature	#	Code	Feature	#	Code	Feature
1	*v*	verb	9	*pron*	pronoun	17	*main*	head of the verb phrase
2	*ind*	indicative	10	*det*	determiner	18	*nh*	head of the noun phrase
3	*sub*	subjunctive	11	*n*	noun	19	*advl*	head of the adverbial phrase
4	*inf*	infinitive	12	*prep*	preposition	20	*sent*	position of the sentence
5	*pcp*	participle	13	*cc*	coord. conj.	21	*punc*	punctuation marks
6	*ger*	gerund	14	*cs*	subord. conj.	22	*num*	numerical expressions
7	*adj*	adjective	15	*prem*	pre-modifier	23	*char*	sentence length in characters
8	*adv*	adverb	16	*postm*	post-modifier	24	*words*	sentence length in words

Table 2: Feature set

Classifier	Corpus A			Corpus B			A tested on B			B tested on A		
	P	R	F	P	R	F	P	R	F	P	R	F
SMO*	0.69	0.83	0.76	0.76	0.87	0.81	0.76	0.87	0.81	0.69	0.83	0.76
NB	0.76	0.81	0.78	0.82	0.62	0.68	0.80	0.83	0.81	0.71	0.67	0.69
JRip	**0.79**	**0.83**	**0.80**	0.81	0.85	0.82	0.76	0.75	0.75	**0.86**	**0.84**	**0.76**
J48	0.77	0.79	0.77	**0.84**	**0.87**	**0.84**	0.76	0.70	0.73	**0.79**	**0.83**	**0.76**

Table 3: Results of the classification between deleted and kept sentences (Key: Corpus A = 10-fold cross-validation with ten repetitions using only corpus A; Corpus B = 10-fold cross-validation with ten repetitions using only corpus B; A on B = training set: corpus A, test set: corpus B; B on A = training set: corpus B, test set: corpus A)

2001; Platt, 1998), JRip (Cohen, 1995), and J48 (Weka implementation of C4.5) (Quinlan, 1993). The experiments were the following:

- Experiment I: Classification between *deleted* ("1-0") and *kept* ("1-1" and "1-n") sentences;
- Experiment II: Classification between *split* and *unsplit* (*same*) sentences.

5 Results and Discussion

Results for each of the experiments are presented and discussed separately in the next two subsections (Sections 5.1 and 5.2).

5.1 Sentence Deletion

The weighted average P (precision), R (recall), and F (F-measure) for each classifier and each setup are given in Table 3.2. It is important to note that the P, R, and F values for the class *deleted* in SMO were 0, and thus can be taken as a baseline which does not delete any sentences (majority class). For each experiment, the results of the classifier which outperformed the baseline (row 'SMO*' in Table 3.2) on all three measures (P, R, and F) are shown in bold.

JRip achieved a significantly better precision (P) than SMO in the cross-validation setup on Corpus A, and when trained on Corpus B and tested on Corpus A. However, when trained on Corpus A and tested on Corpus B, the JRip classifier had a significantly lower performance (P, R, and F) than when used with a 10-fold cross-validation setup only on Corpus A. In general, the 10-fold cross-validation setup on each of the corpora separately, achieved better classification results than the setup with training on one corpus and testing on the other. None of the three classifiers (NB, JRip, and J48) outperformed the baseline (SMO) on any of the two setups ('A on B' and 'B on A') in terms of F-measure, although JRip and J48 achieved a significantly better precision (P) than the baseline.

Two additional experiments were conducted in order to explore whether: (1) elimination of the *reduced* sentences from the Corpus A; or (2) reduction of the feature set to the subset of best features (obtained by using the CfsSubsetEval attribute selection algorithm in Weka (Hall and Smith, 1998)), could improve the classification accuracy. Given that the results of these experiments were not significantly different from the results of the initial experiments (Table 3.2), they are not presented here.

Previous works on deletion decisions in

Classifier	Corpus A			Corpus B			A tested on B			B tested on A		
	P	R	F	P	R	F	P	R	F	P	R	F
SMO	**0.94**	**0.93**	**0.93**	**0.94**	**0.93**	**0.93**	0.94	0.94	0.94	0.94	0.94	0.94
NB	0.93	0.93	0.93	0.93	0.93	0.93	0.94	0.93	0.93	0.94	0.93	0.93
JRip	0.91	0.90	0.91	0.91	0.90	0.91	0.94	0.94	0.94	0.94	0.94	0.94
J48	0.91	0.91	0.91	0.91	0.91	0.91	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**

Table 4: Results of the classification between split and unsplit sentences (Key: Corpus A = 10-fold cross-validation with ten repetitions using only corpus A; Corpus B = 10-fold cross-validation with ten repetitions using only corpus B; A on B = training set: corpus A, test set: corpus B; B on A = training set: corpus B, test set: corpus A)

Spanish using cross-validation achieved F-scores of 0.79 (Drndarević and Saggion, 2012), and 0.82 (Štajner, Drndarević, and Saggion, 2013). We therefore consider the performance of our classification algorithms and feature set reasonable, in spite of not being directly comparable to those previous works because of differences in corpus characteristics.

5.2 Sentence Splitting

For the experiment on classification between *split* and *unsplit* sentences, the *reduced* and *deleted* sentences were excluded from both corpora. The decision not to include *reduced* sentences into either of the two classes (*split* and *unsplit*) arose from the nature of the *reduced* sentences. They could be interpreted as sentences which were first split and then one part was deleted and the second maintained. Therefore, it is expected that the *reduced* sentences contain markers of all three other types of sentences – *deleted*, *split*, and *same*. Also, the percentage of *reduced* sentences in each of the corpora was very unbalanced (Table 3.1 in Section 3).

The results of this classification experiment (Table 4) were quite surprising. All classification algorithms achieved better performances when trained on one corpus and tested on the other corpus. This was particularly accentuated in the case of the J48 classification algorithm which achieved the F-measure of 0.96 in both setups – 'A on B' and 'B on A'. The Support Vector Machines (SMO) performed as the best classifier on each of the corpora separately (columns 'Corpus A' and 'Corpus B' in Table 4). Naive Bayes achieved very similar results as the SMO classifier in all setups. The J48 classifier (Weka implementation of C4.5 decision tree classifier) significantly outperformed all three other classifiers in 'A on B' and 'B on A' setups. Note that a baseline that choses the majority case (split for corpus A and non-split for corpus B) would have obtained F=0.56 on corpus A, F=0.43 on corpus B. Previous work on split decisions by Gasperin et al. (2009), although not directly comparable to ours because of the different language and corpus, achieved an F-score of 0.80. Štajner et al. (2013) achieved an F-measure of 0.92 for the same task on a smaller portion of Corpus A, using a slightly different set of features. We therefore consider the performance of our classifier and set of features on our datasets acceptable.

6 Conclusions and Future Work

In this paper we addressed the issue of sentence deletion and split decisions as a first step in building an automatic text simplification system for Spanish. More particularly, we investigated the adaptability of these decisions across different text genres and two different target populations.

The initial analysis of sentence transformations in two corpora containing different text genres and aimed at different target users revealed some interesting differences in simplification strategies which were applied by human annotators in these two cases. Furthermore, it revealed different distribution of those sentence transformations which were present in both corpora.

The classification of original sentences into those to be *deleted* and those to be *kept* achieved better accuracy when performed on each of the corpora separately using 10-fold cross-validation setup than when trained on one corpus and tested on the other. It also indicated the JRip and J48 classifiers as being the most suitable for this task (out of the four classifiers applied).

The classification of original sentences into those to be *split* and those to be left *unsplit* led to surprising results. All four classifiers achieved better accuracies when trained on one corpus and tested on the other than when performed on each of the corpora separately in a 10-fold cross-validation setup. The difference in the classifier performance between the two setups was most pronounced in the case of the J48 (decision tree) classifier.

In the future, we plan to perform similar experiments on a larger number of corpora aimed at other target populations – second language learners, children, and users with different reading and learning disabilities. The main goal would be to discover how much of the methodology and system components could be shared between the automatic text simplification systems aimed at different target users (and different text genres).

Acknowledgements

The research described in this paper was partially funded by the European Commission under the Seventh (FP7 - 2007-2013) Framework Programme for Research and Technological Development (FIRST 287607). This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein. This work is supported by an Advanced Research Fellowship from Programa Ramón y Cajal (RYC-2009-04291) and by the project SKATER: Scenario Knowledge Acquisition – Knowledge-based Concise Summarization (TIN2012-38584-C06-03) , Ministerio de Economía y Competitividad, Secretaria de Estado de Investigaciṕn, Desarrollo e Innovación, Spain.

References

Barzilay, R. and N. Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bautista, S., C. León, R. Hervás, and P. Gervás. 2011. Empirical identification of text simplification strategies for reading-impaired people. In *European Conference for the Advancement of Assistive Technology*.

Biran, O., S. Brody, and N. Elhadad. 2011. Putting it Simply: a Context-Aware Approach to Lexical Simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 496–501, Portland, Oregon, USA. Association for Computational Linguistics.

Bott, Stefan and Horacio Saggion. 2011. An unsupervised alignment algorithm for text simplification corpus construction. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, MTTG '11, pages 20–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Caseli, H. M., T. F. Pereira, L. Specia, T. A. S. Pardo, C. Gasperin, and S. M. Aluísio. 2009. Building a Brazilian Portuguese parallel corpus of original and simplified texts. In *10th Conference on Intelligent Text PRocessing and Computational Linguistics (CICLing 2009)*.

Chandrasekar, R. 1994. *A Hybrid Approach to Machine Translation using Man Machine Communication*. Ph.D. thesis, Tata Institute of Fundamental Research/University of Bombay, Bombay.

Cohen, W. 1995. Fast Effective Rule Induction. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123.

Coster, W. and D. Kauchak. 2011a. Learning to Simplify Sentences Using Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1–9.

Coster, W. and D. Kauchak. 2011b. Simple English Wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA: Association for Computational Linguistics*, pages 665–669.

Devlin, S. 1999. *Simplifying natural language text for aphasic readers*. Ph.D. thesis, University of Sunderland, UK.

Devlin, S. and G. Unthank. 2006. Helping aphasic people process online information. In *Proceedings of the 8th international*

ACM SIGACCESS conference on Computers and accessibility, Assets '06, pages 225–226, New York, NY, USA. ACM.

Drndarević, B and H. Saggion. 2012. Reducing Text Complexity through Automatic Lexical Simplification: an Empirical Study for Spanish. *SEPLN Journal*, 49.

Freyhoff, G., G. Hess, L. Kerr, B. Tronbacke, and K. Van Der Veken, 1998. *Make it Simple, European Guidelines for the Production of Easy-toRead Information for People with Learning Disability*. ILSMH European Association, Brussels.

Gasperin, C., L. Specia, T. Pereira, and S.M. Aluisio. 2009. Learning When to Simplify Sentences for Natural Text Simplification. In *Proceedings of the Encontro Nacional de Inteligência Artificial (ENIA-2009), Bento Gonçalves, Brazil.*, pages 809–818.

Hall, M. A. and L. A. Smith. 1998. Practical feature subset selection for machine learning. In C. McDonald, editor, *Computer Science '98 Proceedings of the 21st Australasian Computer Science Conference ACSC'98*, pages 181–191. Berlin: Springer.

John, G. H. and P. Langley. 1995. Estimating Continuous Distributions in Bayesian Classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345.

Keerthi, S. S., S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy. 2001. Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation*, 13(3):637–649.

Orasan, C., R. Evans, and I. Dornescu. 2013. Text Simplification for People with Autistic Spectrum Disorders. In D. Tufis, V. Rus, and C. Forascu, editors, *Towards Multilingual Europe 2020: A Romanian Perspective*. Romanian Academy Publishing House, Bucharest, pages 187–312.

Petersen, S. E. and M. Ostendorf. 2007. Text Simplification for Language Learners: A Corpus Analysis. In *Proceedings of Workshop on Speech and Language Technology for Education*.

Platt, J. C. 1998. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In B. Schoelkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*.

Quinlan, R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Saggion, H., E. Gómez-Martínez, A. Anula, L. Bourg, and E. Etayo. 2011. Text simplification in simplext: Making texts more accessible. *Procesamiento del Lenguaje Natural*, 46.

Štajner, S., B. Drndarević, and H. Saggion. 2013. Corpus-based Sentence Deletion and Split Decisions for Spanish Text Simplification. *Computacion y Systemas*, 17(2):251–262.

Witten, I. H. and E. Frank. 2005. *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann Publishers.

Yatskar, M., B. Pang, C. Danescu-Niculescu-Mizil, and L. Lee. 2010. For the sake of simplicity: unsupervised extraction of lexical simplifications from wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 365–368, Stroudsburg, PA, USA. Association for Computational Linguistics.

Reconocimiento y Síntesis del Habla

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 145-152 recibido 30-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Incorporación de n-gramas discriminativos para mejorar un reconocedor de idioma fonotáctico basado en i-vectores

Incorporation of discriminative n-grams to improve a phonotactic language recognizer based on i-vectors

Christian Salamea Palacios[1,2], Luis Fernando D'Haro[1], Ricardo Córdoba[1], Miguel Ángel Caraballo[1]

[1]Grupo de Tecnología del Habla. Departamento de Ingeniería Electrónica
E.T.S.I. Telecomunicación. Universidad Politécnica de Madrid.
Ciudad Universitaria S/N, 28040 - Madrid, España.
{csalamea, lfdharo, cordoba, macaraballo}@die.upm.es

[2]Universidad Politécnica Salesiana del Ecuador
Calle Vieja 12-30 y Elia Liut, Casilla 26, Cuenca, Ecuador
csalamea@ups.edu.ec

Resumen: Este artículo describe una nueva técnica que permite combinar la información de dos sistemas fonotácticos distintos con el objetivo de mejorar los resultados de un sistema de reconocimiento automático de idioma. El primer sistema se basa en la creación de cuentas de posteriorgramas utilizadas para la generación de i-vectores, y el segundo es una variante del primero que tiene en cuenta los n-gramas más discriminativos en función de su ocurrencia en un idioma frente a todos los demás. La técnica propuesta permite obtener una mejora relativa de 8.63% en C_{avg} sobre los datos de evaluación utilizados para la competición ALBAYZIN 2012 LRE.
Palabras clave: Posteriorgrama, i-Vectores, rankings discriminativos, fonotáctico, n-gramas.

Abstract: This paper describes a novel technique that allows the combination of the information from two different phonotactic systems with the goal of improving the results of an automatic language recognition system. The first system is based on the creation of posteriorgram counts used for the generation of i-vectors, and the second system is a variation of the first one that takes into account the most discriminative n-grams as a function of their occurrence in one language compared to all other languages. The proposed technique allows a relative improvement of 8.63% on C_{avg} over the official set used for the ALBAYZIN 2012 LRE evaluation.
Keywords: Posteriorgram, i-Vectors, discriminate rankings, phonotactic, n-grams

1 Introducción

El presente artículo describe una técnica novedosa que permite mejorar las tasas de reconocimiento de idioma mediante la unificación de dos técnicas que emplean información a nivel fonotáctico a partir de la salida de un reconocedor de fonemas que permite determinar las secuencias más probables de éstos para un determinado conjunto de ficheros de audio. En este caso, nos hemos decantando por el uso de técnicas fonotácticas, ya que son ampliamente usadas en el reconocimiento de idioma y/o locutor por las ventajas que presentan, su versatilidad, la posibilidad de incorporar información a alto nivel y el hecho que de forma congruente siempre permite mejorar las tasas de reconocimiento cuando se combina con otras técnicas basadas únicamente en información acústica (Kinnunen and Li, 2010).

La técnica que proponemos, aparte del uso de la información fonotáctica, aprovecha un elemento en común entre las dos técnicas que

ISSN 1135-5948

hacen posible su unificación: son las llamadas "cuentas" que se calculan a partir de la ocurrencia de la secuencia de fonemas (i.e. n-gramas) reconocidos mediante el reconocedor automático de fonemas. En el caso de la primera técnica, esta utiliza las cuentas con el objetivo de entrenar un modelo basado en i-vectores que es la técnica actual que mejores resultados da en reconocimiento tanto usando información acústica como fonotáctica (Dehak et al, 2011)(Martinez et al, 2011)(D'Haro et al, 2012). Este sistema fonotáctico basado en i-vectores en combinación con otros sistemas basados en información acústica fue uno de los factores determinantes en la obtención de los buenos resultados conseguidos durante la evaluación de reconocimiento de idioma Albayzin 2012 LRE, tal y como se describe en (D'Haro et al, 2013).

Por otra parte, en el caso de la segunda técnica, se utilizan las cuentas de los fonemas reconocidos con el objetivo de crear un ranking de los n-gramas más discriminativos para reconocer un idioma frente a los otros. El proceso de creación de los rankings implica la estimación de un valor de discriminación que se utiliza como factor de ordenación de los rankings. Esta técnica también se ha probado previamente con muy buenos resultados (Cordoba et al, 2007) en el contexto de un sistema de reconocimiento que utiliza múltiples reconocedores de fonemas (Zissman, 1996).

En este artículo proponemos una nueva técnica en la que se modifican los valores de las cuentas de posteriorgramas usados por el primer sistema en la generación de los i-vectores mediante la utilización de la información de discriminación de los n-gramas generados para crear los rankings de la segunda técnica. Conviene mencionar que como figura de mérito sobre la eficacia de la técnica propuesta hemos utilizado los mismos datos de la evaluación Albayzin utilizando tanto la métrica oficial de la evaluación (con el objetivo de facilitar la comparación de resultados) como la métrica C_{avg} que es una de la más empleada en las evaluaciones de reconocimiento de idioma. La métrica C_{avg} permite ponderar los errores de falsa aceptación (i.e. reconocer un determinado fichero con un idioma distinto al que realmente es) y el falso rechazo (i.e. no reconocer el idioma real de un determinado fichero).

Este artículo se organiza de la siguiente manera. En la sección 2 se hace una breve descripción del sistema base empleado. Posteriormente se describe cada uno de los dos sistemas empleados por separado, para terminar describiendo la técnica propuesta. Luego, en la sección 3 describimos la base de datos y metodología seguida para la realización de los experimentos. En la sección 4 presentamos y discutimos los resultados obtenidos. Y finalmente, en la sección 5, presentamos las conclusiones y líneas futuras.

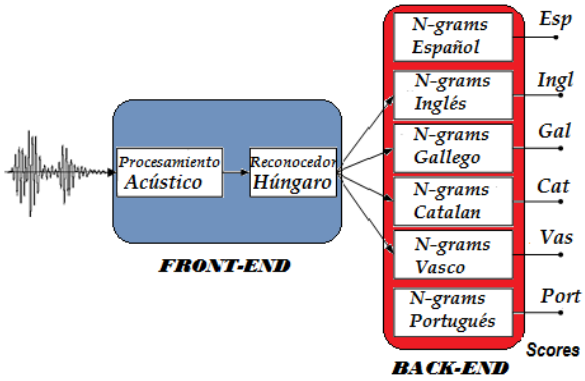

Figura 1: Sistema de reconocimiento de lenguaje basado en PRLM.

2 Descripción de los sistemas

Una gran mayoría de los sistemas de reconocimiento automático de idioma que emplean información fonotáctica lo hacen aplicando una técnica denominada PRLM (Phone Recognition Followed by a Language Model), en el que, como indica la Figura 1, se tienen dos componentes claramente diferenciados: uno denominado Front-End y otro denominado Back-End.

En el primero se realiza la parametrización de los ficheros de audio de entrada y se ejecuta un reconocedor automático de fonemas que se encarga de determinar la secuencia de fonemas más probables. Dado que esta salida contendrá errores de reconocimiento los resultados finales serán menos buenos, pero tiene dos grandes ventajas: 1) que los modelos utilizados por el reconocedor de fonemas no tienen por qué corresponder con los idiomas a reconocer (aunque evidentemente se obtienen mejores resultados cuando hay correspondencia), y 2) porque esto permite reaprovechar reconocedores mucho mejor entrenados, a la vez que se minimiza la necesidad de disponer de una gran cantidad de datos etiquetados para el entrenamiento de dichos reconocedores.

Para los experimentos presentados en este artículo hemos utilizado como Front-End el reconocedor de fonemas de la Universidad de

Brno (Schwarz, 2009), el cual se distribuye libremente e incluye modelos de fonemas de 3 idiomas distintos: checo, húngaro y ruso. En nuestro caso, hemos utilizado únicamente el modelo de fonemas de Húngaro dado que con éste se han conseguido resultados satisfactorios en experimentos previos (D'Haro et al, 2012), además de que fue el que se utilizó durante la evaluación de Albayzin LRE 2012. Este reconocedor permite identificar un total de 61 clases de fonemas aunque para nuestros experimentos hemos reducido este número a un total de 33, iguales a los enumerados en (Diez et al, 2013), unificando tres fonemas que permiten detectar ruidos y pausas en el habla, así como otros con un gran parecido lingüístico.

Por otra parte, en el Back-End se toman las secuencias de fonemas reconocidos y se entrena un modelo de lenguaje por cada idioma a reconocer. Estos modelos se utilizan uno a uno durante la fase de evaluación para calcular la perplejidad de la frase a identificar, siendo finalmente el clasificador quien decide cuál es el idioma reconocido en función al modelo que presente la menor perplejidad.

La diferencia entre el sistema PRLM y el utilizado en este trabajo para el reconocimiento radica principalmente en el Back-End, dado que no se utilizan las "cuentas" de los n-gramas para generar modelos de lenguaje, sino que las utilizamos para generar dos modelos de idioma distintos que luego se combinan. Por una parte, el primer sistema se crea a partir de los i-vectores generados para los ficheros de entrenamiento, mientras que el segundo utiliza las cuentas para crear un ranking de n-gramas discriminativos junto con su valor de discriminabilidad al comparar las ocurrencias de los n-gramas en un idioma frente a los otros. Posteriormente, los dos sistemas se unifican modificando las cuentas usadas por el primer sistema mediante el valor de discriminabilidad calculado en el segundo sistema. Finalmente, como clasificador hemos utilizado un sistema basado en regresión logística que utiliza como entrada los i-vectores reconocidos y les asigna una puntuación (score) según la similitud de cada i-vector con los diferentes modelos de idioma entrenados.

2.1 El sistema fonotáctico basado en cuentas de posteriorgramas

La creación de las cuentas de posteriorgramas se describe en los siguientes pasos (Figura 2):

a. El primer paso consiste en extraer los valores de las probabilidades a posteriori de cada uno de los posibles fonemas a reconocer para cada trama. Estos valores se obtienen directamente del reconocedor de fonemas. En la figura podemos ver que para cada trama del fichero de audio se obtienen 3 valores correspondientes a los 3 posibles fonemas a reconocer (i.e. hemos usado 3 fonemas para simplificar el ejemplo, aunque en el sistema real fueron 33 fonemas distintos).
b. En un segundo paso se suman y se promedian las probabilidades a posteriori de todas las tramas que se consideran que pertenecen a la misma unidad fonética. Esta agrupación de las tramas en fonemas es realizada también por el reconocedor de fonemas empleando el algoritmo de Viterbi sobre las probabilidades a posteriori obtenidas en el paso anterior.
c. El tercer paso es calcular las probabilidades condicionales de que ocurra un determinado fonema considerando los n-1 fonemas previos (i.e. n-gramas). Para ello, en el caso de usar bigramas, como se muestra en la figura, se realiza el producto exterior (outer-product) entre las probabilidades promediadas del fonema previo con las del fonema actual. Para solventar el problema de la primera trama se crea un fonema tipo "dummy" en el que la todos los fonemas son equiprobables.
d. El cuarto paso consiste en sumar todas las matrices producto generadas antes a lo largo de todo el fichero cuidando de sumar adecuadamente los mismos contextos (i.e. la probabilidad condicional $p_{ij}(t-1)$ con la probabilidad $p_{ij}(t)$). El resultado es lo que denominados cuentas de posteriorgrama condicionales.

Por último se convierte la matriz de cuentas de posteriorgramas en un supervector de dimensión [F^n x 1], donde F es el número de fonemas a emplear y n es el orden de los n-gramas. En nuestro caso, al emplear 33 fonemas y usar bigramas obtenemos un vector de dimensión 1089, y en el caso de trigramas tenemos un vector de dimensión 35937. Estos supervectores se crean para cada fichero y para cada uno de los idiomas a reconocer. Luego se utilizan en el entrenamiento de los i-vectores.

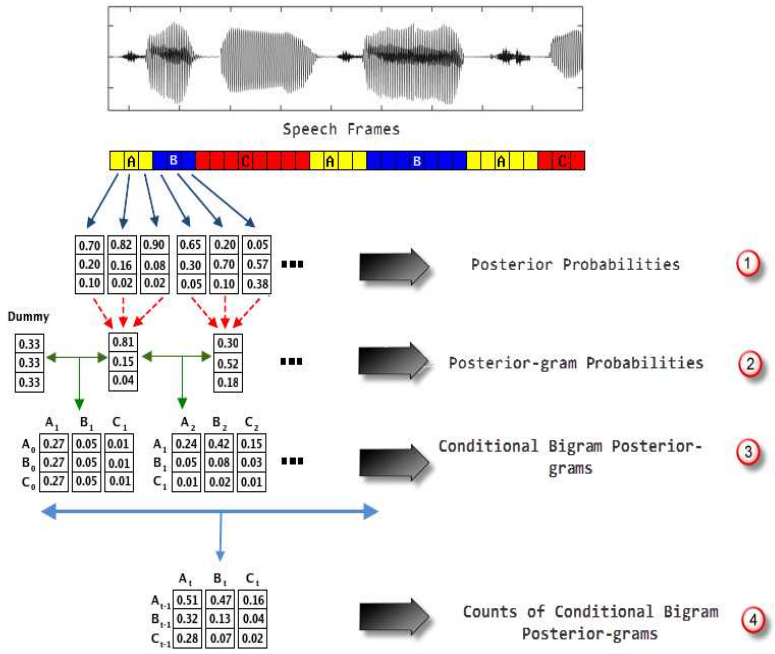

Figura 2. Procedimiento de creación de las cuentas de posteriogramas.

Una vez obtenidas las cuentas de posteriorgramas para todos los ficheros, se procede a modelar las probabilidades globales de ocurrencia de cada una de los n-gramas para posteriormente calcular los i-vectores usando un modelo de subespacios mutinomiales (SMM, Subspace Multinomial Model) propuesto por (Povey, 2010), que permite entrenar vectores de baja dimensión (i-vectores) en el subespacio de la variabilidad total para luego usarlos como vectores de características en el entrenamiento de un clasificador discriminativo de reconocimiento de lenguaje.

El entrenamiento de los i-vectores se realiza mediante el método de estimación y maximización (EM) y la optimización se lleva a cabo aplicando el método de Newton-Raphson. Para mayores detalles acerca de las formulaciones matemáticas de los SMM y su aplicación en sistemas de reconocimiento de idioma se recomienda la lectura de (Kockmann et al, 2010) y (Soufifar et al, 2011). Por otra parte, la formulación matemática de los i-vectores (Dehak et al, 2011) se realiza empleando la siguiente relación matemática:

$$m_x = M + T\omega_x \qquad (1)$$

Donde m_x es un vector de dimensión [F^n x 1] que contiene las medias de las características que se modelan para un fichero determinado *x*. En nuestro caso, a partir de las cuentas de los supervectores de los posteriorgramas hallamos las probabilidades medias de ocurrencia de cada fonema para cada fichero. Por otra parte, la M es un vector de dimensión [F^n x 1], que contiene las medias globales independientes de idioma mejor conocido como UBM (i.e. Universal Background Model). La T se conoce como extractor de i-vectores y tiene una dimensión de [F^n x r] donde r es un valor de baja dimensionalidad que se selecciona de tal forma que r << F^n. Por último, la ω_x son los i-vectores en sí y tienen una dimensión de [r x 1].

El procedimiento para entrenar la matriz T y crear los i-vectores es un proceso iterativo en el que se parte de unos valores de i-vectores inicializados aleatoriamente para con ellos obtener la matriz T; luego, con esta T se regeneran los i-vectores y a partir de los nuevos se vuelve a crear una nueva matriz T, y así sucesivamente. El proceso se detiene cuando entre iteración e iteración no se obtiene una reducción en la verosimilitud global del modelo. Tal como se ha comentado previamente, la gran ventaja de los i-vectores es la posibilidad de trabajar con vectores de baja dimensión ya que esto reduce los problemas de dispersión de datos y facilita el entrenamiento del clasificador. En nuestro caso hemos trabajado con i-vectores de dimensión 400 ya que con ellos se obtuvieron los mejores resultados durante la competición oficial.

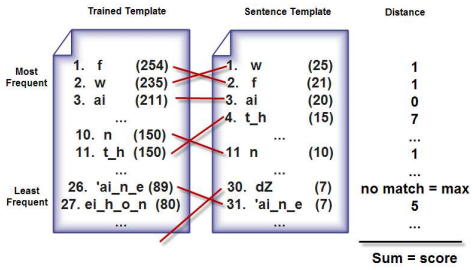

Figura 3. Categorización de texto en base de la ocurrencia de n-gramas

2.2 El sistema fonotáctico basado en rankings discriminativos de n-gramas

Este sistema se basa en el uso de una técnica de categorización de textos propuesta por (Cavnar y Trenkle, 1994) que permite combinar información local (i.e. n-gramas) e información de largo alcance (i.e. las cuentas de n-gramas recogidas a lo largo de una frase). En la Figura 3 se muestra a modo de ejemplo la técnica original en la que se propone durante la fase de entrenamiento la creación de una plantilla con los n-gramas más frecuentes (típicamente los primeros 400) empleando hasta un orden de 5-gramas, obtenidos a partir de las secuencias de caracteres (i.e. fonemas reconocidos en nuestro caso) de los ficheros de entrenamiento para cada idioma y ordenados por su ocurrencia de mayor a menor. Durante la fase de evaluación, se crea una plantilla dinámica a partir de la frase reconocida y se ordena siguiendo el mismo procedimiento que en la fase de entrenamiento. Para realizar la detección del idioma se suma la diferencia absoluta entre las posiciones de los n-gramas de las dos plantillas utilizando la ecuación 2.

$$d^T = \frac{1}{L}\sum_{i=1}^{L} abs(pos\ w_i - pos\ w_i^T) \qquad (2)$$

Donde L es el número de n-gramas generados para la frase a reconocer. Para aquellos n-gramas que no aparecen en los rankings entrenados se aplica una penalización en función del tamaño de la plantilla. Finalmente, el idioma reconocido es aquel que presente la mínima distancia entre las dos plantillas.

En (Cordoba et al, 2007) presentamos diferentes mejoras a la técnica original. Las más importantes fueron:

- Aplicamos lo que denominamos "posición golf" en la que para aquellos n-gramas cuyo número de ocurrencias sean iguales se ubican en el ranking dentro de la misma posición; tal como en el golf, donde aquellos jugadores que tengan el mismo número de golpes ocupan la misma posición en el ranking.
- La creación de diferentes rankings para los diferentes órdenes de n-gramas con lo que no se penalizaban los n-gramas de ordenes mayores en preferencia a aquellos más bajos (e.g. unigramas o bigramas) que suelen aparecer mucho más.
- Finalmente, inspirados en el trabajo presentado en (Lamel et al, 2002), donde se obtuvieron mejores resultados de identificación usando unidades más discriminativas, decidimos incorporar el mismo concepto aquí. En este caso, ubicando en las posiciones más altas del ranking aquellos n-gramas que aparezcan más en un idioma que en los demás y, por tanto, son más discriminativos.

Con el objetivo de calcular el valor de discriminación de cada n-grama probamos diferentes fórmulas basadas en la conocida métrica tf-idf. Para describir la fórmula que usamos, partimos de la ecuación 3 en la que $n_1(w)$ es el número de veces que aparece un n-grama en un idioma concreto, y $n_2(w)$ las veces que ocurre ese mismo n-grama en los otros idiomas, y T son las plantillas creadas para cada idioma.

$$N_1 = \sum_{\forall w; w \in T_1} n_1(w) \qquad N_2 = \frac{1}{|T-1|}\sum_{\forall w; w \in T; T \neq T_1} n_2(w) \qquad (3)$$

Como el número de cuentas será diferente para cada idioma y orden de los n-gramas, antes de hacer las comparativas aplicamos un proceso de normalización utilizando la ecuación 4. En esta fórmula: N_1 es la suma de todas las cuentas de todos los n-gramas para el idioma actual y N_2 es el promedio para el resto de idiomas.

$$n_1'(w) = \frac{n_1(w) \times N_2}{N_1 + N_2} \qquad n_2'(w) = \frac{n_2(w) \times N_1}{N_1 + N_2} \qquad (4)$$

Finalmente, también se puede aplicar un umbral sobre estos valores normalizados con el fin de eliminar n-gramas no representativos que aparecen muy poco. En nuestro caso no usamos estos umbrales. La ecuación 5 muestra las

fórmulas empleadas tanto en el caso de que el n-grama aparezca más en un idioma que en los otros ($n_1' > n_2'$).

$$\mathrm{Disc}(n_1) = \begin{cases} \alpha * \left(\frac{n_1' * (n_1' - n_2')}{(n_1' + n_2')^2} + \delta \right), n_1' > n_2' \\ \alpha * \left(\frac{n_2' * (n_1' - n_2')}{(n_1' + n_2')^2} + \delta \right), else \end{cases} \quad (5)$$

Donde, $\propto$ y δ son valores que normalizan el ranking discriminativo entre 0 y 1. Donde el valor "0" significa que el n-grama no es nada discriminativo o nada relevante para el idioma, y el valor "1" que es muy discriminativo o relevante para el idioma a reconocer.

2.3 Descripción de la técnica propuesta

Tal como hemos comentado en la introducción, la técnica propuesta se basa en la modificación de las cuentas de los posteriorgramas utilizadas para la generación de los i-vectores y le incorpora la información procedente del sistema que genera el ranking discriminativo de las cuentas de n-gramas. Para ello, lo que proponemos es modificar las cuentas de posteriorgramas incrementando su valor en función de cuán discriminativo sea el n-grama; es decir, que aparezca más para un idioma que para los demás. Para ello, utilizamos la ecuación 6:

$$C_{d,n}^{i} = \left(1 + \omega_n^i\right) \times Co_{d,n}^{i} \quad (6)$$

donde $C_{d,n}^{i}$ es el nuevo valor de la cuenta para el n-grama *n* obtenido para el fichero *d* e idioma *i*; $Co_{d,n}^{i}$ es el valor de la cuenta original en el supervector y ω_n^i es el valor de discriminabilidad del n-grama calculado al crear el ranking discriminativo para el idioma i.

Como puede observarse, el resultado es que la cuenta permanece inalterada si el n-grama no es discriminativo y puede llegar a duplicarse en caso de máxima discriminatividad.

El resultado es almacenado como un nuevo supervector denominado (SVs-RkDis) que ahora contendrá el efecto discriminativo de los rankings y que permitirá la generación de nuevos i-vectores en los que las dimensiones relacionadas con los n-gramas discriminativos adquieran mayor relevancia.

3 Condiciones de la evaluación y los experimentos

Para la realización de las pruebas de la técnica propuesta hemos partido del mismo conjunto de datos de entrenamiento, evaluación y desarrollo que se usaron durante la evaluación ALBAYZIN 2012 LRE. Para esta evaluación se distribuyeron ficheros de audio extraídos de un portal de vídeos en la web, con diferentes longitudes, condiciones de canal y número de locutores, así como la presencia de diversas señales sonoras como música y ruido (Rodriguez-Fuentes et al, 2012).

Por otra parte, para la evaluación se propusieron cuatro tipos de condiciones distintas: a) plenty-closed, b) plenty-open, c) empty-closed, y d) empty-open. Donde los términos plenty y empty hacen referencia a que se tiene, o no, un conjunto de datos de entrenamiento, así como una diferenciación en los idiomas a reconocer. En el caso de la condición plenty se debían reconocer los siguientes 6 idiomas: español, inglés, portugués, gallego, vasco y catalán. Para la condición empty los idiomas eran: francés, italiano, alemán y griego. En cuanto a los términos closed y open, se refieren a la posibilidad de reconocer únicamente los idiomas mencionados antes (i.e. closed) o a la posibilidad de que el sistema pudiera detectar que el idioma del fichero es diferente a los incluidos en la condición (i.e. open).

Ficheros	Limpios	Ruidosos	Total
Español	486	312	798
Inglés	322	365	587
Gallego	675	300	975
Catalán	440	209	649
Vasco	579	215	794
Portugués	558	295	853
Total	3060	1596	4656

	Train	Dev	Test	Eval
Ficheros totales	4656	458	457	941

Tabla 1. Estadísticas por idiomas de los ficheros de entrenamiento y de la distribución de todos los datos de la evaluación para la condición plenty-closed

Los experimentos presentados en este artículo se han realizado únicamente para la condición principal de la evaluación, i.e. plenty-

closed. En la Tabla 1, se muestra la distribución y el número de ficheros de entrenamiento disponibles para la condición plenty-closed, así como la distribución que hicimos de todos los datos para realizar ajustes y probar el sistema antes (test) y durante la evaluación oficial (eval).

Finalmente, conviene mencionar que durante la evaluación oficial se propuso la utilización de la métrica F_{act} con el objetivo de medir y comparar la bondad de los sistemas propuestos. Esta métrica se puede entender como una medida del grado de "incertidumbre" que tiene el sistema para detectar los idiomas. De esta manera, un valor de 0.0 significa que el sistema no tiene ninguna duda para reconocer los idiomas, en tanto que un valor igual o superior a 1.0 que no es capaz de mejorar la tasa de un sistema que escogiera de forma equiprobable cualquier idioma. Para mayores detalles se recomienda consultar (Rodriguez-Fuentes et al, 2012) y (Rodriguez-Fuentes et al, 2013).

Finalmente, también hemos decidido incluir en los resultados el cálculo de la medida C_{avg} en porcentaje, ya que esta métrica ha sido ampliamente utilizada en todas las competiciones internacionales de reconocimiento de idioma. Esta medida tiene como objetivo ponderar los errores de falsa aceptación y falso rechazo del sistema por lo que un valor cercano a 0 significa que el sistema no comete ninguno de estos dos tipos de errores.

4 *Resultados*

La Tabla 2 muestra los resultados obtenidos al usar el sistema fonotáctico basado en el uso de i-vectores de tamaño 400 sobre los supervectores de cuentas de trigramas originales (SV) tanto para el conjunto de datos de test, como los proporcionados durante la evaluación. En la segunda línea vemos los resultados tras modificar las cuentas de los supervectores empleando las plantillas discriminativas (SVs+RkDis).

Tal como se puede ver en la Tabla 2, la modificación de las cuentas originales mediante la información discriminativa permite mejorar los resultados de C_{avg} en un 8.63% relativo y un 0.4% en la tasa de F_{act} para los datos de evaluación. En cuanto a los ficheros de test vemos que la mejora en Fact es un poco mayor (0.6%) en tanto que para Cavg empeora sólo un poco (0.43%).

FICHEROS DE TEST			
No.		**Fact**	**Cavg(%)**
1	SVs	0.133658	6.94
2	SVs+RkDis	0.132864	6.97

FICHEROS DE EVALUACIÓN			
No.		**Fact**	**Cavg(%)**
1	SVs	0.181393	9.85
2	SVs+RkDis	0.180704	9.00

Tabla 2. Resultados de los errores de reconocimiento para los ficheros de test y evaluación con trigramas.

5 *Conclusiones y lineas futuras*

En este artículo hemos presentado una técnica novedosa que permite combinar dos tipos de sistemas fonotácticos distintos empleando para ello información de largo alcance e información discriminativa como son las que proveen las plantillas y un sistema basado en i-vectores que es la técnica más exitosa para reconocimiento de idioma actualmente. Los resultados sobre los datos de la evaluación muestran que la técnica propuesta permite mejorar las tasas de reconocimiento hasta un 8.63% relativo, validando así sus prestaciones.

En relación con los trabajos futuros proponemos la inclusión de umbrales de decisión aplicados a los valores discriminativos de las plantillas, de forma que únicamente los n-gramas con un número mínimo de repeticiones vean modificadas las cuentas de sus posteriorgramas. En esta misma línea, consideraremos la creación de nuevas plantillas en las que el valor discriminativo pueda ser calculado a partir de nuevas fórmulas pudiendo también utilizar umbrales. Finalmente, también trabajaremos en ampliar esta técnica utilizando un sistema tipo PPRLM en el que tengamos no sólo un reconocedor de fonemas si no que podamos usar varios reconocedores en paralelo (e.g. aprovechando también los modelos de Checo o Ruso que viene incluido con el reconocedor de la Universidad de Brno).

6 *Agradecimientos*

Este trabajo ha sido posible gracias a la financiación de los siguientes proyectos: MA2VICMR (CC.AA. de Madrid, S2009/TIC-1542), y TIMPANO (TIN2011-28169-C05-03).

7 Referencias bibliográficas

Kinnunen, T., H. Li, 2010. “An overview of text-independent speaker recognition: From features to supervectors”. Speech Communication, Vol 52, Issue 1, pp. 12-40.

Dehak, N., P. Kenny, R. Dehak, P. Dumouchel y P. Ouellet, 2011. “Front-End Factor Analysis for Speaker Verification”. IEEE Transactions on Audio, Speech, and Language Processing, 19(4), pp.788-798.

Martínez, D., O. Plchot, L Burget, O. Glembek y P. Matejka. 2011. “Language Recognition in iVectors Space". Proceedings of Interspeech. pp. 861-864.

D'Haro, L.F., O. Glembek, O. Plchot, P. Matejka, M. Soufifar, R. Cordoba y J. Cernocky. 2012. “Phonotactic language recognition using i-Vectors and phoneme posteriorgram counts”. Proceedings of Interspeech pp. 9-13.

D'Haro, L.F., R. Cordoba, 2013. “Low-Resource language recognition using a fusion of phoneme posteriorgrams counts, acoustic and glottal-based i-Vectors”. ICASSP 2013.

Cordoba, R., L.F. D'Haro, F. Fernandez-Martinez, J. Macias-Guarasa, y J. Ferreiros. 2007. “Language Identification based on n-gram Frequency Ranking”. 8th Annual Conference of the International Speech Communication Association, Interspeech, Vol. 3, pp.1921-1924.

Zissman, M. 1996. “Comparison of four approaches to automatic language identification of telephone speech”. IEEE Transactions on Speech and Audio Processing, vol.4, no.1, pp. 31-44.

Schwarz, P., 2009. “Phoneme Recognition based on Long Temporal Context”, PhD Tesis. Brno University of Technology. Disponible: http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context.

Diez, M., A. Varona, M. Penagarikano, L. Rodriguez-Fuentes, G. Bordel. "Dimensionality Reduction of Phone Log-Likelihood Ratio Features for Spoken Language Recognition". Interspeech 2013; Lyon, France, 25-29 aug., 2013

Povey, D., L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, S. Thomas. 2010. “Subspace Gaussian Mixture Models for Speech Recognition”, Proceedings of ICASSP, Dallas. pp 4330-4333.

Kockmann, M., L. Burget, O. Glembek, L. Ferrer, J. Cernocky, 2010. “Prosodic speaker verification using subspace multinomial models with intersession compensation,” Proceedings of ICSPL, Makuhari, Chiba, Japan.

Soufifar, M., M. Kockmann, L. Burget, O. Plchot, O. Glembek, T. Svendsen. 2011. “IVector approach to phonotactic language recognition”. Proceedings of Interspeech 2011, pp 2913-2916.

Cavnar, W., J. Trenkle. 1994.” N-Gram-Based Text Categorization”. Environmental Research Institute of Michigan.

Lamel, L., J-L. Gauvain, G. Adda, G. 2002. "Lightly Supervised and Unsupervised Acoustic Model Training". Computer Speech and Language, Vol.16, no.1, pp. 115-129.

Rodríguez-Fuentes, L., N. Brummer, M. Penagarikano, A. Varona, M. Diez, G. Bordel. 2012. “The Albayzin 2012 Language Recognition Evaluation Plan (Albayzin 2012 LRE)”.

Rodriguez-Fuentes, L. J., Brümmer, N., Penagarikano, M., Varona, A., Bordel, G. , Diez, M. "The Albayzin 2012 Language Recognition Evaluation". Interspeech 2013; Lyon, France, 25-29 aug., 2013.

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 153-160 recibido 30-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Language Recognition on Albayzin 2010 LRE using PLLR features

Reconocimiento de la Lengua en Albayzin 2010 LRE utilizando características PLLR

M. Diez, A. Varona, M. Penagarikano, L.J. Rodriguez-Fuentes, G. Bordel
University of the Basque Country, UPV/EHU
GTTS, Department of Electricity and Electronics
amparo.varona@ehu.es

Resumen: Los así denominados Phone Log-Likelihood Ratios (PLLR), han sido introducidos como características alternativas a los MFCC-SDC para sistemas de Reconocimiento de la Lengua (RL) mediante iVectors. En este artículo, tras una breve descripción de estas características, se proporcionan nuevas evidencias de su utilidad para tareas de RL, con un nuevo conjunto de experimentos sobre la base de datos Albayzin 2010 LRE, que contiene habla multi-locutor de banda ancha en seis lenguas diferentes: euskera, catalán, gallego, español, portugués e inglés. Los sistemas de iVectors entrenados con PLLRs obtienen mejoras relativas significativas respecto a los sistemas fonotácticos y sistemas de iVectors entrenados con características MFCC-SDC, tanto en condiciones de habla limpia como con habla ruidosa. Las fusiones de los sistemas PLLR con los sistemas fonotácticos y/o sistemas basados en MFCC-SDC proporcionan mejoras adicionales en el rendimiento, lo que revela que las características PLLR aportan información complementaria en ambos casos.
Palabras clave: Reconocimiento de la Lengua, Phone Log-Likelihood Ratios, iVectors

Abstract: Phone Log-Likelihood Ratios (PLLR) have been recently proposed as alternative features to MFCC-SDC for iVector Spoken Language Recognition (SLR). In this paper, PLLR features are first described, and then further evidence of their usefulness for SLR tasks is provided, with a new set of experiments on the Albayzin 2010 LRE dataset, which features wide-band multi speaker TV broadcast speech on six languages: Basque, Catalan, Galician, Spanish, Portuguese and English. iVector systems built using PLLR features, computed by means of three open-source phone decoders, achieved significant relative improvements with regard to the phonotactic and MFCC-SDC iVector systems in both clean and noisy speech conditions. Fusions of PLLR systems with the phonotactic and/or the MFCC-SDC iVector systems led to improved performance, revealing that PLLR features provide complementary information in both cases.
Keywords: Spoken Language Recognition, Phone Log-Likelihood Ratios, iVectors

1. Introduction

In the last years, two complementary types of Spoken Language Recognition (SLR) systems prevail: (1) those using *low-level* (typically, short-term spectral) features; and (2) those using *high-level* (typically, phonotactic) features. Among the first type of systems, the so called Total Variability Factor Analysis approach (also known as *iVector* approach) has been recently introduced, using Mel-Frequency Cepstral Coefficients and Shifted Delta Cepstra (MFCC-SDC) features (Dehak et al., 2011b). The iVector approach maps high-dimensional input data, typically a Gaussian Mixture Model (GMM) supervector, to a low-dimensional feature vector (an iVector), hypothetically retaining most of the relevant information.

Due to its high performance and low complexity, the iVector approach has become a state-of-the-art technique. Besides MFCC-SDC, other alternative features have been already tested under this approach, such as prosodic features (pitch, energy and dura-

ISSN 1135-5948

tion) (Martínez et al., 2012) or speaker vectors from subspace GMM (Plchot et al., 2012). It has been reported that these systems alone do not yield outstanding results, but performance improves significantly when fusing them with a system based on spectral features.

Among *high-level* approaches, best results are reported for the so called *Phone-Lattice-SVM* approach (Campbell, Richardson, and Reynolds, 2007), which uses expected counts of phone n-grams (computed on phone lattices provided by phone decoders) as features to feed a Support Vector Machine (SVM) classifier.

There have been some efforts to use phonotactic features under the iVector approach. In (Soufifar et al., 2012), expected counts of phone n-grams are used as features, reaching the same performance as state-of-the-art phonotactic systems. In (DHaro et al., 2012), phone posteriorgrams (instead of phone lattices) are used to estimate n-gram counts, and the iVector approach is then applied to reduce the high-dimensionality of the resulting feature vectors. Both approaches yield reasonable good results, and the latter is reported to fuse well with a SLR system based on short-term spectral features.

Best results are usually obtained by fusing several acoustic and phonotactic systems. Increasingly sophisticated fusion and calibration techniques have been applied, including generative Gaussian backends (Singer et al., 2003; BenZeghiba, Gauvain, and Lamel, September 2009) and discriminative logistic regression (Brümmer and van Leeuwen, 2006; Brümmer and de Villiers, 2011; Penagarikano et al., 2012).

The development of SLR technology has been largely supported by NIST Language Recognition Evaluations (LRE) (NIST LRE, 2011), held in 1996 and every two years since 2003. As a result, the datasets produced and distributed for such evaluations have become standard benchmarks to test the usefulness of new approaches. NIST LRE datasets consist mostly of narrow-band (8 kHz) conversational telephone speech.

Aiming to fill the gap of SLR technology assessment for wide-band broadcast speech, the Albayzin LREs have been organized (Rodriguez-Fuentes et al., 2010; Rodriguez-Fuentes et al., 2011), with the support of the Spanish Thematic Network on Speech Technologies (RTTH, 2006) and the ISCA Special Interest Group on Iberian Languages (SIG-IL). For the Albayzin 2008 LRE, the four official languages spoken in Spain: Basque, Catalan, Galician and Spanish, were used as target languages. In (Varona et al., 2010) an in depth study was carried out, the main verification system being obtained from the fusion of an acoustic system and 6 phonotactic subsystems.

The set of Iberian languages was completed in the Albayzin 2010 LRE by adding Portuguese as target language. Due to its international relevance and its pervasiveness in broadcast news, English was also added as target language in the Albayzin 2010 LRE. A new condition was introduced, depending on the presence of background noise, music and/or conversations (overlapped speech), leading to two additional tracks which involved clean speech and a mix of clean and noisy speech, respectively.

In a previous work (Diez et al., 2012), we proposed and evaluated the use of log-likelihood ratios of phone posterior probabilities, hereafter called Phone Log-Likelihood Ratios (PLLR), as alternative features to MFCC-SDC under the iVector approach. We found very promising results in language recognition experiments on the NIST 2007 and 2009 LRE datasets.

In this paper, a more detailed study of the PLLR features is undertaken. A new set of experiments has been carried out on the Albayzin 2010 LRE dataset (Rodriguez-Fuentes et al., 2012) to prove their effectiveness. Three iVector systems have been built using three open-source phone decoders to compute the PLLR features. These systems are compared to (and fused with) various state-of-the-art baseline systems, namely: (1) an acoustic iVector system using MFCC-SDC as features; and (2) three Phone-Lattice-SVM systems built on the same decoders used to compute the PLLR features.

The rest of the paper is organized as follows. Section 2 provides some background and describes the computation of the phone log-likelihood ratios used as features in this work. Section 3 describes the experimental setup. Section 4 presents results and compares the performance of the proposed approach to that of state-of-the-art approaches. Finally, conclusions are given in Section 5.

2. Phone Log-Likelihood Ratio (PLLR) features

In (Biadsy, Hirschberg, and Ellis, 2011), a new dialect recognition approach mixing acoustic and phonetic information was presented, based on the assumption that certain phones are realized in different ways across dialects. Acoustic models were trained for different phonetic categories, based on the phonetic segmentation provided by a phone decoder. Scores were computed based on differences between acoustic models corresponding to the same phonetic category in different dialects.

That work encouraged us to search for similar but more sophisticated approaches. After exploring the possibility of using phone posteriors at the frame level to smooth the phonetic segmentation, we came to the idea of using phone posteriors alone as features. The non-Gaussian distribution of phone posteriors was addressed by transforming phone posteriors into phone log-likelihood ratios, which carry the same information but show approximately Gaussian distributions, as illustrated in Figure 1. Under this configuration, phone models perform as a sort of reference system and phone log-likelihood ratios at a given frame can be interpreted as the *location* of the speech segment being analyzed in the space defined by those models.

To compute the PLLRs, let us consider a phone decoder including N phone units, each of them represented typically by means of a model of S states. Given an input sequence of acoustic observations X, we assume that the acoustic posterior probability of each state s ($1 \leq s \leq S$) of each phone model i ($1 \leq i \leq N$) at each frame t, $p_{i,s}(t)$, is output as side information by the phone decoder. Then, the acoustic posterior probability of a phone unit i at each frame t can be computed by adding the posteriors of its states:

$$p_i(t) = \sum_{\forall s} p_{i,s}(t) \qquad (1)$$

Assuming a binary classification task with flat priors, the log-likelihood ratios at each frame t can be computed from posterior probabilities as follows:

$$LLR_i(t) = \log \frac{p_i(t)}{\frac{1}{(N-1)} \sum_{\forall j \neq i} p_j(t)} \qquad i = 1, ..., N \qquad (2)$$

The resulting N log-likelihood ratios per frame are the PLLR features considered in our approach.

3. Experimental setup

3.1. PLLR iVector system

As a first step to get the PLLR features, we applied the open-software Temporal Patterns Neural Network (TRAPs/NN) phone decoders, developed by the Brno University of Technology (BUT) for Czech (CZ), Hungarian (HU) and Russian (RU) (Schwarz, 2008), which include 42, 58 and 49 phonetic units, respectively, plus 3 non-phonetic units. Note that BUT decoders represent each phonetic unit by a three-state model and output the transformed posterior probabilities $p_{i,s}(t)$ (Diez et al., 2012) as side information, for each state s of each phone model i at each frame t.

Before computing PLLR features, the three non-phonetic units —*int* (intermittent noise), *pau* (short pause) and *spk* (non-speech speaker noise)— were integrated into a single 9-state non-phonetic unit model. Then, a single posterior probability was computed for each phone i ($1 \leq i \leq N$), according to Equation 1. Finally, the log-likelihood ratio for each phone i was computed according to Equation 2. In this way, we get 43, 59 and 50 PLLR features per frame using the BUT decoders for Czech, Hungarian and Russian, respectively.

As shown in (Diez et al., 2012), adding first order dynamic coefficients improved significantly the performance of the PLLR-based iVector system. Therefore, PLLR+Δ were used as features also in this work. Voice activity detection was performed by removing the feature vectors whose highest PLLR value corresponded to the integrated non-phonetic unit. A gender independent 1024-mixture GMM (Universal Background Model, UBM) was estimated by Maximum Likelihood using the NIST 2011 LRE training set. The total variability matrix (on which the iVector approach relies) was estimated as in (Dehak et al., 2011a), using only target languages in the NIST 2011 LRE training set. A generative modeling approach was applied in the iVector feature space (as in (Martínez et al., 2011)), the set of iVectors of each language being modeled by a single Gaussian distribution. Thus, the iVector scores were compu-

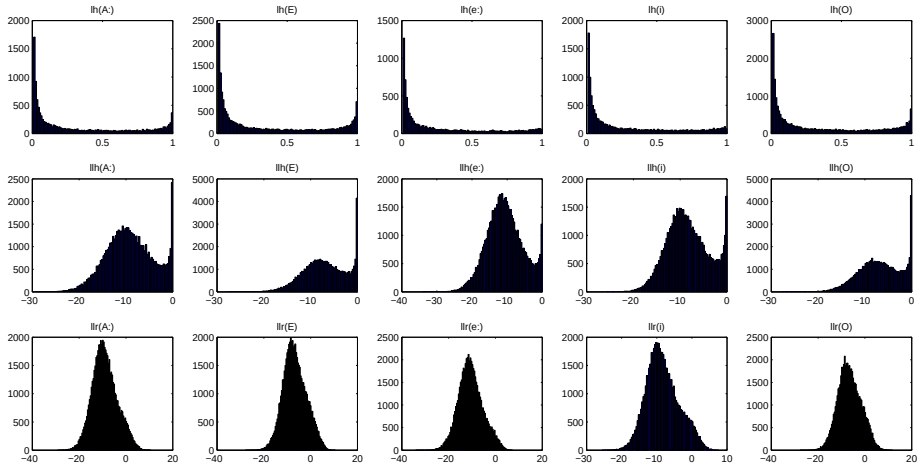

Figure 1: Distributions of frame-level likelihoods (*lh*, first row), log-likelihoods (*llh*, second row) and log-likelihood ratios (*llr*, third row) for five Hungarian phones (A:, E, e:, i and O).

ted as follows:

$$score(f, l) = N(w_f; \mu_l, \Sigma) \qquad (3)$$

where w_f is the iVector for target signal f, μ_l is the mean iVector for language l and Σ is a common (shared by all languages) within-class covariance matrix.

3.2. MFCC-SDC iVector system

In this case, the concatenation of MFCC and SDC coefficients under a 7-2-3-7 configuration was used as acoustic representation. Voice activity detection, GMM estimation and total variability matrix training and scoring were performed as in the PLLR iVector approach.

3.3. Phonotactic systems

The three phonotactic systems used in this work have been developed under the phone-lattice-SVM approach (Campbell, Richardson, and Reynolds, 2007) (Penagarikano et al., 2011). Given an input signal, an energy-based voice activity detector was applied in first place, which split and removed long-duration non-speech segments. Then, the BUT TRAPs/NN phone decoders for Czech, Hungarian and Russian (Schwarz, 2008) were applied. Regarding channel compensation, noise reduction, etc. the three systems relied on the acoustic front-end provided by BUT decoders.

Phone posteriors output by BUT decoders were converted to phone lattices by means of HTK (Young et al., 2006) along with the BUT recipe (Schwarz, 2008). Then, expected counts of phone n-grams were computed using the *lattice-tool* of SRILM (Stolcke, 2002). Finally, a SVM classifier was applied, SVM vectors consisting of expected frequencies of phone n-grams (up to $n = 3$). A sparse representation was used, which involved only the most frequent features according to a greedy feature selection algorithm (Penagarikano et al., 2011). L2-regularized L1-loss support vector regression was applied, by means of LIBLINEAR (Fan et al., 2008).

3.4. Dataset

The Albayzin 2010 LRE dataset (KALAKA-2) contains wide-band 16 kHz TV broadcast speech signals for six target languages. The Albayzin 2010 LRE (Rodriguez-Fuentes et al., 2011) featured two main evaluation tasks, on clean and noisy speech, respectively. In this work, acoustic processing involved down-sampling signals to 8 kHz, since all the systems were designed to deal with narrow-band signals.

The training, development and evaluation datasets used for this benchmark match exactly those defined for the Albayzin 2010 LRE. For the primary clean-speech language recognition task, more than 10 hours of clean speech per target language were used for training. For the noisy-speech language recognition task, besides the clean speech subset, more than 2 hours of noisy/overlapped speech segments were used for each target language. The distribution of training data, which amounts to around 82 hours, is shown

	Clean Speech			Noisy Speech		
	Hours	# 30s segments		Hours	# 30s segments	
Language	Train	Devel	Eval	Train	Devel	Eval
Basque	10.73	146	130	2.25	29	74
Catalan	11.45	120	149	2.18	47	55
English	12.18	133	135	2.53	60	69
Galician	10.74	137	121	2.23	60	83
Portuguese	11.08	164	146	3.28	77	58
Spanish	10.41	136	125	3.70	83	79
TOTAL	66.59	836	806	16.17	356	418

Table 1: Albayzin 2010 LRE: Distribution of training data (hours) and development and evaluation data (# 30s segments).

in Table 1. Only 30-second segments were used for development purposes. The development dataset used in this work consists of 1192 segments, amounting to more than 10 hours of speech. Results reported in this paper were computed on the Albayzin 2010 LRE evaluation corpus, specifically on the 30-second, closed set condition (for both clean speech and noisy speech conditions). The distribution of segments in the development and evaluation datasets is shown in Table 1. For further details, see (Rodriguez-Fuentes et al., 2012).

3.5. Fusion

The *FoCal* multiclass toolkit was applied to perform the calibration and fusion of SLR systems (Brümmer and du Preez, 2006).

3.6. Evaluation measures

In this work, systems are compared in terms of: (1) the average cost performance C_{avg} as defined in NIST evaluations up to 2009; and (2) the Log-Likelihood Ratio Cost (C_{LLR}) (Brümmer and du Preez, 2006).

4. Results

Table 2 shows the performance of the baseline systems (the acoustic MFCC-SDC and phonotactic systems) and the proposed approach (using the BUT Czech (CZ), Russian (RU) and Hungarian (HU) decoders) on the Albayzin 2010 LRE closed-set clean-speech and noisy-speech 30-second task. Regarding clean-speech, most of the systems performed similarly, except for the proposed PLLR iVector system when trained on the HU decoder PLLR features, which clearly stands out as the best single system, yielding 1.41 $C_{avg} \times 100$, which means a 33 % relative improvement with regard to the MFCC-SDC-based iVector approach and a 40 % relative improvement with regard to the respective HU phonotactic approach. Performance differences across decoders were found on both PLLR and phonotactic approaches (e.g. the performance of the phonotactic RU system degraded with regard to that of other phonotactic systems).

When focusing on the noisy speech condition, differences in performance were more noticeable. MFCC-SDC-based iVector system attained great performance (3.95 $C_{avg} \times 100$), but was once again outperformed by the HU PLLR iVector system (3.17 $C_{avg} \times 100$). All PLLR iVector systems outperformed their respective phonotactic counterparts (yielding between 5 % and 56 % relative improvements).

Since the HU PLLR iVector system showed the best performance among individual systems, we selected the HU decoder-based systems to analyze system fusions. Table 3 shows the performance of different fusions involving the baseline MFCC-SDC iVector system, the HU phonotactic system and the HU PLLR iVector system (for a better comparison, single system results are also included in Table 3). All pairwise fusions yielded high performance. The fusion of the MFCC-SDC iVector and phonotactic systems, led to great improvements with regard to single system performance (1.10 $C_{avg} \times 100$). A similar figure was achieved by the fusion of the PLLR iVector and phonotactic system (1.09 $C_{avg} \times 100$), closely followed by the fusion of the MFCC-SDC and PLLR iVector systems (1.20 $C_{avg} \times 100$). The fusion of the three systems yielded great performance: 0.97 $C_{avg} \times 100$, meaning a 31 % relative impro-

System		Clean		Noisy	
		$C_{avg} \times 100$	C_{LLR}	$C_{avg} \times 100$	C_{LLR}
MFCC-SDC iVector		2.12	0.176	3.95	0.325
CZ	Phonotactic	2.15	0.215	7.00	0.664
	PLLR iVector	2.33	0.223	6.66	0.546
HU	Phonotactic	2.35	0.218	7.28	0.621
	PLLR iVector	**1.41**	**0.127**	**3.17**	**0.308**
RU	Phonotactic	2.85	0.244	6.54	0.571
	PLLR iVector	2.34	0.225	4.38	0.352

Table 2: $C_{avg} \times 100$ and C_{LLR} performance for the baseline systems, the PLLR iVector system and different fusions on the Albayzin 2010 LRE primary task on clean and noisy speech.

System			Clean		Noisy	
			$C_{avg} \times 100$	C_{LLR}	$C_{avg} \times 100$	C_{LLR}
MFCC-SDC iVector		(a)	2.12	0.176	3.95	0.325
HU	Phonotactic	(b)	2.35	0.218	7.28	0.621
	PLLR iVector	**(c)**	**1.41**	**0.127**	**3.17**	**0.308**
Fusion	(a)+(b)		1.10	0.106	2.43	0.211
	(a)+(c)		1.20	0.109	2.65	0.227
	(b)+(c)		1.09	0.092	2.65	0.228
	(a)+(b)+(c)		**0.97**	**0.086**	**1.86**	**0.168**
Fusion	ALL (7 systems, Table 2)		0.82	0.075	1.74	0.169

Table 3: $C_{avg} \times 100$ and C_{LLR} performance for the baseline systems, the PLLR iVector system and different fusions on the Albayzin 2010 LRE primary task on clean and noisy speech.

vement with regard to the best individual system (PLLR HU iVector). Finally, the fusion of all the systems led to the best result: 0.82 $C_{avg} \times 100$, that is, a 41 % relative improvement with regard to the PLLR HU system.Note, however, that this improvement was achieved by fusing 7 systems, more than two times the number of systems used to obtain the second best result.

Results for the noisy-speech condition are consistent with the ones attained on clean-speech. The fusion of the acoustic and PLLR iVector systems yielded the best pairwise performance. As on the clean-speech condition, the fusion of the phonotactic and MFCC-SDC iVector systems yielded the same performance than the fusion of the phonotactic and PLLR iVector systems, and the best fusion involved the three systems, with 1.86 $C_{avg} \times 100$, meaning a 41 % relative improvement with regard to the best individual system. Once again, the PLLR iVector system seems to provide complementary information to baseline systems under all configurations. The fusion of the 7 subsystems shown in Table 3 yielded again the best result on the noisy speech condition: 1.74 $C_{avg} \times 100$, that is, a 45 % relative improvement with regard to the best individual system.

Table 4 shows the confusion matrix for the fusion of PLLR HU iVector, HU phonotactic system and MFCC-SD iVector system on the clean condition of the Albayzin 2010 LRE. As expected, the most confused languages were Spanish and Galician, followed by Spanish and Catalan, and Galician and Catalan. On the other hand, significantly low miss and false alarm probabilities were reached for the Basque, Portuguese and English languages.

5. Conclusions and future work

In this paper, further evidence of the suitability of Phone Log-Likelihood Ratio (PLLR) features for improving SLR performance under the iVector approach has been presented. The performance of a PLLR-based iVector system has been compared to that of two baseline acoustic (MFCC-SDC-based iVector) and phonotactic (Phone-Lattice-SVM) systems, using the Albayzin 2010 LRE dataset

		Target Language					
		Basque	Catalan	English	Galician	Portuguese	Spanish
Test audio	Basque	0.00	0.00	0.00	0.00	0.00	0.00
	Catalan	0.00	1.34	0.00	0.00	1.34	1.34
	English	0.00	0.00	0.00	0.00	0.00	0.00
	Galician	2.48	3.31	0.00	3.31	0.00	14.05
	Portuguese	0.00	0.00	0.00	0.00	0.00	0.00
	Spanish	0.00	0.00	0.00	7.20	0.00	0.80

Table 4: Confusion matrix for the fusion of the PLLR HU iVector, Phonotactic HU and MFCC-SDC iVector systems on the clean condition of the Albayzin 2010 LRE. *Miss probabilities (%)* are shown in the diagonal and *false alarm probabilities (%)* out of the diagonal.

as benchmark. The PLLR-based iVector system not only outperformed the baseline systems, but also proved to contribute complementary information in pairwise fusions with both of them. Finally, the fusion of the three approaches led to very competitive performance. The high performance achieved on noisy speech conditions opens a new track for PLLR features, which will be explored in future work on other databases, such as the Albayzin 2012 LRE dataset, featuring speech on more noisy and challenging conditions.

6. Acknowledgments

This work has been supported by the University of the Basque Country under grant GIU10/18 and project US11/06 and by the Government of the Basque Country under program SAIOTEK (project S-PE12UN55). M. Diez is supported by a research fellowship from the Department of Education, Universities and Research of the Basque Country Government.

References

BenZeghiba, M. F., J. L. Gauvain, and L. Lamel. September 2009. Language Score Calibration using Adapted Gaussian Back-end. In *Proceedings of Interspeech 2009*, pages 2191–2194, Brighton, UK.

Biadsy, Fadi, Julia Hirschberg, and Daniel P. W. Ellis. 2011. Dialect and accent recognition using phonetic-segmentation supervectors. In *Interspeech*, pages 745–748.

Brümmer, N. and J. du Preez. 2006. Application-Independent Evaluation of Speaker Detection. *Computer, Speech and Language*, 20(2-3):230–275.

Brümmer, N. and D.A. van Leeuwen. 2006. On calibration of language recognition scores. In *Proceedings of Odyssey - The Speaker and Language Recognition Workshop*, pages 1–8.

Brümmer, Niko and Edward de Villiers. 2011. The BOSARIS Toolkit: Theory, Algorithms and Code for Surviving the New DCF. In *Proceedings of the NIST 2011 Speaker Recognition Workshop*, Atlanta (GA), USA, December.

Campbell, W. M., F. Richardson, and D. A. Reynolds. 2007. Language Recognition with Word Lattices and Support Vector Machines. In *Proc. IEEE ICASSP*, pages 15–20.

Dehak, N., P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. 2011a. Front-end factor analysis for speaker verification. *IEEE Transactions on ASLP*, 19(4):788–798, May.

Dehak, N., P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak. 2011b. Language Recognition via i-vectors and Dimensionality Reduction. In *Interspeech*, pages 857–860.

DHaro, L.F., O. Glembek, O. Plocht, P. Matejka, M. Soufifar, R. Cordoba, and J. Cernocky. 2012. Phonotactic Language Recognition using i-vectors and Phoneme Posteriogram Counts. In *Proceedings of the Interspeech 2012*, Portland, USA.

Diez, M., A. Varona, M. Penagarikano, L.J. Rodríguez Fuentes, and G.Bordel. 2012. On the Use of Phone Log-Likelihood Ratios as Features in Spoken Language Recognition. In *Proc. IEEE Workshop on SLT*, Miami, Florida, USA.

Fan, R.E., K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *J. Machine Learning Research*, 9:1871–1874.

Martínez, D., L. Burget, L. Ferrer, and N.S. Scheffer. 2012. iVector-based Prosodic System for Language Identification. In *Proceedings of ICASSP*, pages 4861–4864, Japan.

Martínez, D., O. Plchot, L. Burget, O. Glembek, and P. Matejka. 2011. Language Recognition in iVectors Space. In *Proceedings of Interspeech*, pages 861–864, Firenze, Italy.

NIST LRE, 2011. *The 2011 NIST Language Recognition Evaluation Plan (LRE11)*. http://www.nist.gov/itl/iad/mig/upload/LRE11_EvalPlan_releasev1.pdf.

Penagarikano, M., A. Varona, M. Diez, L.J. Rodriguez Fuentes, and G. Bordel. 2012. Study of Different Backends in a State-Of-the-Art Language Recognition System. In *Interspeech 2012*, Portland, Oregon, USA, 9-13 September.

Penagarikano, M., A. Varona, L.J. Rodriguez-Fuentes, and G. Bordel. 2011. Dimensionality Reduction for Using High-Order n-grams in SVM-Based Phonotactic Language Recognition. In *Interspeech*, pages 853–856.

Plchot, O., M. Karafiát, N. Brümmer, O. Glembek, P. Matejka, and E. de Villiers J. Cernocký. 2012. Speaker vectors from Subspace Gaussian Mixture Model as complementary features for Language Identification. In *Odyssey: The Speaker and Language Recognition Workshop*, pages 330–333.

Rodriguez-Fuentes, L. J., M. Penagarikano, G. Bordel, and A. Varona. 2010. The Albayzin 2008 Language Recognition Evaluation. In *Proceedings of Odyssey 2010: The Speaker and Language Recognition Workshop*, pages 172–179, Brno, Czech Republic.

Rodriguez-Fuentes, L. J., M. Penagarikano, A. Varona, M. Diez, and G. Bordel. 2011. The Albayzin 2010 Language Recognition Evaluation. In *Proceedings of Interspeech*, pages 1529–1532, Firenze, Italia.

Rodriguez-Fuentes, L. J., M. Penagarikano, A. Varona, M. Diez, and G. Bordel. 2012. KALAKA-2: a TV broadcast speech database for the recognition of Iberian languages in clean and noisy environments. In *Proceedings of the LREC*, Istanbul, Turkey.

RTTH, 2006. *Spanish Network on Speech Technology*. Web (in Spanish): http://lorien.die.upm.es/∼lapiz/rtth/.

Schwarz, P. 2008. *Phoneme recognition based on long temporal context*. Ph.D. thesis, Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic.

Singer, E., P. A. Torres-Carrasquillo, T. P. Gleason, W. M. Campbell, and D. A. Reynolds. 2003. Acoustic, Phonetic and Discriminative Approaches to Automatic Language Identification. In *Proceedings of Eurospeech (Interspeech)*, pages 1345–1348, Geneva, Switzerland.

Soufifar, M., S. Cumani, L. Burget, and J. Cernocky. 2012. Discriminative Classifiers for Phonotactic Language Recognition with iVectors. In *Proc. IEEE ICASSP*, pages 4853–4856.

Stolcke, A. 2002. SRILM - An extensible language modeling toolkit. In *Interspeech*, pages 257–286.

Varona, Amparo, Mikel Penagarikano, Luis Javier Rodriguez Fuentes, Mireia Diez, and Germán Bordel. 2010. Verification of the four spanish official languages on tv show recordings. In *XXV Congreso de la Sociedad Espaï¿½ola para el Procesamiento de Lenguaje Natural (SEPLN)*, Valencia, Spain, 8-10 September.

Young, S., G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Lui, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. 2006. *The HTK Book (for HTK Version 3.4)*. Entropic, Ltd., Cambridge, UK.

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 161-168 recibido 18-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Nueva técnica de fusión de clasificadores aplicada a la mejora de la segmentación de audio

Applying a new classifier fusion technique to audio segmentation

David Tavarez[1], Eva Navas[1], Daniel Erro[1,2], Ibon Saratxaga[1], Inma Hernaez[1]
[1] AHOLAB, University of the Basque Country (UPV/EHU), Bilbao, Spain
[2]Basque Foundation for Science (IKERBASQUE), Bilbao, Spain
{david, eva, derro, ibon, inma}@aholab.ehu.es

Resumen: Este artículo presenta un nuevo algoritmo de fusión de clasificadores a partir de su matriz de confusión de la que se extraen los valores de precisión (*precision*) y cobertura (*recall*) de cada uno de ellos. Los únicos datos requeridos para poder aplicar este nuevo método de fusión son las clases o etiquetas asignadas por cada uno de los sistemas y las clases de referencia en la parte de desarrollo de la base de datos. Se describe el algoritmo propuesto y se recogen los resultados obtenidos en la combinación de las salidas de dos sistemas participantes en la campaña de evaluación de segmentación de audio Albayzin 2012. Se ha comprobado la robustez del algoritmo, obteniendo una reducción relativa del error de segmentación del 6.28 % utilizando para realizar la fusión el sistema con menor y mayor tasa de error de los presentados a la evaluación.
Palabras clave: Fusión de clasificadores, clasificación y segmentación de audio

Abstract: This paper presents a new classifier fusion algorithm based on the confusion matrixes of the classifiers which are used to extract the corresponding precision and recall values. The only data needed to be able to apply this new fusion method are the classes or labels assigned by each of the classifiers as well as the reference classes in the development part of the database. The proposed algorithm is described and it is applied to the fusion of two audio segmentation systems that took part in Albayzin 2012 evaluation campaign. The robustness of the algorithm has been assessed and a relative improvement of 6.28 % has been achieved when combining the results of the best and worst systems presented to the evaluation.
Keywords: Classifier fusion, audio classification and segmentation

1. Introducción

La segmentación de audio consiste en dividir una grabación en regiones homogéneas de acuerdo a su contenido, asignando a cada segmento la etiqueta de la clase a la que pertenece. En función de la aplicación para la que se realice, el objetivo de la segmentación de audio puede ser muy diferente: separar la voz de la música y el ruido (Lu, Zhang, y Jiang, 2002), separar las voces masculinas de las femeninas (Ore, Slyh, y Hansen, 2006), separar los segmentos que corresponden a distintos locutores (Moattar y Homayounpour, 2012), etc. Tiene muchas aplicaciones y comúnmente se utiliza como primer paso de pre-procesado para mejorar los resultados de otros sistemas como los de reconocimiento automático de habla (Rybach y Gollan, 2009), identificación de locutores (Reynolds y Torres-Carrasquillo, 2005), recuperación de información e indexado de audio basada en su contenido (Meinedo y Neto, 2003) (Aguilo et al., 2009), etc.

Las campañas competitivas de evaluación son una herramienta muy adecuada para determinar de manera objetiva la validez de los algoritmos desarrollados. En estas campañas distintos grupos de investigación prueban sus algoritmos sobre una base de datos común, lo que permite comparar el rendimiento de los mismos e identificar las técnicas más adecuadas para cada etapa del sistema. La Red

ISSN 1135-5948

Temática en Tecnologías del Habla [1] organiza las campañas competitivas de evaluación Albayzin que se celebran cada dos años y evalúan distintos aspectos relacionados con las tecnologías del habla. La segmentación de audio se ha incluido en las dos últimas campañas realizadas, Albayzin 2010 (Butko y Nadeu, 2011) y 2012 [2].

En los problemas de clasificación en los que se comparan diferentes métodos, el que obtiene los mejores resultados suele ser el sistema seleccionado para realizar la clasificación. Sin embargo, en general se observa que los errores cometidos por el resto de los sistemas no son comunes y sus resultados podrían utilizarse para mejorar el rendimiento general del sistema seleccionado, mediante técnicas de fusión de clasificadores (Kittler y Hatef, 1998), (Xu, Krzyzak, y Suen, 1992). De hecho, en diferentes campañas de evaluación con objetivos de clasificación muy distintos, la fusión de varios sistemas funciona mejor que cualquiera de ellos por separado (Schuller, 2012). La fusión de clasificadores puede realizarse a varios niveles (Ruta y Gabrys, 2000):

- *a nivel de datos*: se combinan datos provenientes de diferentes fuentes para realizar la clasificación, como sucede cuando se combinan diferentes rasgos biométricos (voz, huella dactilar, imagen facial,...) en la identificación de personas (Jain y Ross, 2004).
- *a nivel de características*: se combinan distinto tipo de características extraídas a partir de los datos de que se dispone para realizar la clasificación. Un típico ejemplo se produce en los sistemas de verificación de locutor que utilizan información segmental y prosódica extraída a partir de la voz de los locutores (Reynolds et al., 2003).
- *a nivel de decisión*: se combinan directamente los resultados de los clasificadores. Esta combinación puede realizarse bien a nivel de etiqueta, cuando únicamente se dispone como dato de la clase asignada por cada clasificador (Asman y Landman, 2011) o bien a nivel de confianza o *score* cuando se dispone no sólo de la clase a la que pertenece cada segmento sino también de la confianza con la que el clasificador ha tomado la decisión (Ross y Jain, 2003).

En este artículo se propone una nueva técnica de fusión de clasificadores a nivel de etiqueta que se ha aplicado con éxito a la fusión de los resultados de dos sistemas de segmentación de audio participantes en la campaña de Albayzin 2012.

La sección 2 del artículo presenta la técnica de fusión de clasificadores propuesta. En la sección 3 se describen resumidamente los datos más relevantes de la campaña Albayzin 2012 de evaluación de sistemas de segmentación de audio. Los resultados obtenidos aplicando la técnica propuesta a dos sistemas participantes en dicha evaluación se presentan y analizan en la sección 4 y finalmente en la sección 5 se exponen las conclusiones del trabajo.

2. Técnica de fusión propuesta

En esta sección se describe el algoritmo propuesto para realizar la fusión de los resultados de dos clasificadores diferentes, en base a su matriz de confusión. Es un algoritmo de fusión que funciona a nivel de etiqueta en el que la única información requerida para poder ser aplicado son las clases asignadas por cada uno de los sistemas y las clases de referencia en la parte de desarrollo de la base de datos

Dados dos clasificadores, c_1 y c_2, en un escenario multiclase, el algoritmo propuesto trata de evaluar la confiabilidad de las decisiones de cada uno de los clasificadores, basándose en sus valores de precisión (*precision*) y cobertura (*recall*) para las clases emitidas. Supongamos que para un caso concreto tenemos como salidas las clases a y b para los clasificadores c_1 y c_2 respectivamente. El algoritmo propuesto plantea que para estimar la confiabilidad de la decisión tomada por el clasificador c_1 hay que tener en cuenta no sólo su precisión para la clase a, sino también la probabilidad de que el clasificador c_2 confunda la clase a con la clase b. Es decir, siendo c_1 y c_2 independientes, la probabilidad de que la respuesta real sea a equivale al producto de la probabilidad de que c_1 haya acertado al elegir a por la probabilidad de que c_2 se haya equivocado al elegir b. Estimaremos esta probabilidad por medio de la tasa de falsos negativos para este segundo clasificador, cal-

[1] http://www.rthabla.es/

[2] http://iberspeech2012.ii.uam.es/index.php/call-for-evalproposals-2/audio-segmentation

culada en la parte de desarrollo de la base de datos.

Con el fin de evaluar dicha confiabilidad de la decisión, en primer lugar es necesario obtener las matrices de confusión de los dos clasificadores. Para ello se utilizan los resultados que los clasificadores logran en la parte de desarrollo de la base de datos. Podemos ver un ejemplo de dicha matriz para un caso en que se consideran tres clases diferentes en la tabla 1, en la que se representan de arriba a abajo las clases reales, a, b y c, y de izquierda a derecha las predicciones realizadas por un clasificador, a', b' y c'. De forma general, V_x representa el número de positivos verdaderos o aciertos del clasificador para la clase x, $F_{xx'}$ el número de falsos negativos o errores cometidos por el clasificador al no identificar la clase real x y predecir en su lugar x', T_x el número total de ejemplos de la clase x en la parte de desarrollo de la base de datos y $T_{x'}$ el número total de ejemplos de la base de datos marcados como clase x' por el clasificador correspondiente.

Real / Pred.	a	b	c	Total
a'	V_a	$F_{ba'}$	$F_{ca'}$	$T_{a'}$
b'	$F_{ab'}$	V_b	$F_{cb'}$	$T_{b'}$
c'	$F_{ac'}$	$F_{bc'}$	V_c	$T_{c'}$
Total	T_a	T_b	T_c	-

Tabla 1: Ejemplo de matriz de confusión para un escenario con tres clases

Para evaluar la confiabilidad de la decisión de cada clasificador tendremos en cuenta además de la precisión del mismo para la clase propuesta, el error cometido por el otro clasificador al escoger una salida diferente, por lo que utilizaremos las matrices de confusión de los dos clasificadores implicados en la fusión a la hora de analizar la confiabilidad de la decisión de cada uno.

Por un lado se considerará la precisión del clasificador para cada clase, que de acuerdo con la nomenclatura anterior es:

$$TP_x = \frac{V_x}{T_{x'}} \tag{1}$$

donde x representa la clase propuesta por el clasificador, con V_x y $T_{x'}$ definidos anteriormente. Mediante este valor se estima la probabilidad de que el segmento clasificado pertenezca realmente a la clase a la que ha sido asignado por el clasificador.

Adicionalmente, podemos analizar el error cometido por cada clasificador por medio de la tasa de falsos negativos, obtenida siguiendo el ejemplo anterior como:

$$FN_{xx'} = \frac{F_{xx'}}{T_x} \tag{2}$$

donde x representa la clase real (la de referencia) y x' la clase propuesta por el clasificador, con $F_{xx'}$ y T_x definidos anteriormente. Este valor se utiliza como estimación de la probabilidad de que el clasificador haya confundido la clase real x con la x'.

Si en un caso concreto las salidas de los sistemas originales coinciden, la clase propuesta por ambos es asignada directamente. En caso de obtener salidas diferentes para los sistemas originales, procedemos a evaluar la confiabilidad de la decisión de cada uno con el fin de seleccionar una de las dos clases propuestas.

Aplicando un razonamiento probabilístico, para calcular la confiabilidad de la decisión tomada por el primer sistema multiplicamos su precisión para la clase propuesta y la tasa de falsos negativos del segundo clasificador en función de la salida de ambos. Siguiendo el ejemplo anterior tenemos:

$$r_{xy'}\big|_1 = TP_x|_1 \cdot FN_{xy'}\big|_2 \tag{3}$$

donde x representa la clase propuesta por el primer clasificador e y' la clase propuesta por el segundo clasificador. En este caso suponemos que la clase propuesta por el primer clasificador, x, es correcta, por lo que la confiabilidad de la decisión del primer clasificador dependerá de su precisión para esta clase, $TP_x|_1$, y del supuesto error cometido por el segundo clasificador al escoger la clase y, es decir, la tasa de falsos negativos del segundo clasificador para la clase y cuando la clase real es x, $FN_{xy'}\big|_2$. Se considera que las decisiones de los dos clasificadores son independientes y por ello se multiplican las probabilidades para estimar la probabilidad conjunta que representa la confiabilidad de la decisión.

Una vez obtenida la confiabilidad de la decisión del primer clasificador, procedemos del mismo modo para evaluar la confiabilidad de la decisión del segundo clasificador. En este caso multiplicamos su precisión para la clase que ha seleccionado y la tasa de falsos negativos del otro clasificador en función de la

salida de ambos. Siguiendo el ejemplo anterior tenemos:

$$r_{yx'}|_2 = TP_y|_2 \cdot FN_{yx'}|_1 \quad (4)$$

donde ahora y representa la clase propuesta por el segundo clasificador, que se supone correcta, y x' la clase propuesta por el primer clasificador.

Una vez evaluada la confiabilidad de la decisión de cada clasificador, se asigna en cada caso la clase propuesta por el clasificador cuya confiabilidad obtenida resulta mayor.

3. Campaña Albayzin 2012

La campaña de evaluación de sistemas de segmentación de audio Albayzin 2012 consistió en la segmentación de audio *broadcast*, asignando a los segmentos obtenidos etiquetas para indicar la presencia de voz, música y ruido en cada uno de ellos, pudiendo existir solapamiento entre las tres clases en cualquier instante.

3.1. Base de datos

Los organizadores de la campaña proporcionaron dos bases de datos de audio diferentes correspondientes a programas de noticias para ser utilizadas en el desarrollo de los sistemas de segmentación.

La primera, utilizada también en la campaña Albayzin 2010 de evaluación de sistemas de segmentación y diarización, está formada por unas 87 horas de grabaciones de programas emitidos por el canal catalán de televisión 3/24. Estos datos podían ser utilizados para realizar el entrenamiento de los sistemas. La distribución de las clases de audio contenidas en esta base de datos es la siguiente: 37 % de voz limpia, 5 % de música, 15 % de voz con música de fondo, 40 % de voz con ruido de fondo y 3 % de otros. En esta última clase se engloba todo el material que no pertenece a las cuatro clases anteriores, incluyendo el ruido.

La segunda base de datos proporcionada por la organización proviene de la Corporación Aragonesa de Radio y Televisión (CARTV), que donó parte de su archivo de Aragón Radio. Está formada por unas 20 horas de audio con la distribución de clases que se describe a continuación: 22 % de voz limpia, 9 % de música, 31 % de voz con música de fondo, 26 % de voz con ruido de fondo y 12 % de otros. En este caso la clase 'otros' contiene tanto los silencios como el ruido y las combinaciones de clases que no se han mencionado. Aproximadamente 4 horas podían ser utilizadas para el entrenamiento de los sistemas y las 16 horas restantes fueron empleadas por la organización para su evaluación.

3.2. Métrica utilizada

Al igual que en las evaluaciones organizadas por el NIST (*National Institute of Standards and Technology*), la métrica utilizada para evaluar el funcionamiento de los sistemas ha sido el SER (Tasa de Error de Segmentación o *Segmentation Error Rate*), que se corresponde con la fracción de tiempo que no ha sido correctamente atribuida a la clase correspondiente (voz, música y ruido en este caso). En las zonas de solapamiento entre clases la duración del segmento se atribuye a todas las clases presentes en el mismo, por lo que un mismo segmento temporal puede ser considerado más de una vez en los cálculos.

El SER se calcula como la suma de tres tipos de errores: el porcentaje de tiempo que es asignado a una clase incorrecta (Error de Clase o *Class Error Time*), el porcentaje de tiempo en el que una clase está presente pero no ha sido etiquetada (Error de Omisión o *Missed Class Time*) y el porcentaje de tiempo en que se ha etiquetado una clase cuando realmente no estaba presente (Error de Inserción o *False Alarm Time*). Todos estos errores se han calculado mediante las herramientas de evaluación proporcionadas por el NIST (NIST, 2009).

3.3. Resultados de la evaluación

En la campaña de evaluación Albayzin 2012 tomaron parte 6 sistemas desarrollados por 5 grupos de investigación diferentes. La tabla 2 recoge los resultados obtenidos por los distintos sistemas en la parte de evaluacion de la base de datos.

Por respeto a los otros participantes y dado que la técnica de fusión no depende de sistemas concretos sino sólo de las etiquetas que proporcionan, nos referiremos a los sistemas ajenos con los nombres ficticios S2...S6 en atención al puesto que ocuparon en la evaluación.

3.4. Sistemas seleccionados para la fusión

Para comprobar los resultados de la técnica de fusión propuesta se seleccionaron los sistemas que mejor y peor resultados obtuvieron

Sistema	SER (Test)
AHOLAB-EHU	25.78 %
S2	26.53 %
S3	28.12 %
S4	33.30 %
S5	39.55 %
S6	40.01 %

Tabla 2: Resultados de los sistemas de segmentación de audio presentados en la campaña de Albayzin 2012

en la evaluación Albayzin 2012, suponiendo que si de este modo se consiguen mejoras en los resultados, realizando la fusión entre sistemas con menor tasa de error la mejora sería mayor. Por lo tanto se utilizarán en los experimentos de fusión el sistema presentado por Aholab (Tavarez et al., 2012) y el sistema S6. Además, con el fin de comprobar los resultados cuando los dos clasificadores originales presentan bajas tasas de error, se seleccionaron dos de los sistemas que mejores resultados obtuvieron en la evaluación, Aholab y el sistema S3.

4. Experimentos y resultados

En esta sección se muestran los resultados obtenidos al aplicar el algoritmo de fusión descrito anteriormente a los sistemas de segmentación de audio seleccionados.

En primer lugar, se ha realizado un mapeo de las clases para evitar el solapamiento de las mismas. Como se ha comentado en la sección 3, en el problema de segmentación planteado en Albayzin 2012 las clases podían solaparse. Sin embargo, para llevar a cabo la fusión de los sistemas con el método propuesto, la salida de cada sistema debe ser única en cada segmento, por lo que las clases originales (voz, música y ruido), han sido sustituidas por las diferentes combinaciones posibles entre ellas: voz limpia, música, ruido, voz con música, voz con ruido, etc.

A continuación se calcula la matriz de confusión de cada sistema usando la parte de desarrollo de la base de datos de Aragón Radio. Posteriormente se realiza la combinación de las salidas de los sistemas originales mediante la técnica de fusión propuesta y se obtienen las marcas finales para cada segmento deshaciendo el mapeo realizado inicialmente. La tabla 3 muestra el resultado, en términos de SER, de la fusión de los dos sistemas seleccionados tanto en la parte de desarrollo y como en la de evaluación de la base de datos.

Sistema	Desarrollo	Evaluación
AHOLAB	19,97 %	25,78 %
S6	35,93 %	40,01 %
Fusión	**18,71 %**	**24,16 %**

Tabla 3: Resultado de la fusión de los sistemas Aholab y S6

Podemos observar cómo se logra una mejora de los resultados obtenidos en ambos casos, con una reducción relativa del SER del 6.3 % en la parte de desarrollo y del 6.28 % en la parte de evaluación respecto al mejor de los sistemas. Se trata de una mejora significativa teniendo en cuenta que uno de los sistemas utilizados parte con un SER del 40 %, el mayor de la campaña de evaluación. El método de fusión propuesto es capaz de obtener información suficiente de éste sistema y de utilizarla para mejorar los resultados del sistema propuesto por AHOLAB, lo que demuestra la robustez del algoritmo desarrollado.

Con el fin de analizar el origen de la mejora de los resultados, se ha estudiado el comportamiento de la fusión en cada una de las clases consideradas. Para ello se ha evaluado el error cometido para cada evento original (voz, música y ruido) considerados individualmente. Se incluyen además el error de omisión (MC) y el error de inserción (FA) de cada clase, referidos al tiempo total asignado a dicha clase en la referencia, tal y como son calculados por las herramientas de evaluación de NIST.

Sistema		MC	FA	SER
AHOLAB	Desarrollo	3,9 %	0,8 %	4,63 %
S6		0,8 %	1,6 %	2,34 %
Fusión		**4,3 %**	**0,4 %**	**4,12 %**
AHOLAB	Evaluación	3,3 %	0,9 %	4,19 %
S6		0,6 %	3,0 %	3,56 %
Fusión		**3,6 %**	**0,5 %**	**4,12 %**

Tabla 4: Detalle del error cometido por los sistemas Aholab y S6 y la fusión de ambos para la clase 'voz'

La tabla 4 muestra el resultado obtenido para la clase de voz en la parte de desarrollo y de evaluación de la base de datos. Se pue-

de observar cómo el resultado de la fusión en este caso es prácticamente nulo, debido principalmente a los buenos resultados para esta clase en cada uno de los sistemas originales, con un SER en torno a sólo el 4 % en el sistema AHOLAB, el peor sistema en este caso.

Sistema		MC	FA	SER
AHOLAB	Desarrollo	26,8 %	4,1 %	30,86 %
S6		56,6 %	1,8 %	58,41 %
Fusión		**25,4 %**	**4,5 %**	**29,94 %**
AHOLAB	Evaluación	36,9 %	6,7 %	43,59 %
S6		64,3 %	1,4 %	65,76 %
Fusión		**33,9 %**	**5,4 %**	**39,34 %**

Tabla 5: Detalle del error cometido por los sistemas Aholab y S6 y la fusión de ambos para la clase 'música'

El resultado obtenido para la clase de música se muestra en la tabla 5. En este caso se consigue una reducción del SER del 10 % en la parte de test de la base de datos. El algoritmo desarrollado permite utilizar las diferencias para esta clase (en realidad cuatro clases tras el mapeo realizado: música, voz con música, música con ruido y música con voz y ruido) entre los dos sistemas para mejorar el resultado final, a pesar de que los resultados del sistema S6 son considerablemente inferiores a los del sistema propuesto por el grupo AHOLAB.

Por último, la tabla 6 muestra el resultado obtenido para la clase de ruido. Se puede observar cómo, al igual que en el caso de la voz, el resultado de la fusión en este caso es casi inapreciable, debido principalmente a los resultados del sistema S6 con un 162,18 % de SER (recordemos que el porcentaje está referido al tiempo total asignado a cada clase en la referencia, por lo que puede superar el 100 %). En este caso no es posible extraer información de utilidad con la que mejorar los resultados. Sin embargo, también cabe resaltar que el resultado final no se ve comprometido a pesar de estas tasas de error superiores al 100 % y se mantiene del orden de lo logrado por el mejor de los sistemas, lo que demuestra la robustez del método desarrollado.

Tras realizar este estudio del comportamiento respecto a cada etiqueta, se puede observar cómo la mejora obtenida al aplicar el método de fusión propuesto se debe al resultado obtenido en la clase 'música', que fue la que más dificultades de clasificación planteó en la campaña Alabyzin 2012.

Sistema		MC	FA	SER
AHOLAB	Desarrollo	33,3 %	9,5 %	42,85 %
S6		52,7 %	61,3 %	113,95 %
Fusión		**36,0 %**	**7,5 %**	**43,57 %**
AHOLAB	Evaluación	34,8 %	28,2 %	63,03 %
S6		42,5 %	119,6 %	162,18 %
Fusión		**38,8 %**	**24,8 %**	**63,61 %**

Tabla 6: Detalle del error cometido por los sistemas Aholab y S6 y la fusión de ambos para la clase 'ruido'

A continuación, este algoritmo se ha aplicado también a la fusión entre el sistema propuesto por el grupo AHOLAB y el sistema S3 (tercer mejor sistema de la campaña de evaluación de Albayzin 2012). El objetivo de este experimento es comprobar los resultados de la técnica de fusión propuesta cuando los dos clasificadores originales presentan bajas tasas de error. La tabla 7 muestra el resultado, en términos de SER, de la fusión de los dos sistemas.

Sistema	Desarrollo	Evaluación
AHOLAB	19,97 %	25,78 %
S3	21,24 %	28,12 %
Fusión	**16,13 %**	**18,86 %**

Tabla 7: Resultado de la fusión de los sistemas Aholab y S3

Podemos observar cómo en este caso, en el que se cuenta con buenos resultados de partida en ambos sistemas, el algoritmo desarrollado obtiene una reducción relativa del SER del 19.5 % en la parte de desarrollo y del 26.8 % en la parte de evaluación respecto al mejor de los sistemas, lo que demuestra la validez del método de fusión propuesto, cuando se utilizan como datos de partida los de un clasificador tipo.

Al igual que en el caso anterior, se ha estudiado el comportamiento de la fusión en cada una de las clases consideradas. Para ello evaluamos de nuevo el error cometido para cada evento original considerados individualmente.

Sistema		MC	FA	SER
AHOLAB	Desarrollo	3,9 %	0,8 %	4,63 %
S3		0,1 %	5,5 %	5,65 %
Fusión		**1,5 %**	**0,8 %**	**2,32 %**
AHOLAB	Evaluación	3,3 %	0,9 %	4,19 %
S3		0,2 %	8,5 %	8,69 %
Fusión		**1,7 %**	**1,4 %**	**3,11 %**

Tabla 8: Detalle del error cometido por los sistemas Aholab y S3 y la fusión de ambos para la clase 'voz'

La tabla 8 muestra el resultado obtenido para la clase de voz en la parte de desarrollo y de evaluación de la base de datos. Se puede observar cómo en este caso sí se obtiene cierta mejora, a pesar de contar con peores resultados para esta clase que en el caso de utilizar los resultados de S6. El reparto diferente de clases entre los dos sistemas permite mejorar el resultado final.

Sistema		MC	FA	SER
AHOLAB	Desarrollo	26,8 %	4,1 %	30,86 %
S3		25,9 %	4,1 %	29,93 %
Fusión		**10,3 %**	**6,3 %**	**16,59 %**
AHOLAB	Evaluación	36,9 %	6,7 %	43,59 %
S3		37,5 %	5,4 %	42,91 %
Fusión		**19,2 %**	**6,7 %**	**25,92 %**

Tabla 9: Detalle del error cometido por los sistemas Aholab y S3 y la fusión de ambos para la clase 'música'

En la tabla 9 se presenta el resultado obtenido para la clase de música. Al igual que en el caso anterior, se consigue una reducción considerable del SER, 40 % en la parte de test de la base de datos. En este caso el error original de los sistemas es menor y la mejora obtenida tras aplicar el algoritmo de fusión es más elevada.

La tabla 10 muestra el resultado obtenido para la clase de ruido. Se puede observar cómo, al igual que en el caso de la voz y la música, se ha conseguido una importante reducción del SER, 22 % en la parte de test. En este caso los dos clasificadores aportan información de utilidad, a pesar del elevado error de ambos para esta clase, y el algoritmo de fusión propuesto es capaz de mejorar el resultado de ambos.

Sistema		MC	FA	SER
AHOLAB	Desarrollo	33,3 %	9,5 %	42,85 %
S3		14,3 %	20,8 %	35,09 %
Fusión		**30,3 %**	**4,1 %**	**34,42 %**
AHOLAB	Evaluación	34,8 %	28,2 %	63,03 %
S3		19,0 %	65,9 %	84,87 %
Fusión		**29,5 %**	**19,6 %**	**49,07 %**

Tabla 10: Detalle del error cometido por los sistemas Aholab y S3 y la fusión de ambos para la clase 'ruido'

Realizado el estudio respecto a cada etiqueta, se puede observar cómo hemos obtenido una mejora considerable de los resultados en todas las clases, con una reducción importante del SER, salvo en la clase de voz en la que los resultados originales eran buenos y el margen de mejora era menor. Esto demuestra la validez del método de fusión propuesto.

5. Conclusiones

Se ha descrito un nuevo sistema de fusión de clasificadores a nivel de etiqueta en base a las matrices de confusión de cada clasificador, obtenidas a partir de la parte de desarrollo de la base de datos utilizada. Además se ha probado con éxito su uso en la combinación de las salidas de sistemas de segmentación de audio propuestos en la campaña de evaluación Albayzin 2012.

Asimismo, se ha comprobado la robustez del algoritmo, obteniendo una mejora de los resultados aún cuando uno de los sistemas utilizados presenta originalmente una tasa de error considerablemente elevada, sin empeorar los resultados del mejor de los dos sistemas considerados.

Este algoritmo de fusión puede ser utilizado para combinar los resultados de más de dos sistemas de clasificación, sin más que aplicarlo de manera jerárquica. También es posible ampliar el algoritmo a la fusión de varios clasificadores considerando a la hora de valorar la confiabilidad de la decisión la tasa de falsos negativos de más de un clasificador.

6. Agradecimientos

Este trabajo ha sido financiado parcialmente por la UPV/EHU (Ayudas para la Formación de Personal Investigador), el Gobierno Vasco (proyecto Ber2Tek, IE12-333) y el Ministerio de Economía y Competitividad (Proyecto SpeechTech4All,

http://speechtech4all.uvigo.es/, TEC2012-38939-C03-03).

Bibliografía

Aguilo, M., T. Butko, A. Temko, y C. Nadeu. 2009. A hierarchical architecture for audio segmentation in a broadcast news task. En *Proc. I Iberian SLTech*, páginas 17–20, Porto Salvo, Portugal.

Asman, A. J. y B. A. Landman. 2011. Robust statistical label fusion through consensus level, labeler accuracy, and truth estimation (COLLATE). *IEEE Transactions on Medical Imaging*, 30(10):1779–1794.

Butko, T. y C. Nadeu. 2011. Audio segmentation of broadcast news in the Albayzin-2010 evaluation: overview, results, and discussion. *EURASIP Journal on Audio, Speech, and Music Processing*, 2011(1):1–10.

Jain, A. K. y A. Ross. 2004. Multibiometric systems. *Communications of the ACM*, 47(1):34–40, Enero.

Kittler, J. y M. Hatef. 1998. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239.

Lu, L., H. J. Zhang, y H. Jiang. 2002. Content analysis for audio classification and segmentation. *IEEE Transactions on Speech and Audio Processing*, 10(7):504–516.

Meinedo, H. y J. Neto. 2003. Audio segmentation, classification and clustering in a broadcast news task. En *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volumen 2, páginas 5–8, Hong-Kong, China.

Moattar, M. H. y M. M. Homayounpour. 2012. A review on speaker diarization systems and approaches. *Speech Communication*, 54(10):1065–1103, Junio.

NIST. 2009. The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan. Online: http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meetingeval-plan-v2.pdf, accessed on 15 April 2013.

Ore, B. M., R. E. Slyh, y E. G. Hansen. 2006. Speaker Segmentation and Clustering using Gender Information. En *Proceedings IEEE Odyssey'06 Conference*, páginas 1–8.

Reynolds, D., W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adomi, D. Kluracek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, y S. Xiang. 2003. The SuperSID Project: Exploiting High-level Information. En *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volumen 4, páginas 784–787, Hong-Kong, China.

Reynolds, Douglas A y P. Torres-Carrasquillo. 2005. Approaches and applications of audio diarization. En *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, páginas 953–956, Philadelphia, USA.

Ross, A. y A. K. Jain. 2003. Information fusion in biometrics. *Pattern Recognition Letters*, 24(13):2115–2125, Septiembre.

Ruta, D. y B. Gabrys. 2000. An overview of classifier fusion methods. *Computing and Information systems*, 7:1–10.

Rybach, D. y C. Gollan. 2009. Audio segmentation for speech recognition using segment features. En *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, páginas 4197 – 4200, Taipei, Taiwan.

Schuller, B. 2012. The Computational Paralinguistics Challenge. *Signal Processing Magazine, IEEE*, (July):97–101.

Tavarez, D., E. Navas, D. Erro, y I. Saratxaga. 2012. Audio Segmentation System by Aholab for Albayzin 2012 Evaluation Campaign. En *Iberspeech*, páginas 577–584, Madrid, Spain.

Xu, L., A. Krzyzak, y C. Y. Suen. 1992. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(3):418–435.

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 169-175 recibido 29-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Sistema de Conversión Texto a Voz de Código Abierto Para Lenguas Ibéricas

Open-Source Text to Speech Synthesis System for Iberian Languages

Agustin Alonso[1], Iñaki Sainz[1], Daniel Erro[1,2], Eva Navas[1], Inma Hernaez[1]
[1]AHOLAB, University of the Basque Country (UPV/EHU), Bilbao, Spain
[2]Basque Foundation for Science (IKERBASQUE), Bilbao, Spain
{agustin,inaki,derro,eva,inma}@aholab.ehu.es

Resumen: Este artículo presenta un conversor texto a voz basado en síntesis estadística que por primera vez permite disponer en un único sistema de las cuatro lenguas oficiales en España además del inglés. Tomando como punto de partida el sistema AhoTTS existente para el castellano y el euskera, se le han añadido funcionalidades para incluir el catalán, el gallego y el inglés utilizando módulos disponibles en código abierto. El sistema resultante, denominado AhoTTS multilingüe, ha sido liberado en código abierto y ya está siendo utilizado en aplicaciones reales.
Palabras clave: Texto a Voz, Multilingüismo, Herramienta Software, Síntesis Estadística, Código Abierto

Abstract: This paper presents a text-to-speech system based on statistical synthesis which, for the first time, allows generating speech in any of the four official languages of Spain as well as English. Using the AhoTTS system already developed for Spanish and Basque as a starting point, we have added support for Catalan, Galician and English using the code of available open-source modules. The resulting system, named multilingual AhoTTS, has also been released as open-source and it is already being used in real applications.
Keywords: Text-to-Speech, Multilingualism, Software Tool, Statistical Synthesis, Open Source

1 Introducción

Un sistema de conversión texto a voz (CTV) es un sistema que convierte una entrada de texto en una salida en forma de señal de audio cuyo contenido se corresponde con el mensaje del texto de entrada.

Los sistemas CTV actuales pueden descomponerse en dos módulos tal y como muestra la Figura 1. El primero, fuertemente dependiente del idioma, toma como entrada el texto a sintetizar y genera a su salida etiquetas que describen lingüística, fonética y prosódicamente dicha entrada. Estas etiquetas alimentan el segundo módulo, el motor de síntesis que en sí mismo es independiente del idioma. Este módulo emplea voces entrenadas a partir de bases de datos que son dependientes del idioma para sintetizar la señal de voz de salida en función de las etiquetas de entrada. Respecto a los métodos de síntesis, las tecnologías más empleadas actualmente son dos: la selección y concatenación de unidades y la síntesis estadística paramétrica. La selección de unidades (Hunt y Black, 1996) consiste en generar la señal de voz concatenando segmentos de voz real. La síntesis estadística consiste en reconstruir la señal a partir de parámetros acústicos extraídos de modelos matemáticos previamente entrenados con señales de voz real (Zen et al., 2009). También se han desarrollado sistemas híbridos que combinan ambas técnicas (Ling et al., 2007).

ISSN 1135-5948

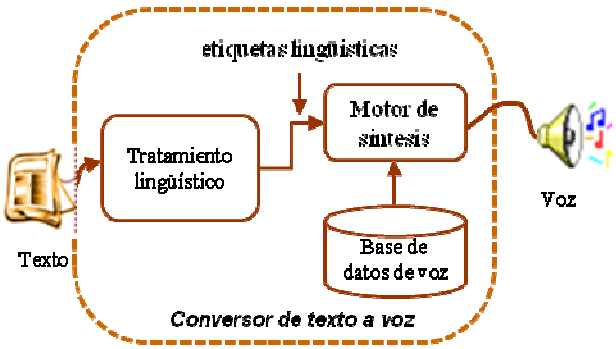

Figura 1: Estructura genérica de un sistema de conversión de texto a voz

Actualmente existen diversos conversores en código abierto desarrollados por equipos de investigación universitarios de distintas universidades para las diferentes lenguas oficiales de España que emplean múltiples métodos tanto en el procesado lingüístico como en la síntesis. Entre ellos destacan para los idiomas castellano y euskera AhoTTS[1], para el castellano y el gallego Cotovia[2], y para el catalán Festcat[3].

Aunque existen trabajos realizados para disponer en un único CTV de todas las lenguas oficiales (Rodríguez, Escalada y Torre, 1998) se trata de sistemas propietarios no disponibles libremente. La necesidad de disponer de un CTV multilingüe surge en el contexto del proyecto TVSocial[4], en el que se pretendía crear una plataforma de difusión de contenidos de televisión con bajo coste, utilizando para ello un CTV multilingüe de las características mencionadas.

En este artículo se explica el procedimiento que se ha seguido para la creación de un conversor de texto a voz de código abierto con la posibilidad de elegir entre las cuatro lenguas oficiales de España más el inglés La base de la que se ha partido es el sistema AhoTTS desarrollado por el grupo Aholab de la UPV/EHU.

Primero, en la sección 2, se describen las características básicas del sistema AhoTTS. A continuación, en la sección 3, se explican los pasos seguidos para integrar módulos de procesado lingüístico y voces nuevas para los idiomas no disponibles inicialmente. Finalmente en la sección 4 se resumen las conclusiones y se mencionan las líneas futuras de trabajo que se tiene pensado seguir.

2 *Sistema AhoTTS*

El sistema AhoTTS es la plataforma de conversión texto a voz de Aholab que lleva siendo desarrollada desde 1992. Programado en C/C++, es un sistema modular, multiplataforma y multilingüe. Inicialmente los idiomas disponibles eran el castellano y el euskera, aunque el esfuerzo de investigación se ha centrado principalmente en el segundo (Hernaez, 1995) (Navas et al., 2002a) (Navas et al., 2002b) (Navas, 2003).

AhoTTS integra la práctica totalidad de las tecnologías actuales de síntesis. Así, permite sintetizar la señal de voz empleando tanto la selección y concatenación de unidades (Sainz et al., 2008) como la síntesis estadística paramétrica (Erro et al, 2010). Para este segundo caso, Aholab ha desarrollado tecnología propia de parametrización y reconstrucción de señales (Erro et al, 2011). También se han hecho experimentos utilizando tecnologías híbridas (Sainz et al., 2011) con resultados muy satisfactorios.

De entre estas tecnologías de síntesis, la estadística presenta varias ventajas prácticas respecto al resto (Zen et al, 2009):

- Produce una voz de características más estables y con mayor inteligibilidad, especialmente cuando las bases de datos son pequeñas.
- El entrenamiento de la voz es automático, robusto y no requiere el ajuste manual de múltiples parámetros.
- El tamaño en disco de las voces generadas es menor, lo cual facilita su almacenamiento e integración en sistemas embebidos o con recursos limitados.
- Es fácil modificar las características acústicas de la voz en tiempo de síntesis y da mayor flexibilidad para generar nuevas voces mediante técnicas de adaptación, interpolación, etc.

Debido a estas ventajas se ha optado por este método para la inclusión de los nuevos idiomas en AhoTTS.

La calidad de AhoTTS viene avalada por los excelentes resultados cosechados en

[1] http://sourceforge.net/projects/ahotts/
[2] http://sourceforge.net/projects/cotovia/
[3] http://festcat.talp.cat/download.php
[4] TVSocial ETORGAI Televisión Social - Low Cost Telebista (ER-2010/00003)

evaluaciones competitivas tanto a nivel nacional como internacional. En la campaña Albayzin 2010 obtuvo el primer puesto (Sainz et al., 2010) y en la edición de 2012 obtuvo la mejor valoración para síntesis de voz neutra (Sainz et al., 2012a). En la campaña Blizzard Challenge 2011 obtuvo el 5º puesto a nivel mundial. (Sainz et al, 2011). Recientemente obtuvo los mejores resultados en varias categorías del Hurricane Challenge (Erro et al., 2013), concretamente en aquellas en las que se evaluaba la inteligibilidad de la voz sintética en condiciones ruidosas extremas.

3 AhoTTS Multilingüe

Para añadir nuevos idiomas a los ya disponibles se han integrado los módulos lingüísticos de código abierto diseñados por otras universidades. Aunque habría sido posible desarrollar un procesador lingüístico único para todos los idiomas esta posibilidad fue descartada debido a la dificultad de su implementación. Así, el procesador lingüístico del catalán se ha tomado de Festcat (Bonafonte et al., 2009), el del gallego de Cotovía (Rodríguez et al., 2012) y el del inglés de Festival (Taylor y Black, 1998). Para la síntesis, el motor que se emplea en todos los casos es el estadístico basado en HTS (Zen et, al., 2007). Debido a la manera peculiar en que cada uno de estos analizadores lingüísticos extrae la información y a la dependencia que de ella tienen las correspondientes voces sintéticas, ha sido necesario desarrollar nuevas voces para los idiomas recién incorporados.

Un diagrama general del sistema AhoTTS multilingüe diseñado puede verse en la Figura 2, en la que se observan los distintos módulos lingüísticos integrados para cada lengua.

El código fuente del sistema CTV multilingüe, junto con las voces disponibles, puede encontrarse en el siguiente repositorio de SourceForge http://sourceforge.net/projects/ahottsmultiling/.

3.1 Integración del Catalán

Para el idioma catalán se ha optado por el sistema Festcat. Festcat es el sistema CTV desarrollado por el grupo TALP de la UPC para la lengua catalana. Está desarrollado en base a Festival y sus diferentes módulos están escritos en el lenguaje de programación Lisp.

En AhoTTS multilingüe se emplean los módulos de procesado lingüístico de Festcat para obtener las etiquetas lingüísticas contextuales correspondientes al texto de entrada. Para poder emplear dichos módulos sin la necesidad de instalar previamente Festival, en el repositorio de AhoTTS multilingüe se suministran estas librerías ya compiladas para máquinas Linux de 64bits. El uso en otros sistemas operativos requerirá la compilación de las librerías adecuadas.

Para las llamadas a Festcat y evaluación de comandos Lisp dentro de AhoTTS se emplea la API C/C++ de la universidad de Edimburgo. Una vez obtenidas las etiquetas contextuales se

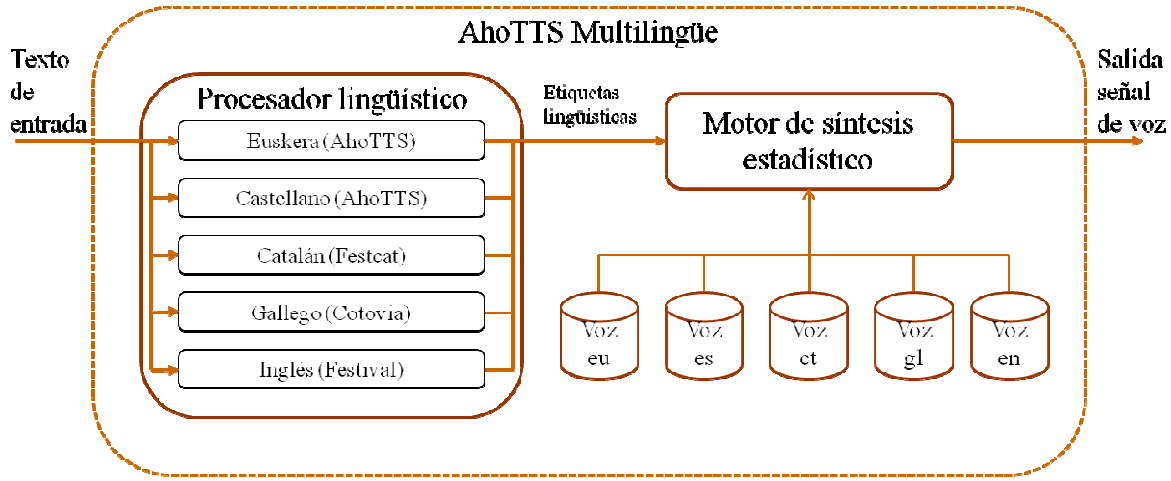

Figura 2: Diagrama general del AhoTTS Multilingüe

modifica el formato de las mismas para adecuarlo a la entrada del motor de síntesis. Después se introducen directamente en este módulo de igual manera que las generadas por los módulos lingüísticos propios de AhoTTS.

3.2 Integración del Inglés

En este caso se han incluido los módulos de procesado lingüístico de inglés proporcionados por la universidad de Edimburgo (Taylor y Black, 1998). De este modo se ha aprovechado el trabajo realizado para el catalán añadiendo fácilmente el inglés. Al igual que en el catalán es necesario adecuar el formato de las etiquetas lingüísticas al requerido por el motor de síntesis.

Es interesante destacar que cualquier módulo de procesado lingüístico desarrollado para Festival en otro idioma puede integrarse de manera sencilla en AhoTTS Multilingüe.

3.3 Integración del Gallego

El sistema escogido para la integración del gallego es Cotovia del grupo GTM de la UVIGO. Como el código del CTV está escrito en C/C++, simplemente se han integrado las funciones correspondientes al procesado lingüístico y generación de las correspondientes etiquetas en AhoTTS Multilingüe.

La salida que genera esté módulo sigue el formato ECESS (Pérez et al., 2006) basado en XML. Por tanto se ha adecuado dicha salida al formato de etiquetas de entrada del motor de síntesis.

3.4 Entrenamiento de Nuevas Voces

Previamente a la integración de los nuevos idiomas (el catalán, el gallego y el inglés) en AhoTTS, ya se disponía de modelos estadísticos para voces femenina y masculina en castellano y euskera generadas usando las bases de datos AhoSyn (Sainz et al., 2012b). También se disponía de una voz femenina para inglés entrenada a partir de la base de datos CMU ARCTIC (Kominek y Black, 2008). Por tanto, para el catalán y el gallego ha sido necesario desarrollar las correspondientes voces.

El entrenamiento de una nueva voz requiere un corpus fonéticamente balanceado y la transcripción de cada frase, así como las correspondientes grabaciones. A partir del audio se extraen los parámetros acústicos de la voz, y del texto se extraen las etiquetas lingüísticas correspondientes de forma que el sistema pueda aprender una relación estadística entre ambas. Las bases de datos para realizar dichos entrenamientos han sido cedidas por el grupo TALP de la UPC en el caso del catalán y por el grupo GTM de la UVIGO para el gallego.

Para entrenar voces para los nuevos idiomas se emplea HTS 2.2 (http://hts.sp.nitech.ac.jp/). Los datos principales sobre el tamaño de las bases de datos para la construcción de todas voces del sistema se resumen en la Tabla 1. En el caso de castellano y euskera el corpus para ambos géneros es el mismo en cada idioma.

Voz	Nº Frases	Nº Palabras	Duración aprox.
Voces castellano	3995	51380	6 horas
Voces euskera	3799	38544	6 horas
Inglés femenina	1132	10002	1 hora
Catalán femenina	3974	62314	6 horas
Catalán masculina	3692	58154	6 horas
Gallego masculina	1316	11235	1 hora

Tabla 1: Tamaño de las bases de datos usadas para la construcción de las voces

La parametrización acústica empleada consta de 39 coeficientes cepstrales en escala Mel junto con sus diferencias y sus segundas diferencias. También se extrae la frecuencia fundamental junto con su primera y segunda diferencia, así como un parámetro que indica el grado de sonoridad (más concretamente, la frecuencia máxima a la que la señal muestra armonicidad).

Como puede verse en la tabla 1, la cantidad de material disponible para el desarrollo de la voz gallega e inglesa es bastante inferior al disponible para las otras voces, a pesar de lo cual se ha considerado que es suficiente para obtener una calidad aceptable debido al método de síntesis empleado.

3.5 Transformación de Voces

Uno de los objetivos del proyecto en el que se enmarca el desarrollo de este sistema era la obtención de voces femeninas y masculinas

para todos los idiomas ya mencionados. Sin embargo, no fue posible obtener las bases de datos abiertas y libres de licencia en todos los casos. Por ello, y también por razones de economía de trabajo, se optó por aplicar técnicas de transformación de voces para completar el catálogo de las voces.

La transformación se ha llevado a cabo modificando los modelos estadísticos de la voz original a dos niveles: (i) se ha modificado el nivel medio de la frecuencia fundamental; (ii) se han aplicado técnicas de normalización del tracto vocal, lo que en el domino cepstral se traduce en un simple producto por una matriz como demuestran Pitz y Ney (2005). Para un valor adecuado de los parámetros de esta transformación, el resultado es una voz perceptualmente distinta a la original y que además mantiene un nivel de naturalidad comparable al de ésta.

La versión del sistema liberado dispone de las voces femeninas desarrolladas para los cinco idiomas.

3.6 API de desarrollo

En el repositorio de SourceForge junto con el código se proporciona además una API de desarrollo. Esta API permite incluir las funcionalidades básicas de AhoTTS multilingüe de manera sencilla en otros programas. También están incluidas aplicaciones de ejemplo para ilustrar el uso de esta API: un sistema autónomo y otro con arquitectura cliente/servidor.

De entre las características de las que dispone la API, las principales son:

- Permite cambiar la velocidad de lectura del texto en tiempo de síntesis.
- Proporciona las muestras de la salida para que el desarrollador las gestione de la manera que le convenga, ya sea guardándolas en un archivo de audio o enviándolas directamente a la tarjeta de sonido usando librerías del sistema operativo.
- Realiza el procesado frase a frase lo que permite el uso del sistema en aplicaciones en las que es necesario el procesado en tiempo real.

4 Conclusiones y Trabajos Futuros

Este artículo describe el sistema de conversión de texto a voz de código abierto desarrollado para las cuatro lenguas oficiales del estado más el inglés. El sistema del repositorio incluye además del código fuente del conversor multilingüe, las voces femeninas y una API de desarrollo para facilitar su integración en otras aplicaciones.

El hecho de que sea código abierto permite que cualquier persona interesada pueda descargarlo desde el repositorio donde se encuentra y utilizarlo para aprender, investigar o mejorarlo.

Este sistema multilingüe se ha desarrollado en el contexto del proyecto TV SOCIAL (http://tvsocial.ibercom.com/).

En el futuro se tiene pensado incluir como parte del procesado lingüístico un módulo previo de detección del idioma. De este modo se detectará automáticamente la lengua en la que está escrito el texto y se llamará directamente al módulo de procesado lingüístico correspondiente.

También se pretende crear una única voz multilingüe que incluya las particularidades fonéticas de todos los idiomas de manera que pueda usarse como voz única de todo el sistema.

5 Agradecimientos

Agradecemos al grupo TALP de la UPC y al grupo GTM de la UVIGO su ayuda y el material cedido para la creación de las voces catalanas y gallega respectivamente.

Queremos reconocer el trabajo de todas las personas que han colaborado en algún momento en el desarrollo de AhoTTS durante los últimos 20 años.

También agradecemos el trabajo realizado a todos los grupos que han liberado el código de sus sistemas CTV.

La migración del sistema a código abierto ha sido parcialmente financiada por el Gobierno Vasco (proyectos Ber2Tek, IE12-333 y Etorgai, ER-2010/00003), la empresa Eleka Ing. Ling. S.L. y por el Ministerio de Economía y Competitividad (Proyecto SpeechTech4All, TEC2012-38939-C03-03).

Bibliografía

Bonafonte, A., L. Aguilar, I. Esquerra, S. Oller, A. Moreno, 2009 "Recent Work on the FESTCAT Database for Speech Synthesis", Proc. SLTECH pp. 131-132.

Erro, D., I. Sainz, I. Luengo, I. Odriozola, J. Sánchez, I. Saratxaga, E. Navas, I. Hernáez, 2010, "HMM-based Speech Synthesis in Basque Language using HTS", Proc. FALA 2010 (VI Jornadas en Tecnología del Habla & II Iberian SLTech), pp. 67-70, (Vigo).

Erro, D., I. Sainz, E. Navas, I. Hernaez, 2011, "Improved HNM-based Vocoder for Statistical Synthesizers", Proc. Interspeech, pp. 1809-1812, (Florence).

Erro, D., T.C. Zorila, Y. Stylianou, E. Navas, I. Hernáez, 2013 "Statistical Synthesizer with Embedded Prosodic and Spectral Modifications to Generate Highly Intelligible Speech in Noise", Proc. Interspeech, (Lyon).

Hunt, A., A. Black, 1996 "Unit selection in a concatenative speech synthesis system using a large speech database", Proc. ICASSP, vol. 1, pp. 373–376.

Kominek, J., A Black, 2004 ”The CMU Arctic speech databases”, Proc. 5th ISCA Speech Synthesis Workshop, pp 223-224, Pittsburgh, PA.

Hernaez, I. 1995 “Conversión de texto a voz para el euskera basada en un sintetizador de formantes”, Tesis doctoral, UPV/EHU.

Ling, Z.H., L. Qin, H. Lu, Y. Gao, L.R. Dai, R.H. Wang, Y. Jiang, Z.W. Zhao, J.H. Yang, Y.J. Chen, G.P. Hu, 2007 "The USTC and iFlytek speech synthesis systems for Blizzard Challenge 2007", Proc. Blizzard Challenge Workshop, Aug .

Navas, E., I. Hernáez, J. Sánchez, 2002 "Basque Intonation Modelling For Text To Speech Conversion", Proc. 7th International Conference on Spoken Language Processing (ICSLP), pp. 2409-2412, Denver.

Navas, E., I. Hernáez, J. Sánchez, 2002 "Modelo de duración para conversión de texto a voz en euskera", Procesamiento del Lenguaje Natural, vol. 29, pp. 147-152.

Navas, E , 2003 “Modelado prosódico del euskera batua para conversión de texto a habla”, Tesis doctoral, UPV/EHU.

Pérez, J., A. Bonaforte, H.U. Hain, E. Keller, S. Breueur, J. Tian, 2006 “ECESS Inter-Module Interface Specification for Speech Synthesis”, Proceedings of LREC Conference.

Pitz, M., H. Ney, 2005 “Vocal tract normalization equals linear transformation in cepstral space”, IEEE Trans. Speech and Audio Process., vol. 13(5), pp. 930-944.

Rodríguez, E., C. García, F. Méndez, M. Gozález, C. Magariños, 2012 “Cotovía: an Open Source Text-to-Speech System for Galician and Spanish”, Proc. Iberspeech 2012 (VII Jornadas en Tecnología del Habla & III Iberian SLTech), pp. 308-315, (Madrid).

Rodríguez, M.A., J.G. Escalada, D. Torre, 1998 ”Conversor multilingüe para castellano, catalán, gallego y euskera”, Procesamiento del lenguaje natural, Revista nº 23 pp19-23.

Sainz, I., D. Erro, E. Navas, J. Adell, A. Bonafonte, 2011 "BUCEADOR Hybrid TTS for Blizzard Challenge 2011", Proc. Blizzard Challenge Workshop, (Torino).

Sainz, I., D. Erro, E. Navas, I. Hernáez, J. Sánchez, I. Saratxaga, I. Odriozola, I. Luengo, 2010 "Aholab Speech Synthesizers for Albayzin2010", Proc. FALA 2010 (VI Jornadas en Tecnología del Habla & II Iberian SLTech), pp. 343-347, (Vigo).

Sainz, I., D. Erro, E. Navas, I. Hernáez, 2011 "A Hybrid TTS Approach for Prosody and Acoustic Modules", Proc. Interspeech, pp. 333-336.

Sainz, I., D. Erro, E. Navas, I. Hernáez, J. Sánchez, I. Saratxaga, , 2012a "Aholab Speech Synthesizer for Albayzin 2012 Speech Synthesis Evaluation", Proc. Iberspeech 2012 (VII Jornadas en Tecnología del Habla & III Iberian SLTech), pp. 645-652, (Madrid).

Sainz, I., D. Erro, E. Navas, I. Hernáez, J. Sánchez, I. Saratxaga and I. Odriozola, 2012b “Versatile Speech Databases for High Quality Synthesis for Basque”, Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12), pp. 3308-3312.

Taylor, P., Black, A. and Caley, R, 1998 “The architecture of the Festival Speech Synthesis System”, Proc. 3rd ESCA Workshop on Speech Synthesis, pp. 147-151, Jenolan (Caves, Australia).

Zen, H., T Nose, J Yamagishi, S Sako, T Masuko, AW Black, K Tokuda, 2007 “The HMM-based speech synthesis system (HTS)

version 2.0", Proc. ISCA Workshop on Speech Synthesis (SSW6), pp. 294-299.

Zen, H., K. Tokuda, A. W. Black, 2009 "Statistical parametric speech synthesis", Speech Communication, Volume 51, Issue 11, pp. 1039-1064.

Análisis Automático del Contenido Textual

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 179-186 recibido 30-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Improving Subjectivity Detection using Unsupervised Subjectivity Word Sense Disambiguation

Mejoras en la Detección de Subjetividad usando Desambiguación Semántica del Sentido de las Palabras

Reynier Ortega
Adrian Fonseca
Center for Pattern Recognition
and Data Mining
Santiago de Cuba, Cuba

Yoan Gutiérrez
Andrés Montoyo
DLSI,University of Alicante
Alicante, Spain

Resumen: En este trabajo se presenta un método para la detección de subjetividad a nivel de oraciones basado en la desambiguación subjetiva del sentido de las palabras. Para ello se extiende un método de desambiguación semántica basado en agrupamiento de sentidos para determinar cuándo las palabras dentro de la oración están siendo utilizadas de forma subjetiva u objetiva. En nuestra propuesta se utilizan recursos semánticos anotados con valores de polaridad y emociones para determinar cuándo un sentido de una palabra puede ser considerado subjetivo u objetivo. Se presenta un estudio experimental sobre la detección de subjetividad en oraciones, en el cual se consideran las colecciones del corpus MPQA y Movie Review Dataset, así como los recursos semánticos SentiWordNet, Micro-WNOp y WordNet-Affect. Los resultados obtenidos muestran que nuestra propuesta contribuye de manera significativa en la detección de subjetividad.
Palabras clave: detección de subjetividad, desambiguación semántica, análisis de sentimiento

Abstract: In this work, we present a sentence-level subjectivity detection method, which relies on Subjectivity Word Sense Disambiguation (SWSD). We use an unsupervised sense clustering-based method for SWSD. In our method, semantic resources tagged with emotions and sentiment polarities are used to apply subjectivity detection, intervening Word Sense Disambiguation sub-tasks. Through an experimental study, we empirically validated the proposed method over two subjectivity collections, MPQA Corpus and Movie Review Dataset, using three widely popular opinion-mining resources SentiWordNet, WordNet-Affect and Micro-WNOp. The results show that our proposal performs significantly better than our proposed baseline.
Keywords: subjectivity detection, subjective word sense disambiguation, sentiment analysis

1 Introduction

Subjectivity detection consists in identifying whether a phrase, word or sentence is used to express opinion, emotion, evaluation, speculation, etc., (Wiebe and Riloff, 2005). It besides contributes in many Natural Language Processing (NLP) tasks. For instance, Information Retrieval systems incorporate subjectivity detection to provide opinionated and factual information, separately (Pang and Lee, 2008); and Question Answering systems increase their performances when using criteria for discrimination among types of factual versus opinionated questions (Lloret et al., 2011). On the other hand, Summarization systems pretends to resume factual and subjective content differently (Murray and Carenini, 2008).

Motivated by the usability and applicability of this task, some researchers have pro-

ISSN 1135-5948

posed methods for deal with Subjective Detection Resolution (SDR). Many approaches rely on lexicons[1] of words that may be used to express subjectivity. These approaches do not make distinction between different senses of a word, so terms included in such lexicons are treated as subjective regardless of their sense. Moreover, most subjectivity lexicons are compiled as keyword lists, rather than word meanings. However, many keywords have both subjective and objective senses, depending on the context where the corresponding word appears.

Recent approaches have proposed to profit from Word Sense Disambiguation (WSD) in subjective analysis. It could be either by adding semi-automatically subjectivity tags to annotated senses in WSD corpora, or training a supervised classifier to determine whether a word is being used in a subjective sense or not, without explicitly tagging senses. The WSD uses in this area have been necessaries to know the context meaning to provide a better efficiency SDR.

In this paper, we propose using subjectivity annotated resources to solve the SDR, unlike previous approaches, which depend heavily on manual or semiautomatic annotation for training supervised classifiers. We use an unsupervised strategy consisting in a coarse-grained clustering-based WSD method that differentiates objective, subjective and highly subjective uses of every word, and classify sentences as subjective or objective. Our method is able to integrate the affective usabilities of SentiWordNet "SWN" (Baccianella, Esuli, and Sebastiani, 2010), Micro-WNOp "WNOp" (Cerini and Gandini, 2007) and WordNet-Affect "WNA" (Valitutti, 2004) to resolve Subjective Analysis Task.

The paper is organized as follows. We review related works in Section 2. Section 3 is dedicated to describing our approach, whereas Section 4 contains the descriptions and results analysis of the conducted experiments. Finally, we present in Section 5 our conclusions and further works.

2 Related Work

Methods for subjectivity detection span a wide range of viewpoints. An early work proposed by Hatzivassiloglou and Wiebe (2000) examined the effects of adjective orientation and gradability on sentence subjectivity. Its goal has been to determine whether a given sentence is subjective or not, judging from the adjectives involved in current sentence. An attempt to classify subjective and objective sentences have been introduced in (Riloff and Wiebe, 2003), which explores syntactic pattern extraction using semi-supervised learning.

Other works have focused in annotating senses with emotion labels or polarity values. For instance, a WordNet (Miller, 1995) extension has been presented by (Valitutti, 2004), where every sense is annotated with one of the six basic emotion labels "anger", "happiness", "surprise", "digust", "sadness" and "fear". Esuli and Sebastiani (2006) determine the polarity of word senses in WordNet, distinguishing among positive, negative and objective. They manually annotate a seed set of positive/negative senses and by following the internal relations in WordNet expand a small set using a supervised approach. They extend their work (Baccianella, Esuli, and Sebastiani, 2010) by applying the PageRank algorithm for ranking the WordNet senses in terms of how strongly a sense possesses a given semantic property (e.g., positive or negative).

A large number of works have applied WSD in sentiment analysis for instance, Rentoumi et al. (2008) determine the polarity by disambiguating the words and then mapping the senses to models of positive and negative polarity. To compute these models and produce the mappings of senses, they adopt a graph-based method which takes into account contextual and sub-word information. Similarly to earlier work, Martín-Wanton et al. (2010a) exploits full word sense disambiguation for determining the correct sense of a word and assigning polarity using SentiWordNet and General Inquirer (Stone et al., 1966). Martín-Wanton et al. (2010b) study the behavior of SWN, WNA, and WNOp in polarity detection.

Recently, Akkaya, Wiebe, and Mihalcea (2009) introduced Subjectivity Word Sense Disambiguation (SWSD), which consists in automatically determining which word instances in a corpus are being used in subjective senses, and which are being used in objective senses. They use a supervised system for SWSD, and exploit the SWSD output

[1] Are a stock of words used in a particular profession, subject or domain.

to improve the performance of multiple contextual opinion analysis tasks. Akkaya et al. (2010) carried out a pilot study where a subjectivity sense-tagged dataset was created for eight SENSEVAL[2] words through MTurk[3], a web-based non-expert manual annotation interface.

These works have focused in creating new datasets for subjectivity contextual analysis by using existing polarity classification resources. They all rely heavily on manual or semiautomatic annotation for training supervised classifiers.

3 Our Proposal

As we mentioned previously, we use an unsupervised strategy consisting in a coarse-grained clustering-based WSD method that differentiates objective, subjective and highly subjective uses of every word.

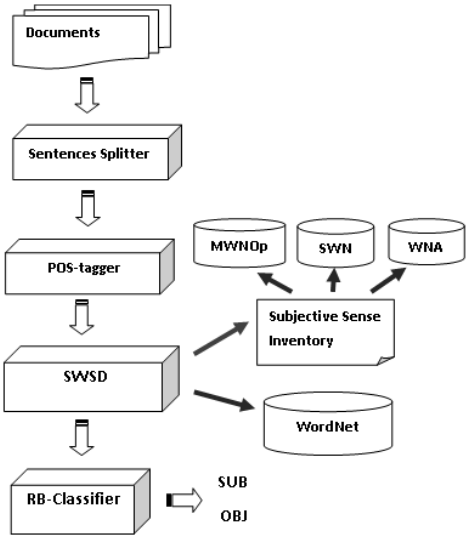

Figure 1: Overall architecture of contextual subjective classifier.

In this work, we evaluate the behavior of our proposal using different opinion mining resources. The overall architecture of our contextual subjective classifier is shown in Figure 1.

Firstly, the text is segmented into sentences, lemmatized and POS-tagged using TreeTagger tool (Schmid, 1994) being removed the stopwords. Then a Subjectivity Word Sense Disambiguation (SWSD) method is applied to content words (nouns, adjectives, verbs and adverbs). Once all content words are disambiguated (Section 3.2), we apply a rule-based classifier (Section 3.3) to decide whether the sentence is subjective or objective.

The use of SWSD in our proposal is motivated by considerations exposed in (Akkaya, Wiebe, and Mihalcea, 2009), where they explain that a same word may be used subjectively or objectively in different contexts. For example, the word *"earthquake"* is used in a subjective sense in the sentence:

"Selling the company caused an earthquake among the employees".

Whereas it is used in an objective sense in the sentence:

"An earthquake is the result of a sudden release of energy in the Earth's crust that creates seismic waves".

We adapted the unsupervised word sense disambiguation method proposed by Anaya-Sánchez, Pons-Porrata, and Berlanga-Llavori (2006) which is based on clustering as a manner of identifying related senses, for SWSD. Unlike the authors, who aim at obtaining the correct sense of a word; we use the method to determine when a word is subjective or objective relying on a subjective sense inventory. We constructed subjective senses inventories based on affective and polarity annotations in opinion mining resources.

3.1 Subjective Sense Inventories

Creating subjective sense-tagged data is a hard and expensive task. For this reason, we decided to use existent sense-level resources for fine-grained and coarse-grained subjective sense labeling. We considered three different resources for building our subjective sense inventory: SWN, WNOp and WNA. These resources have not explicit subjectivity labels; therefore we mapped polarity or affect labels to subjectivity labels.

SWN and WNOp contain positive, negative and objective scores between 0 and 1. In this case the mapping was defined in the following manner: senses whose sum of positive and negative scores is greater than or equal to 0.75 are considered to be highly subjective (HS); whereas those whose sum is lower than 0.75 and greater than or equal to 0.5 are con-

[2]http://www.senseval.org/
[3]http://mturk.amazon.com

sidered to be subjective (S). In the remaining cases, the senses are considered to be objective (O). In WNOp is important to clarify that this resource only contains 1105 WordNet senses annotate manually, the remainder WordNet senses were considered as objective.

In WNA, the senses are annotated with emotion labels. In order to match these labels with ours, we apply a similar strategy to (Balahur et al., 2009). Here, senses labeled with the following emotions: "anger", "disgust" and "surprise", are considered as highly subjectives. Others like "guilt", "sadnes" and "joy" are considered as subjectives; and the rest are considered as objectives. In Table 1, we show the distribution of the subjectivity labels assigned for each resource.

Resources	HS	S	O
SentiWordNet	1766	6429	107229
Micro-WNOp	216	118	115090
WordNet-Affect	110	148	115166

Table 1: Senses highly subjective (HS), subjective(S) and objective (O) distributions by resources.

For all three resources, exists a notable unbalance between the number of objective and subjective senses, which is particularly strong in the case of WNA.

Once tagged the sense with subjectivity label, these are grouped for building the coarse-grained sense. For instance, considering the following adjective, *"sad"*, using SentiWordNet, this adjective has three word senses in WordNet 2.0, from which we can obtain its lemma, part-of-speech, sense offset (id), definition and subjective label assigned.

i. sad#a#1 – *experiencing or showing sorrow or unhappiness* – (HS)

ii. sad#a#2 – *of things that make you feel sad* – (O)

iii. sad#a#3 – *bad; unfortunate* – (HS)

As we can see, the first and third senses are considered as highly subjective and second is considered as objective. These considerations were taken using the defined mapping above. For this reason sense 1 and 3 are merged in only one sense representing an highly subjective unique sense, keeping sense 2 as objective sense.

3.2 Adaptations introduced in the WSD

As we expressed, the selected disambiguation method was developed for the traditional WSD task. In this WSD method, the senses are represented as topic signatures (Lin and Hovy, 2000) built from WordNet concept repositories. The disambiguation process starts from a clustering distribution of all possible senses of the ambiguous words by applying the Extended Star clustering algorithm (Gil-García, Badía-Contelles, and Pons-Porrata, 2003). Such clustering tries to identify cohesive groups of word senses, which are assumed to represent different meanings for the set of words. Then, clusters that best match with the context are selected. If the selected clusters disambiguate all words, the process stops and the senses belonging to the selected clusters are interpreted as the disambiguating ones. Otherwise, the clustering is performed again (regarding the remaining senses) until a complete disambiguation is achieved.

Thus, it does not distinguish between highly subjective, subjective and objective senses. We propose two strategies to adapt this method for the task at hand. The first strategy is based in fine-grained WSD. It consists in applying the original WSD method (Anaya-Sánchez, Pons-Porrata, and Berlanga-Llavori, 2006) and searching the subjective sense inventory for subjectivity labels for each fine-grained sense.

The second strategy is based on coarse-grained WSD. Many authors (Chan and Ng, 2007; Navigli, Litkowski, and Hargraves, 2007) have demonstrated that WSD methods increase their performance by using coarse-grained senses. In this paper, we defined new coarse-grained senses. All highly subjective senses of a word are collapsed into a single sense, as well as all subjective senses. On the other hand, objective senses are kept separated.

For our selected WSD method, word senses are represented by means of topic signatures (Lin and Hovy, 2000). The topic signature for coarse-grained senses is the sum of the topic signature of the corresponding original fine-grained senses. To take again the example in section 3.1 referring the adjective *"sad"*, it represents an instance of coarse topic signatures, where first and third senses, were grouped in a coarse-grained

sense (highly subjective HS). The signature for this new sense is obtained in following manner:

$$Tsign(sad\#a\#HS) = Tsign(sad\#a\#1) + Tsign(sad\#a\#3)$$
$$Tsign(sad\#a\#O) = Tsign(sad\#a\#2)$$

Where $Tsign(sense_i)$ compute the related topic signature with the $sense_i$.

3.3 Subjective Sentence Classifier

We use a rule-based classifier to classify sentences into subjective or objective. A voting scheme is used. Every word disambiguated as highly subjective has assigned a score of 4 and every word disambiguated as subjective has assigned a score of 2. If the sum of all scores is greater than a threshold, the sentence is classified as subjective. This method is similar to that proposed by (Riloff, Wiebe, and Wilson, 2003). Equation 1 is used to classify a new sentence:

$$RL(f) = \begin{cases} subjective & if \quad \sum_{i=1}^{n} Score(w_i) \geq \lambda \\ objective & e.o.c \end{cases} \tag{1}$$

Where:

$$Score(w_i) = \begin{cases} 4.0 & if \quad w_i \quad is \quad high \quad subjective \\ 2.0 & if \quad w_i \quad subjective \\ 0.0 & if \quad w_i \quad objective \end{cases} \tag{2}$$

In both equations (1, 2) w_i is the sense which the word is using in the sentence f. In our proposal the threshold used was $\lambda = 4.0$. This value was estimated using empirical evaluation over a subset of the SemCor corpus for English, being it automatically annotated with OpinionFinder tool by (Carmen Banea and Hassan, 2008). Thus, we employ this rule-based classifier with the aim to obtain an unsupervised method to classify sentences in subjective or objective categories.

4 Result and Discussion

We conducted a series of experiments in order to evaluate the validity of our proposal. The aim has been focused on the impact of using different resources for constructing the sense inventories to solve the SWSD.

In our experiments, we use two collections of subjectivity detection: the manually annotated MPQA Corpus (Wilson, 2005) and the automatically annotated collection over movie domain, Movie Review Dataset (Pang and Lee, 2004).

MPQA Corpus contains news (for version 1.2 contains 11115 sentences) where opinions are spreaded at sentence level. They are annotated with sentiment polarities and its respective strength value. In order to experimenting our SWSD method, we used the approach presented by Riloff and Wiebe (2003), obtaining 8026 subjective and 3089 objective sentences respectively.

Movie Review Dataset covers the movie domain. It contains 5000 subjective sentences extracted from movie reviews collected from the Rotten Tomatoes web site, and 5000 objective sentences collected from movie plot summaries from the Internet Movie Database (IMDB). The underlying assumption is that all the snippets from the Rotten Tomatoes pages are subjective (as they come from a review site), while all sentences from IMDB are objective (as they focus on movie plot descriptions).

In order to constructing a baseline which to evaluate the effect of applying SWSD on each resource, we use the same classification scheme, but without applying word sense disambiguation.

The polarity score of the words were defined as the average of the positive and negative score sum, for all associated senses to each word. All words with a new score above 0.75 were tagged as highly subjective, the words with score in the range 0.5 and 0.75 were tagged as subjective, and the rest were tagged as objective.

As score measures we computed precision, recall and F1 for both subjective and objective classes, moreover we compute the average of the F1 over subjective and objective sentences.

The behavior of all variants are shown in Tables 2 and 3. In these tables we can observe that except for one case, the classification is improved when is used both forms of SWSD, respect to the baseline ([resource] without WSD). The improvement is higher than 25% in both collections when we use WNOp (see Tables 2 and 3). This fact confirms our prior hypothesis that taking into account the individual subjectivity levels of different senses

Strategy	Ps	Po	Rs	Ro	Fs	Fo	F1-avg
SWN with Fine WSD	0.9305	0.270	0.5677	0.7903	0.7052	0.4225	**0.5538**
SWN with Coarse WSD	**0.9322**	0.2861	0.6052	0.7823	0.7339	0.4190	**0.5765**
SWN without WSD	0.8996	0.3309	0.8483	0.3710	0.8558	0.3498	**0.6043**
WNOp with Fine WSD	0.**9305**	0.2700	0.5677	0.7903	0.7052	0.4025	0.**5538**
WNOp with Coarse WSD	0.9237	0.2533	0.5334	0.7823	0.6763	0.3826	**0.5295**
WNOp without WSD	0.9302	0.1744	0.06525	0.9758	0.1220	0.2958	0.2089
WNA with Fine WSD	1.0	0.1710	0.0196	1.0	0.0384	0.2921	**0.1653**
WNA with Coarse WSD	1.0	0.1710	0.0196	1.0	0.0384	0.2921	**0.1653**
WNA without WSD	0.9091	0.1694	0.0163	0.9919	0.0321	0.2894	0.1607

Table 2: Experimental evaluation using MPQA Corpus.

Strategy	Ps	Po	Rs	Ro	Fs	Fo	F1-avg
SWN with Fine WSD	0.6303	0.5204	0.7745	0.35	0.6950	0.4185	**0.5568**
SWN with Coarse WSD	0.6203	0.4988	0.7734	0.3226	0.6884	0.3918	**0.5401**
SWN without WSD	0.6066	0.5268	0.89064	0.1742	0.7218	0.2618	0.4918
WNOp with Fine WSD	**0.6303**	0.5204	0.7745	0.35	0.6950	0.4185	**0.5568**
WNOp with Coarse WSD	0.6267	0.5011	0.7475	0.3629	0.6818	0.4210	**0.5514**
WNOp without WSD	0.5044	0.4046	0.0643	0.9097	0.1140	0.9097	0.5118
WNA with Fine WSD	0.7436	0.4245	0.0.981	0.9516	0.1733	0.5871	**0.3802**
WNA with Coarse WSD	**0.7479**	0.4251	0.1003	0.9516	0.1769	0.5876	**0.3821**
WNA without WSD	0.6724	0.4148	0.04397	0.9694	0.0825	0.5810	0.3317

Table 3: Experimental evaluation using Movie Review Dataset.

of a word, it may helping in SDR. Surprisingly, very small differences are observed between fine-grained and coarse-grained SWSD variants. We suppose that this situation is due to the high correlation among sense clusters obtained by the WSD method and those manually defined for the coarse-grained variant.

On the other hand the results using SWN are higher than the rest, whereas are lower those obtained using WNA. We may observe that as the unbalance between subjective and objective senses is higher, less accuracy have the obtained results with this resource. In case of WNA, we should note additionally that the mapping established between affect categories and subjectivity labels does not reflect all circumstances under which words are used subjectively.

In case of WNOp, despite being significantly smaller than SWN and suffering an unbalance between objective and subjective senses, when it is used for SWSD, the obtained results are similar to the obtained when SWN is used. This fact is encouraging, as it suggests that a small resource with a high-quality of annotated data is able to perform at the same level than annotated resources much bigger. A further exploration could be required to determine if the growing of WNOp may resulting in an improving of the performance of our proposal.

5 Conclusion and Further Works

In this work, we have presented an unsupervised SWSD-based approach to subjectivity detection, which relies on sense-level polarity and emotion-labeled resources. We conduct an experimental study, where the behavior of our proposed method is evaluated using three widely used resources: SWN, WNOp and WNA. As a result of our experiments, we show that subjectivity detection using our unsupervised SWSD-based approach outperforms a baseline where disambiguation techniques are not used. Besides, we obtain a characterization of the method's behavior using different resources, and remarking that SWN and WNOp are the most suitable for the task.

On the other hand, in order to find out other ways to obtain semantic labels of coarse-grained, we will adapt our method to the use (Gutiérrez, Vázquez, and Montoyo, 2011) proposal, which is able to obtain relevant domains associated to the sentences,

where these domains involve polarity values.

Another attractive direction for future works is determining the influence of subjectivity-annotated resources, rather than approximating a subjectivity annotation from existing polarity or affect annotations.

Acknowledgments

This research work has been partially funded by the Spanish Government through the project TEXT-MESS 2.0 (TIN2009-13391-C04), "Análisis de Tendencias Mediante Técnicas de Opinión Semántica" (TIN2012-38536-C03-03), and SAM - Dynamic Social & Media Content Syndication for 2nd Screen (FP7-611312); and by the Valencian Government through the project PROMETEO (PROMETEO/2009/199).

References

Akkaya, Cem, Alexander Conrad, Janyce Wiebe, and Rada Mihalcea. 2010. Amazon mechanical turk for subjectivity word sense disambiguation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT 10, pages 195–203, Stroudsburg, PA, USA. Association for Computational Linguistics.

Akkaya, Cem, Janyce Wiebe, and Rada Mihalcea. 2009. Subjectivity word sense disambiguation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 190–199, Stroudsburg, PA, USA. Association for Computational Linguistics.

Anaya-Sánchez, Henry, Aurora Pons-Porrata, and Rafael Berlanga-Llavori. 2006. Word sense disambiguation based on word sense clustering. In *Proceedings of the 10th Ibero-American Conference on AI 18th Brazilian conference on Advances in Artificial Intelligence*, IBERAMIA-SBIA '06, pages 472–481, Berlin, Heidelberg. Springer-Verlag.

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, LREC '10, Valletta, Malta, may.

Balahur, Alexandra, Ralf Steinberger, Erik van der Goot, Bruno Pouliquen, and Mijail Kabadjov. 2009. Opinion mining on newspaper quotations. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 03*, WI-IAT '09, pages 523–526, Washington, DC, USA. IEEE Computer Society.

Carmen Banea, Rada Mihalcea, Janyce Wiebe and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii.

Cerini, S., Compagnoni V. Demontis A. Formentelli M. and G Gandini. 2007. Language resources and linguistic theory: Typology, second language acquisition, english linguistics (forthcoming), chapter micro-wnop: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining.

Chan, Yee Seng and Hwee Tou Ng. 2007. Word sense disambiguation improves statistical machine translation. In *In 45th Annual Meeting of the Association for Computational Linguistics*, ACL '07, pages 33–40.

Esuli, Andrea and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation*, LREC '06, pages 417–422.

Gil-García, R. J., J. M. Badía-Contelles, and A. Pons-Porrata. 2003. Extended star clustering algorithm. *Lecture Notes on Computer Sciences*, 2905:480–487.

Gutiérrez, Yoan, Sonia Vázquez, and Andrés Montoyo. 2011. Sentiment classification using semantic features extracted from wordnet-based resources. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment*

Analysis (WASSA 2.011), pages 139–145, Portland, Oregon, June. Association for Computational Linguistics.

Hatzivassiloglou, Vasileios and Janyce M. Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. pages 299–305.

Lin, Chin-Yew and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics - Volume 1*, COLING '00, pages 495–501, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lloret, Elena, Alexandra Balahur, Manuel Palomar, and Andrés Montoyo. 2011. Towards a unified approach for opinion question answering and summarization. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, pages 168–174, Stroudsburg, PA, USA. Association for Computational Linguistics.

Martín-Wanton, Tamara, Aurora Pons-Porrata, Andrés Montoyo-Guijarro, and Alexandra Balahur. 2010a. Opinion polarity detection - using word sense disambiguation to determine the polarity of opinions. In *ICAART (1)*, pages 483–486.

Martín-Wanton, Tamara, Aurora Pons-Porrata, Andrés Montoyo-Guijarro, and Alexandra Balahur. 2010b. Word sense disambiguation in opinion mining: Pros and cons. In *Research in Computing Science*, pages 119 – 129.

Miller, George A. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41.

Murray, Gabriel and Giuseppe Carenini. 2008. Summarizing spoken and written conversations. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, EMNLP '09, pages 773–782.

Navigli, Roberto, Kenneth C. Litkowski, and Orin Hargraves. 2007. Semeval-2007 task 07: coarse-grained english all-words task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 30–35, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pang, Bo and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*.

Pang, Bo and Lillian Lee. 2008. Opinion mining and sentiment analysis.

Rentoumi, Vassiliki, George Giannakopoulos, Vangelis Karkaletsis, and George A. Vouros. 2008. Sentiment analysis of figurative language using a word sense disambiguation approach.

Riloff, Ellen and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 105–112.

Riloff, Ellen, Janyce Wiebe, and Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. pages 25–32.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees.

Stone, Philip J., Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA.

Valitutti, Ro. 2004. Wordnet-affect: an affective extension of wordnet. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086.

Wiebe, Janyce and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *In Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics*, CICLing '05, Mexico City, Mexico.

Wilson, Theresa. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *In Proceedings of HLT-EMNLP*, pages 347–354.

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 187-194 recibido 30-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Una Nueva Técnica de Construcción de Grafos Semánticos para la Desambiguación Bilingüe del Sentido de las Palabras *

A New Technique for Cross Lingual Word Sense Disambiguation based on Building Semantic Graphs

Andres Duque Fernandez[1], **Lourdes Araujo**[2], **Juan Martinez-Romo**[3]
NLP & IR Group
Universidad Nacional de Educación a Distancia (UNED)
28040 Madrid, Spain
{[1]aduque, [2]lurdes, [3]juaner}@lsi.uned.es

Resumen: En este trabajo presentamos unos resultados preliminares obtenidos mediante la aplicación de una nueva técnica de construcción de grafos semánticos a la tarea de desambiguación del sentido de las palabras en un entorno multilingüe. Gracias al uso de esta técnica no supervisada, inducimos los sentidos asociados a las traducciones de la palabra ambigua considerada en la lengua destino. Utilizamos las traducciones de las palabras del contexto de la palabra ambigua en la lengua origen para seleccionar el sentido más probable de la traducción. El sistema ha sido evaluado sobre la colección de datos de una tarea de desambiguación multilingüe que se propuso en la competición SemEval-2010, consiguiendo superar los resultados de todos los sistemas no supervisados que participaron en aquella tarea.
Palabras clave: Desambiguación lingüística, bilingüismo, métodos basados en grafos

Abstract: In this paper we present preliminary results obtained by the application of a new technique for building semantic graphs to the task of cross-lingual word sense disambiguation. Through the use of this unsupervised technique, we induce the senses associated with the translations of the ambiguous word in the target language. For this purpose, we use the translation of the words in the context of the ambiguous word in the source language to select the most likely sense. The system has been evaluated on a dataset from a cross-lingual word sense disambiguation task proposed in the SemEval-2010 competition, outperforming all unsupervised systems participating in that task.
Keywords: Cross-lingual, word-sense disambiguation, graph-based approaches

1. Introducción

Una gran parte de las palabras de un idioma son polisemas, es decir, tienen más de un significado y se interpretan de distintas formas según el uso que se hace de ellas. La Desambiguación del Sentido de las Palabras o Word Sense Disambiguation (WSD) se ha convertido por ello en uno de los problemas del Procesamiento del Lenguaje Natural (PLN) que ha atraído más atención, ya que también es un paso necesario para numerosos procesos de PLN (Ide y Veronis, 1998): traducción automática, recuperación de información, clasificación de textos, etc.

El problema de la desambiguación del sentido de las palabras se ha tratado frecuentemente como un problema de aprendizaje supervisado (Màrquez et al., 2006; Mihalcea, 2006). Sin embargo, este tipo de métodos requiere disponer de textos etiquetados semánticamente. Estos recursos son muy costosos y de hecho muy escasos. Por ello ha surgido un interés creciente en abordar el problema de forma no supervisada. Estos trabajos, que no requieren textos anotados semánticamente, se centran en la denominada *Inducción del sentido de las palabras* (ISP). El objetivo es distinguir los distintos usos o sentidos de una palabra determinada en un texto dado, pero sin clasificar dichos sentidos en base a un inventario de sentidos preexistente. La distinción se hace en base a grupos de palabras que presentan alguna relación acentuada con un sentido particular. Generalmente esta relación es la coaparición con la palabra considerada en los contextos observados en los textos. Esta es una característica muy interesante, como apunta Pedersen (2006), ya que por una parte no existe un inventario de sentidos para todas las palabras, y por otra parte, incluso cuando existe este inventario, la naturaleza y el grado de distinción de los sentidos que

* Trabajo financiado parcialmente por los proyectos Holopedia (TIN2010-21128-C02-01) y MA2VICMR (S2009/TIC-1542)

ISSN 1135-5948

nos interesa varía con las aplicaciones.

De forma muy genérica los métodos aplicados en ISP se pueden clasificar en dos categorías principales, los basados en vectores y los basados en grafos. Estos últimos (Veronis, 2004; Agirre et al., 2006; Agirre y Soroa, 2007; Klapaftis y Manandhar, 2008) son los más directamente relacionados con el trabajo que presentamos. Generalmente estos enfoques representan como un vértice a cada palabra que coaparece con la palabra objetivo (palabra ambigua que se quiere traducir) dentro de una ventana predefinida. Dos vértices están conectados por una arista si coaparecen en uno o más contextos de la palabra objetivo. Una vez que se ha construido el grafo para la palabra objetivo se aplican distintos algoritmos de detección de comunidades para inducir los sentidos. Cada comunidad de palabras se toma como uno de los sentidos inducidos.

Otro enfoque que también se ha investigado para tratar la desambiguación es la utilización de textos paralelos. Como ha señalado Resnik (2004), no sólo para este problema, sino en general para el procesamiento del lenguaje, el significado oculto que comparten las traducciones paralelas permite inferir conocimiento de una lengua a partir de otra en la que se disponga de más recursos.

En este trabajo proponemos inducir los sentidos de las palabras ambiguas en distintos idiomas mediante un nuevo algoritmo de construcción de grafos y aplicarlos a la desambiguación semántica de textos.

La hipótesis de partida del algoritmo de construcción de grafos que utilizamos (Martinez-Romo et al., 2011) es que un documento tiene un contenido coherente, por lo que tiene sentido hacer el supuesto básico de que todas las palabras que aparecen en el mismo documento tienden a compartir un sentido común. El objetivo es enlazar cada par de palabras que compartan un sentido común, condición que aproximamos por la coaparición de ambas palabras en un mismo documento. Sin embargo, esto no es siempre cierto. Algunas palabras pueden aparecer en un documento sin tener realmente relación con el sentido general de dicho documento. Por lo tanto, se considera que dos palabras realmente comparten un sentido común si coaparecen *frecuentemente* en los mismos documentos.

Concretamente, generaremos un grafo en el idioma de destino (español) del que extraemos comunidades de palabras que representarán los sentidos inducidos. Un diccionario de traducciones entre lenguas nos permite identificar las comunidades asociadas a las traducciones que están más relacionadas en las lenguas consideradas. En concreto hemos utilizado un diccionario (López-Ostenero, 2002) bilingüe español-inglés creado en el Grupo de Procesamiento de Lenguaje Natural de la UNED (http://nlp.uned.es). Considerando textos alineados, esta identificación nos permitirá seleccionar el sentido más acertado de las palabras ambiguas de los textos buscando un sentido común entre las traducciones de los idiomas considerados.

El resto del artículo se organiza de la siguiente forma: en la sección 2 se citan los principales trabajos dentro del área de la desambiguación multilingüe del sentido de las palabras. En la sección 3 se describirán cada una de las etapas que componen la metodología seguida por el algoritmo propuesto. En la sección 4 se muestran los principales resultados. Finalmente, en la sección 5 se extraen una serie de conclusiones y se expone la línea de trabajo futuro.

2. *Estado del Arte*

Ha habido algunas propuestas de explotar los corpora paralelos para tratar la desambiguación del sentido de las palabras. Resnik y Yarowsky (1999) presentaron uno de los primeros análisis de la potencialidad de los recursos multilingües para la desambiguación, proponiendo un marco de evaluación y una medida de distancia multilingue entre sentidos. Diab y Resnik (2002) propusieron un método para anotar automáticamente el sentido de las palabras en grandes corpora paralelos. Este método requiere disponer de un inventario de sentidos para una de las lenguas y se apoya en el alineamiento a nivel de palabra para identificar las traducciones de palabras entre las lenguas consideradas. Ng, Wang, y Chan (2003) propusieron un método para adquirir datos de entrenamiento de desambiguación de sentidos basado en la selección manual de traducciones entre corpora paralelos que se utilizaban para entrenar un clasificador. Banea y Mihalcea (2011) también proponen un método supervisado, que en este caso utiliza rasgos multilingües para entrenar un clasificador. Los rasgos se obtienen traduciendo el contexto de las palabras ambiguas a varias lenguas. Fernandez-Ordonez, Mihalcea, y Hassan (2012) hacen una propuesta no supervisada suponiendo que la única información disponible es un diccionario con las definiciones de los distintos significados de las palabras ambiguas. En esta propuesta se utiliza una variante del algoritmo Lesk, que dada una secuencia de palabras intenta identificar las combinaciones de sentidos

que maximizan el solapamiento entre las definiciones correspondientes.

En relación a la evaluación, en la edición de 2010 de la campaña de SemEval en la que se proponen competiciones de sistemas sobre tareas relacionadas con la desambiguación del sentido de las palabras (WSD), se incluyó una dedicada a la desambiguación en un contexto multilingüe (Lefever y Hoste, 2010). En dicha tarea los participantes debían determinar automáticamente la traducción apropiada para el contexto de un nombre inglés dados cinco idiomas: Holandés, Alemán, Italiano, Español y Francés. Para la compilación de la colección, fueron empleados dos tipos de datos: un corpus paralelo extraído de Europarl (http://www.statmt.org/europarl/) sobre el que se construyó el "gold standard" y una colección de frases en inglés que contienen las palabras de la muestra léxica anotadas con sus correspondientes traducciones en cinco idiomas. En el idioma *español*, que es en el que nos centraremos en este trabajo, participaron cuatro sistemas. En cuanto a los sistemas supervisados, los sistemas UvT-WSD (van Gompel, 2010) y FCC (Vilariño et al., 2010) emplearon un clasificador basado en el algoritmo K-Nearest y Naive Bayes respectivamente. Los sistemas no supervisados utilizaron algoritmos basados en grafos con ciertas diferencias respecto al algoritmo presentado en este trabajo. El sistema T3-COLEUR (Guo y Diab, 2010) hizo uso de tablas de probabilidad de traducción bilingües que se derivan a partir del corpus Europarl. Por su parte el sistema UHD (Silberer y Ponzetto, 2010) construye para cada palabra objetivo un grafo de coapariciones multilingüe basado en los contextos alineados de la palabra objetivo, disponibles en los corpora paralelos. Los términos con una traducción cruzada por cada idioma se unen por un enlace específico de "traducción" entre los diferentes grafos correspondientes a cada idioma. Finalmente el grafo se transforma en un árbol de expansión mínimo y es utilizado para seleccionar las palabras más relevantes en el contexto y desambiguar cada instancia de evaluación.

En este trabajo también utilizaremos estos datos para su evaluación y aplicaremos una nueva técnica de construcción de grafos semánticos que se describe a continuación.

3. Descripción del Algoritmo

En esta sección se describirán cada una de las etapas que componen la metodología seguida para la aplicación del algoritmo basado en grafos de coaparición a una tarea de desambiguación semántica multilingüe.

3.1. Preprocesado del Corpus

El algoritmo basado en grafos de coaparición que se expone en el presente artículo es un algoritmo no supervisado, que utiliza el conocimiento extraído de los documentos completos que contienen alguna instancia de la palabra a desambiguar. En este caso, se ha utilizado el corpus paralelo multilingüe Europarl (Koehn, 2005), el cuál se extrae de las actas del Parlamento Europeo correspondientes a los años comprendidos entre 1996 y 2011, y se encuentra alineado a nivel de frase. Los idiomas utilizados para realizar la desambiguación han sido el inglés como idioma origen, y el español como idioma destino.

El corpus inicial está separado en documentos que representan las actas del Parlamento. Cada uno de los documentos contiene diferentes etiquetas que indican las distintas partes del mismo. Tras analizar los documentos que aparecen en el corpus utilizado, se decidió separar los documentos mediante la etiqueta "<SPEAKER>" que separan cada una de las intervenciones acontecidas a lo largo de la sesión. De esta forma, cada uno de los documentos que el algoritmo tiene en cuenta para construir un grafo de coaparición, representa la intervención de un único miembro del Parlamento Europeo en una sesión concreta y por lo tanto respeta nuestra hipótesis de que todas las palabras que aparecen en el mismo documento tienden a compartir un sentido común.

Una vez que se ha realizado la separación de los documentos, es necesario realizar un etiquetado de los mismos para clasificar las palabras que en ellos aparecen, en función de su categoría gramatical. Esta tarea se realizó de forma automática mediante la herramienta TreeTagger (Schmid, 1994), para los dos idiomas utilizados. Una vez realizado este etiquetado, disponemos de un conjunto de documentos etiquetados gramaticalmente, alineados a nivel de frase, en los dos idiomas que se van a utilizar para esta tarea.

3.2. Construcción del grafo semántico

El siguiente paso consiste en la construcción del grafo de coaparición de palabras en la lengua destino (español), a partir de los documentos etiquetados. Para ello se han utilizado grafos dedicados, es decir, para construir el grafo de palabras en español, sólo se han tenido en cuenta aquellos documentos que contienen al menos una de las posibles traducciones de la palabra ambigua. Estas posibles traducciones se obtienen del diccionario de referencia (López-Ostenero, 2002).

Para comprobar si la coaparición de dos palabras en un documento es significativa, se define un modelo nulo en el que las palabras se distribuyen aleatoria e independientemente entre un conjunto de documentos de un corpus. Concretamente, se calcula la probabilidad de que dos palabras coincidan por puro azar. Este valor nos permite determinar un p-valor p para la coaparición de dos palabras. Si $p \ll 1$ se puede considerar que la aparición de las dos palabras en el mismo documento es significativa, y por lo tanto, es probable que su significado esté relacionado.

Concretamente, si dos palabras que se encuentran respectivamente en dos documentos n_1 y n_2 de entre los N que componen el corpus, para contar cuantos casos existen en los que dos palabras coincidan en exactamente k documentos, debemos tener en cuenta que hay cuatro tipos de documentos: k documentos que contienen ambas palabras, $n_1 - k$ documentos que contienen sólo la primera palabra, $n_2 - k$ documentos que contiene sólo la segunda palabra, y $N - n_1 - n_2 + k$ documentos que no contienen ninguna de las dos palabras. Por lo tanto, el número de disposiciones que buscamos viene dado por el coeficiente multinomial:

$$\binom{N}{k}\binom{N-k}{n_1-k}\binom{N-n_1}{n_2-k} \tag{1}$$

Así, la probabilidad de que dos palabras que aparecen en los documentos n_1 y n_2 respectivamente y que están distribuidas de forma aleatoria e independiente entre N documentos, coincidan en exactamente k de ellos viene dada por:

$$p(k) = \frac{\binom{N}{k}\binom{N-k}{n_1-k}\binom{N-n_1}{n_2-k}}{\binom{N}{n_1}\binom{N}{n_2}} \tag{2}$$

si $\max\{0, n_1 + n_2 - N\} \leq k \leq \min\{n_1, n_2\}$ y cero en otro caso.

Podemos escribir la ecuación (2) de una forma más fácil de tratar computacionalmente. Para ello introducimos la notación $(a)_b \equiv a(a-1)\cdots(a-b+1)$, para cualquier $a \geq b$, y sin pérdida de generalidad suponemos que la primera palabra es la más frecuente, es decir $n_1 \geq n_2 \geq k$. Entonces:

$$\begin{aligned} p(k) &= \frac{(n_1)_k (n_2)_k (N-n_1)_{n_2-k}}{(N)_{n_2}(k)_k} \\ &= \frac{(n_1)_k (n_2)_k (N-n_1)_{n_2-k}}{(N)_{n_2-k}(N-n_2+k)_k (k)_k}, \end{aligned} \tag{3}$$

donde en la segunda forma se ha usado la identidad $(a)_b = (a)_c(a-c)_{b-c}$ valida para $a \geq b \geq c$. La ecuación (3) se puede reescribir como

$$\begin{aligned} p(k) = &\prod_{j=0}^{n_2-k-1}\left(1 - \frac{n_1}{N-j}\right) \\ &\times \prod_{j=0}^{k-1} \frac{(n_1-j)(n_2-j)}{(N-n_2+k-j)(k-j)}. \end{aligned} \tag{4}$$

Esto nos permite determinar un p-valor para la coaparición de dos palabras como

$$p = \sum_{k \geq r} p(k), \tag{5}$$

donde r es el número de documentos en el corpus en el que se han encontrado realmente las dos palabras juntas. Si $p \ll 1$ podemos considerar que la aparición de las dos palabras en el mismo documento es significativa, y por lo tanto es probable que su significado esté relacionado. Podemos cuantificar aún más esta significancia tomando la mediana (correspondiente a $p = 1/2$) como una referencia y calculando el peso de un enlace como $\ell = -\log(2p)$, es decir una medida de cuanto se desvía de la mediana el valor real de r.

Con esto podemos construir un grafo que tiene las palabras como nodos, y conectar con un enlace los pares de palabras que coinciden en al menos un documento y con un peso de coaparición por debajo de un valor umbral. El peso ℓ asignado a los enlaces mide la desviación de la coaparición de las dos palabras con respecto al caso nulo. Al grafo resultante le denominamos grafo de coaparición.

3.3. Detección de comunidades

Una vez que se ha definido el grafo de coaparición se puede pasar a agrupar los nodos, por ejemplo, aplicando una descomposición en comunidades. Las comunidades son subgrafos cuyos nodos presentan algún tipo de afinidad estructural o dinámica, y por lo tanto es plausible suponer que cada comunidad comparte un sentido común, diferente del de las restantes comunidades. Existen diversas propuestas de algoritmos de extracción de comunidades. En los experimentos realizados se ha utilizado *Walktrap*, presentado por Pons y Latapy (Pons y Latapy, 2005). Este algoritmo es de los que se apoyan en la idea de que un *camino aleatorio* o *random walk* queda atrapado más fácilmente en las partes del grafo densamente conectadas, que corresponden a

las comunidades. También utiliza heurísticas para fusionar las comunidades iterativamente, hasta conseguir un conjunto óptimo de comunidades. Dado que la optimización de este algoritmo de comunidades se escapa de los objetivos principales de nuestro trabajo, hemos decidido adoptar los valores por defecto de su implementación (http://www-rp.lip6.fr/ latapy/PP/walktrap.html) para obtener la descomposición óptima.

3.3.1. Grafo de Comunidades

Las comunidades que se han obtenido son posteriormente representadas en un nuevo grafo para poder calcular las distancias entre ellas. Para la construcción de este nuevo grafo GC, se recorrerá el grafo de coaparición de palabras GP, y se generará un enlace entre dos comunidades C_1 y C_2 siempre que una palabra $x \in C_1$ esté enlazada en el grafo de coaparición GP con una palabra $y \in C_2$.

3.4. Desambiguación de la Palabra Objetivo

A través de la construcción de los grafos de coaparición y la detección de comunidades, se consigue una estructura de representación del conocimiento que nos permitirá realizar la tarea de desambiguación de las palabras objetivo.

3.4.1. Extracción del contexto

La competición del SemEval-2010 que estamos utilizando para evaluar nuestro sistema, proporciona tan solo una frase (en la que aparecía la palabra ambigua) cómo única información de contexto. Dicho contexto es el que se va a utilizar para buscar la traducción más probable como conocimiento auxiliar y seleccionar las posibles traducciones de la palabra ambigua. En la Figura 1 se muestra un esquema del funcionamiento de la desambiguación de la palabra objetivo.

En primer lugar, se realiza un análisis de la frase en la que aparece la palabra objetivo, y se extraen las palabras más representativas. En los experimentos realizados hemos tenido en cuenta los nombres, adjetivos y verbos, aunque el uso de verbos ha sido descartado debido a la reducción significativa del rendimiento del algoritmo. Una vez hecho esto, utilizando el diccionario, se obtienen todas las posibles traducciones de cada una de las palabras del contexto. Dado que los grafos con los que trabajamos están formados únicamente por nombres o nombres y adjetivos, se seleccionan las traducciones de las palabras del contexto que corresponden a estas categorías gramaticales.

Posteriormente se identifican aquellas comunidades que contienen al menos una traducción ya sea de las palabras del contexto o de la palabra objetivo. Como resultado, tenemos un conjunto de comunidades en las que aparece al menos una traducción de la palabra objetivo M_T, y otro conjunto de comunidades en las que aparece al menos una traducción de las palabras del contexto M_C. Utilizando el grafo de comunidades, se calculan las distancias entre cada comunidad dentro de M_T y cada comunidad de M_C. Teniendo en cuenta que una traducción de la palabra objetivo puede pertenecer a la misma comunidad que otras traducciones de las palabras del contexto, la distancia en este caso sería 1, y por tanto se normalizan el resto de las distancias sumando 1 a su valor original.

Nuestra hipótesis en este punto, se basa en que aquella traducción de la palabra objetivo que se encuentre más cerca de las traducciones de las palabras del contexto, tiene una mayor probabilidad de ser la traducción correcta. De esta forma hemos establecido una ponderación de cada una de las traducciones de la palabra objetivo en base a dos factores. El primero de ellos es la distancia entre las comunidades que contienen las traducciones de la palabra objetivo y las comunidades que contienen las traducciones de las palabras del contexto. El segundo se basa en la cantidad de traducciones de las palabras del contexto que contiene la comunidad considerada. El peso asignado a cada traducción, w_{t_j}, viene dado por la fórmula

$$w_t = \max_{M_C^i \in M_C} \frac{A_C^i}{(d_{M_C^i M_T^t} + 1)} \qquad (6)$$

donde A_C^i es el número de traducciones del contexto que contiene M_C^i, y $d_{M_C^i M_T^t}$ es el número de pasos (distancia) existente entre la comunidad M_C^i y la comunidad M_T^t, es decir, aquella en la que se encuentra la traducción analizada.

Esta ponderación obtenida por cada traducción de la palabra ambigua, se utiliza posteriormente para ordenar dichas traducciones.

4. Experimentación y resultados

En este apartado se expondrá el método que se ha seguido para realizar la evaluación del sistema propuesto, así como los resultados obtenidos y su comparación con otros sistemas.

4.1. Método de evaluación

La metodología seguida para realizar la evaluación del sistema es la misma utilizada en la ta-

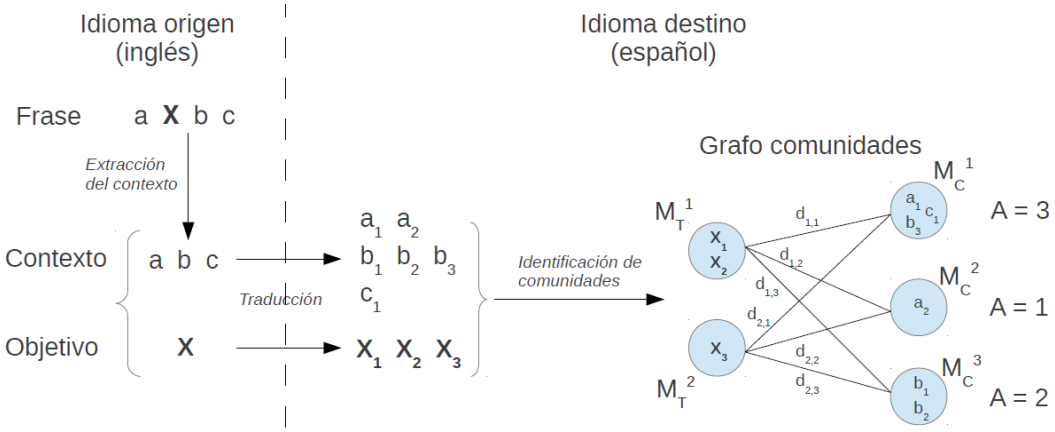

Figura 1: Diagrama del funcionamiento del algoritmo de desambiguación multilingüe de palabras.

rea 3 de la competición SemEval 2010, denominada *Cross-Lingual Word Sense Disambiguation* (Lefever y Hoste, 2010). Los conjuntos de datos proporcionados para la tarea consisten en un subconjunto *Trial*, compuesto por cinco palabras y veinte frases etiquetadas con la palabra a desambiguar para cada una de ellas, y un subconjunto *Test*, compuesto por veinte palabras, y cincuenta frases etiquetadas con la palabra ambigua. En (Lefever y Hoste, 2010) se puede encontrar una descripción más detallada del proceso de construcción de estos conjuntos de datos.

La evaluación se realiza sobre las palabras de test, a partir de un *Gold-Standard* que ofrece las traducciones más probables, ordenadas según su peso, para cada una de las palabras en cada uno de los contextos proporcionados. En la competición SemEval 2010 se realizaron dos tipos de evaluaciones. La primera denominada *Best*, permitía al sistema evaluado proponer tantas traducciones como decidiese que eran adecuadas, pero la puntuación se dividía por el número de palabras propuestas. En la segunda evaluación, denominada *Out-Of-Five*, se permitía al sistema proponer hasta un máximo de cinco traducciones para cada palabra, y la puntuación se obtiene únicamente en función de si las palabras propuestas se encuentran en el *Gold-Standard* y en qué posición. El sistema propuesto en el presente trabajo está enfocado hacia la evaluación *Out-Of-Five*, por lo que se proponen las cinco traducciones que el sistema considera más probables.

La puntuación de los sistemas se ofrece en términos de precisión y cobertura (*precision* y *recall*), y el ranking de los sistemas se obtiene a partir de la Medida-F (*F-Measure*).

4.2. Resultados

Como ya hemos indicado, en este trabajo nos centramos únicamente en la traducción del inglés al español. El sistema ha sido evaluado utilizando, por una parte, únicamente nombres para construir el grafo de coaparición de palabras, y por otra parte, utilizando nombres y adjetivos. También se ha variado el umbral a partir del cuál se considera que una coaparición de palabras es estadísticamente significativa, es decir, el valor máximo que puede tomar el *p-valor* explicado en la sección 3.2 para generar un enlace entre dos palabras en el grafo.

La Figura 2 nos muestra gráficamente los valores que toma la Medida-F para cada una de las configuraciones de parámetros. Los nombres se representan como “NN”, mientras que los adjetivos se representan como “JJ”. Como se puede observar, los mejores resultados, tanto para el grafo que utiliza sólo los nombres, como para el grafo que utiliza nombres y adjetivos, se consiguen con los umbrales de $10^{(-13)}$ y $10^{(-11)}$ respectivamente. La evolución de estos valores confirma la utilidad del algoritmo empleado, ya que el aumento del umbral repercute en la permanencia de unas relaciones cada vez más significativas. De esta forma, se prueba el hecho de que el grafo establece unos vínculos entre términos tan representativos que el hecho de eliminar enlaces superfluos se traduce en una mejora de los resultados. Esta mejora se mantiene hasta que el grafo comienza a perder enlaces realmente relevantes como se puede apreciar en la figura. También se muestra que los mejores resultados se obtienen cuando se utilizan nombres y adjetivos, que enriquecen la representación frente a la utilización únicamente de nombres.

En el Cuadro 1 se puede observar la compara-

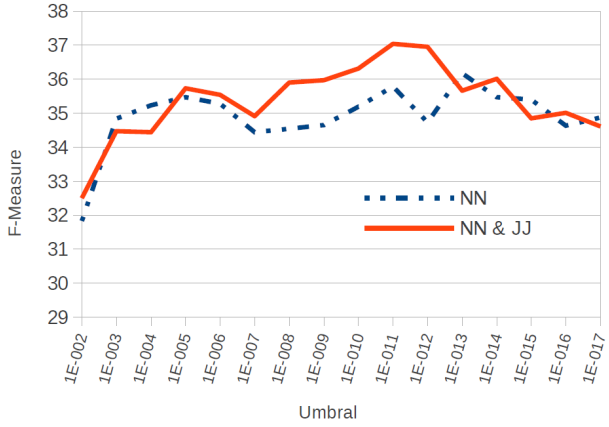

Figura 2: Evolución de la Medida-F según el umbral utilizado para los métodos basados en el uso de nombres (NN), y nombres y adjetivos (NN & JJ)

ción entre los resultados obtenidos por el sistema propuesto, y los obtenidos por los diversos sistemas no supervisados que compitieron en la tarea 3 del SemEval-2010. El sistema propuesto, en su configuración óptima (utilizando nombres y adjetivos para construir el grafo y con un umbral de 10^{-11}) supera a todos los sistemas no supervisados que presentaron resultados para el idioma español en dicha competición. Como se puede observar, nuestro algoritmo presenta el mejor valor de cobertura, en detrimento de algo de precisión. Esto es debido a que el método de evaluación elegido para desarrollar el sistema fue el *Out-Of-Five*, en donde se valora la proposición de un número de hasta cinco traducciones por encima de la selección de una única traducción.

Sistema	Medida-F	P	C
Propuesto	**37.04**	37.04	**37.04**
T3-COLEUR	35.67	35.84	35.46
UHD-1	34.95	**38.78**	31.81
UHD-2	34.22	37.74	31.30

Cuadro 1: Resultados en función de la Medida-F, Precisión (P) y Cobertura (C) obtenidos por nuestro algoritmo, en comparación con los sistemas no supervisados participantes en la tarea 3 del SemEval-2010.

5. *Conclusiones y Trabajo Futuro*

En este trabajo hemos abordado la tarea de la desambiguación del sentido de las palabras en un contexto multilingüe y con un enfoque no supervisado. La idea subyacente ha sido inducir los sentidos de las palabras en el idioma destino, en este caso el español, mediante una nueva técnica de construcción de grafos. Después hemos utilizado las traducciones de las palabras del contexto de la palabra origen para identificar la correspondencia más probable con las traducciones de la palabra considerada. Esta metodología, aplicada a los datos de la tarea 3 de la competición Semeval-2010, ha conseguido superar a todos los sistemas no supervisados que participaron en la tarea. Sin embargo, hay muchos aspectos de la propuesta cuya investigación nos proponemos abordar. Uno de ellos es refinar la medida de distancia entre comunidades utilizando los pesos de los enlaces del grafo de coaparición. Otro aspecto a estudiar, consiste en analizar distintas alternativas para tratar los casos de empates que se puedan producir al seleccionar la mejor traducción. También queremos investigar otros algoritmos de detección de comunidades así como aplicar nuestro algoritmo a otras lenguas.

Bibliografía

Agirre, Eneko, David Martínez, Oier Lopez de Lacalle, y Aitor Soroa. 2006. Two graph-based algorithms for state-of-the-art wsd. En *EMNLP*, páginas 585–593.

Agirre, Eneko y Aitor Soroa. 2007. Ubc-as: A graph based unsupervised system for induction and classification. En *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, páginas 346–349, Prague, Czech Republic, June. Association for Computational Linguistics.

Banea, Carmen y Rada Mihalcea. 2011. Word sense disambiguation with multilingual features. En *Proceedings of the Ninth International Conference on Computational Semantics*, IWCS '11, páginas 25–34. Association for Computational Linguistics.

Diab, Mona T. y Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. En *ACL*, páginas 255–262.

Fernandez-Ordonez, Erwin, Rada Mihalcea, y Samer Hassan. 2012. Unsupervised word sense disambiguation with multilingual representations. En *LREC*, páginas 847–851.

Guo, Weiwei y Mona Diab. 2010. Coleur and colslm: A wsd approach to multilingual lexical substitution, tasks 2 and 3 semeval 2010. En *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, páginas 129–133, Strouds-

burg, PA, USA. Association for Computational Linguistics.

Ide, Nancy y Jean Veronis. 1998. Word sense disambiguation: The state of the art. *Computational Linguistics*, 24:1–40.

Klapaftis, Ioannis P. y Suresh Manandhar. 2008. Word sense induction using graphs of collocations. En *Proceeding of the 2008 conference on ECAI 2008*, páginas 298–302, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. En *MT summit*, volumen 5.

Lefever, Els y Veronique Hoste. 2010. Semeval-2010 task 3: Cross-lingual word sense disambiguation. En *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, páginas 15–20, Stroudsburg, PA, USA. Association for Computational Linguistics.

López-Ostenero, Fernando. 2002. *Un sistema interactivo para la búsqueda de información en idiomas desconocidos por el usuario*. Ph.D. tesis, Departamento de Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia.

Màrquez, Lluís, Gerard Exsudero, David Martínez, y German Rigau. 2006. Supervised corpus-based methods for wsd. En *Word Sense Disambiguation: Algorithms and Applications*, volumen 33 de *Text, Speech and Language Technology*. Springer, Dordrecht, The Netherlands, páginas 167–216.

Martinez-Romo, Juan, Lourdes Araujo, Javier Borge-Holthoefer, Alex Arenas, José A. Capitán, y José A. Cuesta. 2011. Disentangling categorical relationships through a graph of co-occurrences. *Phys. Rev. E*, 84:046108, Oct.

Mihalcea, Rada. 2006. Knowledge-based methods for wsd. En *Word Sense Disambiguation: Algorithms and Applications*, volumen 33 de *Text, Speech and Language Technology*. Springer, Dordrecht, The Netherlands, páginas 107–132.

Ng, Hwee Tou, Bin Wang, y Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. En *ACL*, páginas 455–462.

Pedersen, Ted. 2006. Unsupervised corpus-based methods for WSD. En *Word Sense Disambiguation: Algorithms and Applications*. Springer, páginas 133–166.

Pons, P. y M. Latapy. 2005. Computing communities in large networks using random walks. *Lect. Notes Comput. Sci.*, 3733:284.

Resnik, Philip. 2004. Exploiting hidden meanings: Using bilingual text for monolingual annotation. En *Int. Conf. Computational Linguistics and Intelligent Text Processing (CICLing)*, páginas 283–299.

Resnik, Philip y David Yarowsky. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. En *Proceedings of international conference on new methods in language processing*, volumen 12, páginas 44–49. Manchester, UK.

Silberer, Carina y Simone Paolo Ponzetto. 2010. Uhd: Cross-lingual word sense disambiguation using multilingual co-occurrence graphs. En *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, páginas 134–137, Stroudsburg, PA, USA. Association for Computational Linguistics.

van Gompel, Maarten. 2010. Uvt-wsd1: A cross-lingual word sense disambiguation system. En *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, páginas 238–241, Stroudsburg, PA, USA. Association for Computational Linguistics.

Veronis, Jean. 2004. Hyperlex: lexical cartography for information retrieval. *Computer Speech & Language*, 18(3):223–252.

Vilariño, Darnes, Carlos Balderas, David Pinto, Miguel Rodríguez, y Saul León. 2010. Fcc: Modeling probabilities with giza++ for task #2 and #3 of semeval-2. En *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, páginas 112–116, Stroudsburg, PA, USA. Association for Computational Linguistics.

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 195-202 recibido 31-03-2013 revisado 16-06-2013 aceptado 21-06-2013

A social tag-based dimensional model of emotions: Building cross-domain folksonomies

Un modelo dimensional de emociones basado en etiquetas sociales: Construcción de folcsonomías en dominios cruzados

Ignacio Fernández-Tobías, Iván Cantador
Universidad Autónoma de Madrid
C/ Francisco Tomás y Valiente, 11
28049 Madrid, Spain
{i.fernandez,ivan.cantador}@uam.es

Laura Plaza
UNED NLP & IR Group
C/ Juan del Rosal, 16
28040 Madrid, Spain
lplaza@lsi.uned.es

Resumen: En este trabajo se presenta un modelo dimensional de emociones basado en etiquetas sociales. El modelo se construye sobre un léxico generado automáticamente que caracteriza emociones por medio de términos sinónimos y antónimos. Este léxico se enlaza con diversas folcsonomías emocionales específicas de dominio. Se propone una serie de métodos para transformar perfiles de objetos basados en etiquetas sociales en perfiles emocionales. El objetivo de estos perfiles es su uso por parte de sistemas adaptativos y de personalización que permitan recuperar o recomendar contenidos en función del estado de ánimo del usuario. Para validar el modelo, se muestra que la representación de un conjunto de emociones básicas se corresponde con la del aceptado modelo de Russell. También se reportan resultados de un estudio de usuario que demuestran una alta precisión de los métodos propuestos para inferir emociones evocadas por objetos en los dominios del cine y la música.
Palabras clave: emociones, léxico afectivo, etiquetado social, folcsonomías

Abstract: We present an emotion computational model based on social tags. The model is built upon an automatically generated lexicon that describes emotions by means of synonym and antonym terms, and that is linked to multiple domain-specific emotion folksonomies extracted from entertainment social tagging systems. Using these cross-domain folksonomies, we develop a number of methods that automatically transform tag-based item profiles into emotion-oriented item profiles, which may be exploited by adaptation and personalization systems. To validate our model, we show that its representation of a number of core emotions is in accordance with the well known psychological circumplex model of affect. We also report results from a user study that show a high precision of our methods to infer the emotions evoked by items in the movie and music domains.
Keywords: emotions, affective lexicon, social tagging, folksonomies

1. Introduction and background

The study and development of computational systems aimed to recognize and interpret human feelings is usually referred to as *Affective Computing* (Picard, 1995). In Natural Language Processing, this discipline - often known as *Sentiment Analysis* - is becoming increasingly important with the development of the Social Web and the growing popularity of online forums, social networks, and collaborative tagging systems (Carrillo-De-Albornoz, Plaza, and Gervás, 2010; De Choudhury and Gamon, 2012).

Focusing on User Modeling and Recommender Systems, emotions (and moods) can be efficiently used in a wide range of applications; e.g. constructing user behaviour models (Hastings et al., 2011), tailoring the search results (Meyers, 2007) and filtering the recommending items (Winoto and Ya Tang, 2010). The user's mood has proven to have an important influence on the choice of items that the user is more likely to consume, and therefore the system should be able to suggest items according to that mood.

In this context, modeling and exploiting emotions present challenging problems. First, there is not agreement on the categorization of emotions to be used. Focusing on computational models of emotion, three main psy-

ISSN 1135-5948

chological theories have been adopted, namely the *categorical emotion theory* (James, 1984) - which characterizes emotions as discrete units with boundaries -, the *emotional dimension theory* (Russell, 1980) - which conceptualizes emotions as points in a continuous space -, and the *appraisal theory* (Scherer, Shorr, and Johnstone, 2001) - which represents emotions as outcomes of certain events and situations. Second, detecting emotions in text is extremely difficult. Most approaches to the problem use emotion lexicons that provide specific vocabularies for describing emotions. SentiWordNet (Baccianella, Esuli, and Sebastiani, 2010), for instance, associates to each WordNet synset three numerical scores *obj*, *pos* and *neg*, describing how objective, positive, and negative the terms in the synset are. More fine-grained is SentiSense (Carrillo-De-Albornoz, Plaza, and Gervás, 2012), which attaches specific emotions (e.g. *sad* or *like*) to WordNet concepts.

Existing emotion lexicons, usually developed for polarity classification of texts, present several drawbacks that limit their use in personalization and recommendation systems. First, they use a single set of generic emotions for categorizing all terms in the lexicon, but, as we will see, the emotions that are evoked by items usually are domain-specific - such as *scare* in the movie domain, and *peacefulness* in the music domain. Second, the vocabulary employed by users to characterize the different emotions varies with the application domain, and thus it is necessary to develop domain-specific emotional lexicons. Third, in order to perform cross-domain recommendation (i.e., to suggest items in a target domain using user feedback about items in a different source domain), an automatic method for translating emotional information from one domain to another is required.

In this paper, we propose an automatic approach that generates a core lexicon and different folksonomies to represent both generic and domain-dependent emotion categories. These resources are generated from a generic thesaurus and social tagging systems in entertainment domains, namely the movie and music domains. More specifically, we propose a model in which emotions are represented as vectors of weighted synonym and antonym terms, and which enables computing (dis)similarities between emotions. In this way, it is possible to relate core emotions and domain-specific emotions, and thus to extrapolate emotional information from one domain to another. We think our model could be exploited by adaptive and personalized systems that are based on a keyword- or concept-based knowledge representations.

2. A core domain-independent emotion lexicon

Our model adopts the emotional dimension theory, and is based on the Russell's circumplex model of affect (Russell, 1980). This model understands emotions as a linear combination of two dimensions, *pleasure* and *arousal*, as shown in Figure 1. Arousal (in the vertical axis) reflects the intensity of an emotion; and pleasure (in the horizontal axis) reflects whether an emotion is positive or negative. With this representation, any emotion can be represented at any level of arousal and pleasure. Hence, for instance, *happiness* and *sadness* can be considered as emotions with the highest and lowest levels of pleasure, respectively, but with neutral levels of arousal, with respect to other emotions such as *tension* (with high arousal) and *calmness* (with low arousal). Russell proposed a set of 16 core (basic) emotions (see Figure 1). We will use this set of emotions in our model.

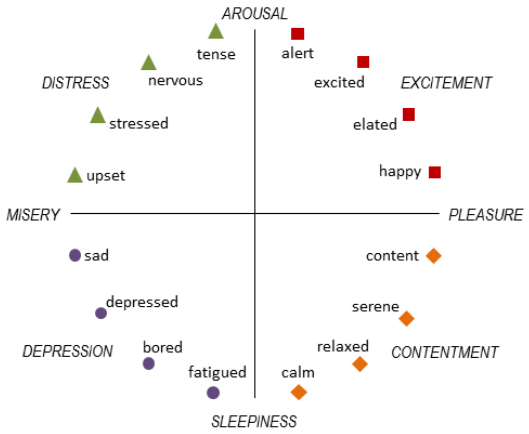

Figure 1: Distribution of core emotions in the circumplex model

Our dimensional model is built upon an automatically generated lexicon $L = \{t_1, ..., t_K\}$ composed of synonym and antonym terms t_k of the core emotions' names (e.g. *happy*, *sad*). The synonym and antonym terms of each emotion are obtained from the online thesaurus provided by Dic-

tionary.com.[1] Table 1 shows some of the obtained synonyms for each emotion.

Emotion	Synonym terms
alert	alert, active, animated, lively
excited	excited, stimulated, agitated, moved
elated	elated, jubilant, overjoyed, exhilarated
happy	happy, merry, cheerful, joyful, bright
content	content, satisfied, gratified, pleased
serene	serene, quiet, placid, tranquil
relaxed	relaxed, moderated, mitigated, loose
calm	calm, mild, appeased, smooth, soften
fatigued	fatigued, tired, fatigued, drained
bored	bored, apathetic, exasperated
depressed	depressed, dejected, despondent
sad	sad, sorrowful, doleful, downcast
upset	upset, bother, disturbed, troubled
stressed	stressed, tormented, harassed, vexed
nervous	nervous, apprehensive, uneasy
tense	tense, restless, uptight, jittery

Table 1: Core emotions and their synonyms

Once the lexicon L is generated, a core emotion $e_i \in E$ is represented as a vector $e_i = (e_{i,1}, ..., e_{i,k}) \in R^K$, in which the component $e_{i,k}$ corresponds to the term $t_k \in L$ and is computed as shown in eq. (1). The component $e_{i,k}$ is greater than 0 if the term t_k is a synonym of the emotion e_i, lower than 0 if t_k is an antonym of e_i, and 0 otherwise. Its absolute value corresponds to the TF-IDF weight (Baeza-Yates and Ribeiro-Neto, 2011) of t_k computed by considering the lexicon L as the collection vocabulary, and the set E of emotions (described as sets of synonym and antonym terms) as the collection documents.

$$e_{i,k} = \begin{cases} tf \times idf(e_{i,k}) \; si \; t_k \in synonyms(e_i) \\ -tf \times idf(e_{i,k}) \; si \; t_k \in antonyms(e_i) \\ 0 \text{ otherwise} \end{cases} \quad (1)$$

With the proposed vector representation of emotions, we can measure (dis)similarities between emotions. Specifically, we can use the well known cosine similarity (Baeza-Yates and Ribeiro-Neto, 2011). Figure 2 shows the cosine similarity between each pair of core emotions. The cell colors indicate the sign of the similarity values - being black for positive values and white for negative ones -, and the cell intensities correspond to the similarity absolute values - being dark for values close to 1, and light for values close to 0. The emotions are sorted according to the quadrants of Russell's model (Figure 1). We can observe that emotions in the same quadrant have high similarities (e.g. *alert* and *excited*), while emotions in opposite quadrants have low similarities (e.g. *calm* and *excited*). These results show that our tag-based model is in accordance with the circumplex model.

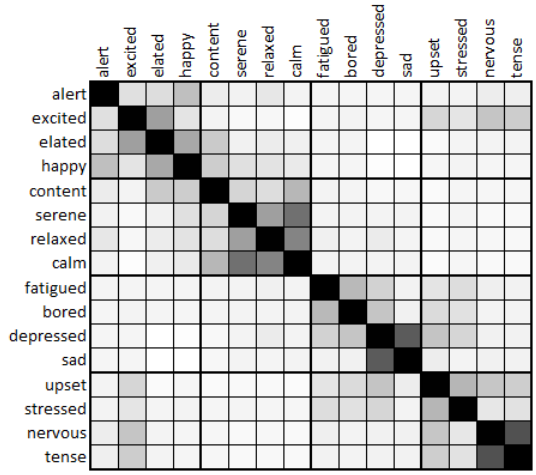

Figure 2: Similarity between core emotions

To better show the correspondences between our computational model and the theoretical circumplex model, Figure 3 shows the projections of our emotion vectors into a two-dimension space by applying *Principal Component Analysis*. We observe that our model locates all 16 core emotions in their corresponding quadrants.

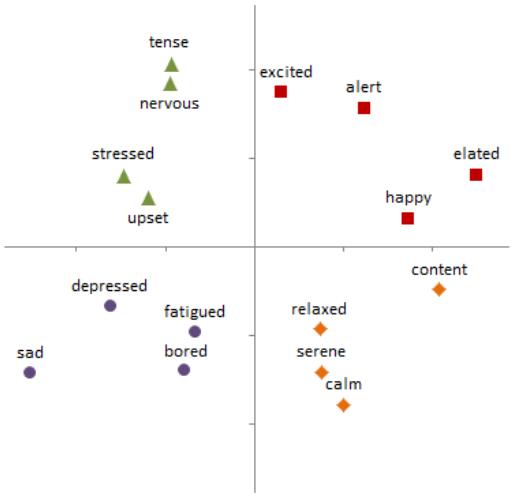

Figure 3: Distribution of core emotions in our tag-based model

More interestingly, we observe that in our model the axes defined by the two most informative principal components are related to the arousal and pleasure factors of the circumplex model. Hence, positive emotions

[1] Dictionary.com thesaurus, http://thesaurus.com

(e.g. *happy*, *content*) are in right quadrants, while negative emotions (e.g. *sad*, *upset*) are in left quadrants; and more intense emotions (e.g. *tense*, *excited*) are in the upper quadrants, while less intense emotions (e.g. *relaxed*, *fatigated*) are in the lower quadrants.

3. Cross-domain emotion folksonomies

In a social tagging system users create items and annotate them with freely chosen tags. The whole set of tags constitutes an unstructured knowledge classification scheme that is known as *folksonomy*. This implicit classification is then used to search and recommend items (Cantador, Brusilovsky, and Kuflik, 2011). Within the set of tags that express qualities and opinions about the items, there are tags that refer to emotions caused by the items. In most cases, however, such emotions are not the core emotions presented in the previous section, but domain-specific emotional categories - such as *suspense* in the movie domain, and *nostalgia* in the music domain -, which indeed may be related to one or more core emotions. In this section we extend our emotion model by linking the core emotions with domain-specific emotional categories described by tags in different folksonomies. Specifically, we focus on the movie and music entertainment domains by exploiting the MovieLens and Last.fm folksonomies provided in the HetRec'11 workshop (Cantador, Konstas, and Jose, 2011). With the extended model we propose to build emotion-oriented item profiles and cross-domain folksonomies. This process is illustrated in Figure 4. We make all the data publicly available.[2]

3.1. An emotion folksonomy for the movie domain

To build the emotion folksonomy for movies, we first select a total of 15 emotional categories listed under the mood topic in Jinni [3] movie search and recommendation system. We describe each category by 4 to 6 associated feeling terms, and use them as seed terms (see Table 2). Next, we extend the seed terms with their synonyms and antonyms from Thesaurus.com, but restricted to those existing as social tags in the MovieLens dataset. Finally, we repeat the process in Section 2 to represent an emotional category as a vector of weighted terms. Table 2 shows the number of terms per category that we collected.

Category	Seed terms	#
clever	clever, cerebral, reflective	71
offbeat	offbeat, quirky, surreal	83
exciting	exciting, energetic, frantic	104
suspenseful	suspenseful, tense	34
captivating	captivating, rousing, poignant	83
emotional	emotional, passionate, romantic	185
feel good	cute, merry, happy	41
humorous	humorous, funny, comical	101
sexy	sexy, erotic, sensual	39
sexual	sexual, lascive, horny	16
uplifting	uplifting, inspirational, hope	32
bleak	bleak, grim, depressing	84
gloomy	gloomy, sad, melancholic	85
rough	rough, brutal, lurid, macabre	126
scary	scary, creepy, menacing	57

Table 2: Movie emotional categories, seed terms and number of terms per category

Figure 5 shows the cosine similarity between each pair of emotional categories. It can be observed that close emotional categories, such as *gloomy* and *bleak*, present high similarity, while very distinct categories, such as *gloomy* and *feel good*, present low similarity.

3.2. An emotion folksonomy for the music domain

To generate an emotion folksonomy in the music domain, we select as emotional categories the 9 emotions proposed in the GEMS (Geneva Emotional Music Scales) model (see Table 3). As initial seed terms we use the category names and their associated feeling terms given in (Zentner, Grandjean, and Scherer, 2008). Next, we extend these terms with their synonyms and antonyms in Thesaurus.com, but restricted to those existing as social tags in the Last.fm dataset. The emotional category vectors are then created as for the movie domain. Table 3 shows some of the most informative tags for each emotional category, along with the total number of tags we collected for each category.

Figure 5 shows the similarity between each pair of emotional categories. Again, close categories, such as *tenderness* and *nostalgia*, present high similarity, while very distinct categories, such as *sadness* and *joy*, present low similarity.

[2] http://ir.ii.uam.es/emotions/

[3] http://www.jinni.com

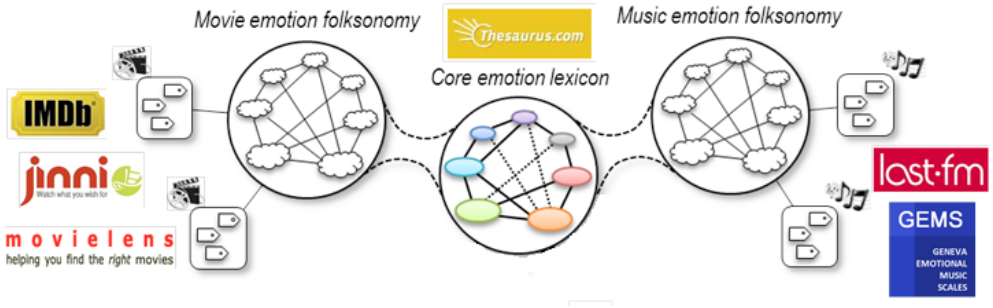

Figure 4: Crossing domain-dependent folksonomies

Category	Seed terms	#
joy	funny, happy, amusing, jolly	84
nostalgia	nostalgic, melancholic, sentimental	49
peacefulness	peaceful, quiet, calm, gentle	71
power	powerful, strong, energetic, intense	97
sadness	sad, sorrowful, unhappy, dismal	51
tenderness	tender, gentle, mellow, romantic	41
tension	tense, edgy, angry, fierce	58
transcendence	fascinating, enchanting	45
wonder	wonderful, strange, fantastic	24

Table 3: Music emotional categories, seed terms and number of terms per category

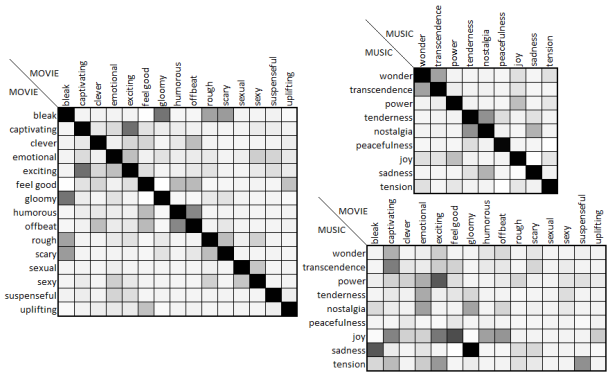

Figure 5: Similarity values between movie and music emotional categories

3.3. Emotion-oriented tag-based profiles

The proposed representation of emotions lets transform tag-based item profiles (i.e., the items' annotation sets) into emotion-oriented profiles. To this end, we first transform the tag-based profiles into domain emotion-oriented profiles. Next, the domain emotion-oriented profiles are transformed into core emotion-oriented profiles. Formally, let a core emotion $e_i^C \in E$ and a domain-specific emotional category $e_j^D \in E_D$ be defined as in eq. (1). For an item (o_n), let $o_n^T = (o_{n,1}, ..., o_{n,|\tau|}) \in \Re^{|\tau|}$ be the item's tag-based profile. Then, from such profile, we define:

- the item's **domain emotion-oriented profile** as $p_n^D = (p_{n,1}, ..., p_{n,|E_D|}) \in [-1, 1]$, where each component represents a domain emotion, and its weight is computed as $p_{n,i} = cos(o_n^\tau, e_i^D)$, and
- the item's **core emotion-oriented profile** as $q_n^C = (q_{n,1}, ..., q_{n,|E|}) \in [-1, 1]$, where each component corresponds a core emotion, and its weight is computed as $q_{n,i} = \sum_{k=1}^{|E_D|} p_{n,k} \times cos(e_i^C, e_k^D)$.

Moreover, for each of these types of emotion-oriented profiles, we consider two alternatives for defining the emotion vectors: **basic vectors**, whose components correspond to terms of the lexicon, and **extended_N vectors** , whose components correspond to the N folksonomy tags that cooccur most frequently (in the tag-based item profiles) with the terms of the basic vectors.

3.4. Crossing folksonomies in different domains

The proposed model let us to relate core emotions and domain-specific emotional categories by computing the cosine similarity between their vector representations. Figure 6 shows the relation between some domain-specific emotional categories and the different core emotions for both the movie and music domains. It can be observed that, for instance, the emotional category *suspenseful* in the movies domain strongly overlaps with the *tense* and *nervous* core emotions, while the *peacefulness* category in the music domain intersects tightly with the *calm*, *relaxed* and *serene* core emotions.

Moreover, the intersection between cross domain-specific emotional categories could

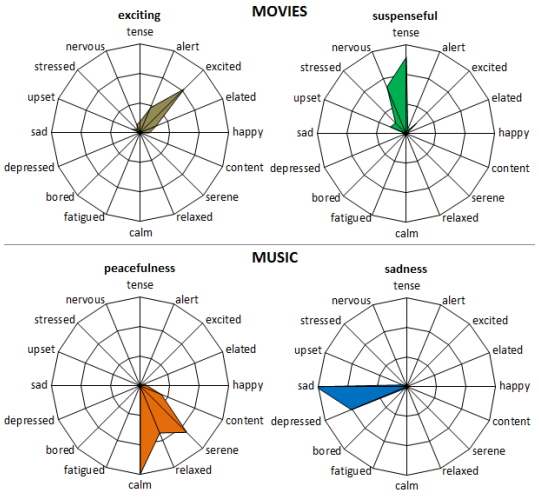

Figure 6: Relation between core emotions and domain-specific emotional categories

be computed to obtain a measure of similarity between them. Figure 5 shows the cosine similarity between pairs of cross-domain emotional categories. It can be seen that pairs of categories such as *feel good-joy* and *gloomy-sadness*, which are very close in pleasure and arousal, present very high similarity, while very distinct categories, such *joy-gloomy* and *sadness-uplifting*, present very low similarity.

4. Experiments and results

To evaluate our emotional model, we conducted a user study in which participants, recruited via social networking sites, were presented with sets of movies or musicians (no combinations of both), and were requested to freely select one or more domain-specific and core emotions for each item. A total of 71 users participated, evaluating 165 movies and 129 musicians. They generated 703 evaluation cases, assigning an average of 4.08 and 3.38 domain-specific emotional categories, and 3.30 and 4.18 core emotions, to items in the movies and music domains, respectively. To facilitate the evaluation, the users could select preferred movie and music genres and the language - English or Spanish - of the online evaluation tool[4] (see Figure 7), and skip any item they did not want to evaluate. We note that, as expressed by some of the participants, there are cases in which it is difficult to assign certain emotions to an item. Opposite emotions (e.g. *happiness* and *sadness*) can be evoked in different parts of a particular movie, and by different compositions of the same musician.

In the user study participants stated which core and domain-specific emotions they consider as relevant for each item (movie or musician), thus manually (and collectively) creating emotion-oriented item profiles, which we consider as ground truth. To evaluate the quality of the emotion-oriented profiles generated by our methods (Section 3.3) with respect to the ground truth profiles, we compared them by means of IR precision metrics. Specifically, we computed Precision at position k, $P@k$ (Baeza-Yates and Ribeiro-Neto, 2011), which, for a particular item, is defined as the percentage of the top k emotions returned by a method that are relevant for the item, as stated by the users of our study. We also computed $R-precision$ (Buckley and Voorhees, 2005), which is defined as the precision of the top R emotions returned by a method for an item, being R the number of emotions that are relevant for the item, as stated by the users of our study. That is, $R-precision$ is $P@R$, i.e. it is the break-even point in the precision-recall curve where precision is equal to recall.

Table 4 shows average precision for the different methods (and a random method) on the movie and music domains. The basic method was the best performing one in both domains (with highest $P@1$ values around 70 %), only outperformed by the *extended_10* method in the movie domain for the core emotion-oriented profiles. In general, the methods performed in the music domain better than in the movie domain, and were able to identify domain emotional categories more effectively than core emotions in both domains.

Table 5 shows the top two emotional categories assigned by the users to items belonging to some of the 26 genres considered from the Jinni and Last.fm systems. This table also shows the two predominant emotional categories for each genre, according to our emotion-tag based profiles. It can be seen that the emotions assigned by our model are very similar to that assigned by the users.

Finally, and concerning the frequency with which the different emotions are associated with the movies and musicians in our experiment, Figure 8 shows the percentage of items that have been assigned a given emotion (both core and domain-specific). As it

[4] Evaluation tool, url omitted to ensure anonymity

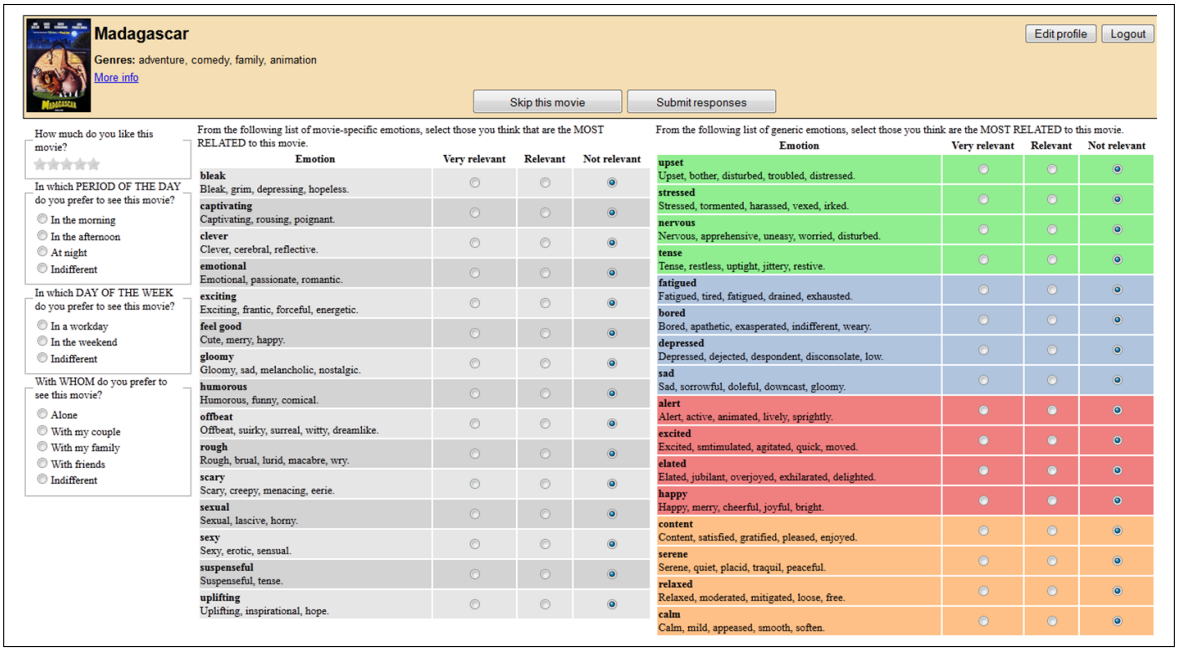

Figure 7: User study - online evaluation tool

Profile type	Vector model	Movie domain #evals	*P*@1	**P@2**	**P@3**	**R-Pr**	Music domain #evals	*P*@1	*P*@2	*P*@3	**R-Pr**
core emotion-oriented	random	165	0.297	0.305	0.302	0.300	129	0.327	0.339	0.345	0.348
	basic	107	0.598	0.528	0.514	0.481	109	0.606	**0.670**	**0.636**	**0.547**
	extended_10	77	**0.675**	**0.643**	**0.589**	**0.519**	11	**0.636**	0.636	0.546	0.497
	extended_50	142	0.373	0.324	0.406	0.365	44	0.546	0.625	0.568	0.502
	extended_100	155	0.419	0.390	0.411	0.399	79	0.557	0.620	0.582	0.546
domain emotion-oriented	random	165	0.379	0.382	0.377	0.380	129	0.418	0.416	0.414	0.414
	basic	108	**0.722**	0.625	**0.571**	**0.579**	109	**0.743**	**0.587**	**0.532**	**0.546**
	extended_10	77	0.675	**0.656**	0.554	0.399	11	0.727	0.546	0.455	0.503
	extended_50	144	0.507	0.490	0.463	0.412	44	0.682	0.443	0.394	0.428
	extended_100	158	0.551	0.532	0.513	0.449	79	0.696	0.494	0.426	0.463

Table 4: Avg. $P@k$ and $R-precision$ values of the considered emotion-oriented profiles

Movie domain		
action	**Users**	exciting, suspenseful
	Model	suspenseful, captivating
comedy	**Users**	humorous, feel good
	Model	humorous, feel good
horror	**Users**	scary, rough
	Model	scary, exciting
Music domain		
classical	**Users**	nostalgia, peacefulness
	Model	nostalgia, peacefulness
rock	**Users**	power, tension
	Model	power, joy
jazz	**Users**	nostalgia, peacefulness
	Model	tension, peacefulness

Table 5: Top emotional categories assigned to some movie and music genres by (a) the users (b) inferred using the proposed model

can be observed, the most frequent core emotion is *content*, followed by *happy*. In the music domain, the predominant emotional category is *power*, while the most frequent one in the movie domain is *humorous*.

5. Conclusions and future work

We have presented a computational model that represents emotions as vectors of weighted synonym and antonym terms, which are automatically obtained from an online thesaurus and social tagging systems in different entertainment domains. Our model distinguishes and relates generic core emotions (e.g. *happiness*, *sadness*) with domain-specific emotional categories (e.g. *suspense* in the movie domain, and *nostalgia* in the music domain). This lets transform tag-based profiles into emotion-oriented profiles, and build cross-domain emotion folksonomies.

The next step in our research is to exploit the generated emotion-oriented profiles in adaptation and personalization systems. In

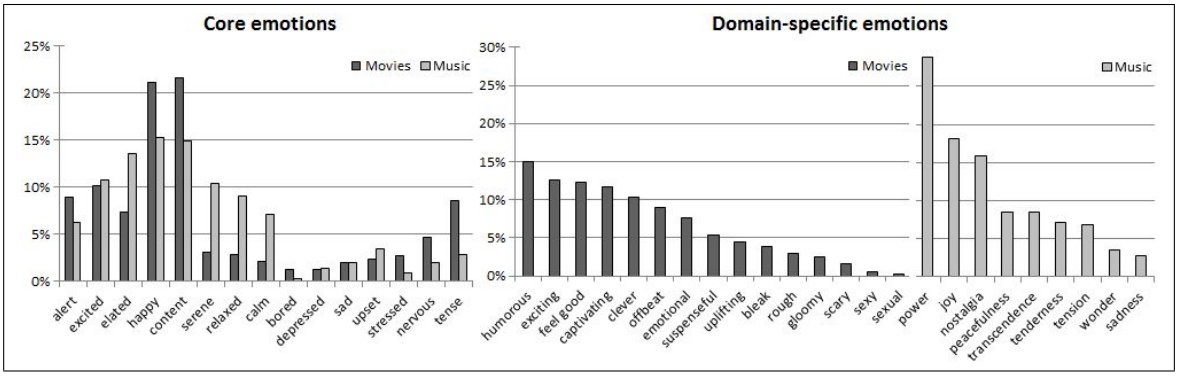

Figure 8: Distribution of core and domain-dependent emotions in the evaluation collection

particular, we plan to use them for developing mood-based and cross-domain recommendation strategies.

References

Baccianella, S., A. Esuli, and F. Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th Conference on Language Resources and Evaluation (LREC'10)*.

Baeza-Yates, R. A. and B.A. Ribeiro-Neto. 2011. *Modern Information Retrieval - The Concepts and Technology behind Search*. Pearson Education.

Buckley, C. and E. Voorhees, 2005. *TREC: Experiment and Evaluation in Information Retrieval*, chapter Retrieval System Evaluation. MIT Press.

Cantador, I., P. Brusilovsky, and T. Kuflik. 2011. Second workshop on information heterogeneity and fusion in recommender systems. In *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys'11)*, pages 387–388.

Cantador, I., I. Konstas, and J. Jose. 2011. Categorising social tags to improve folksonomy-based recommendations. *Journal of Web Semantics*, 9(1):1–15.

Carrillo-De-Albornoz, J., L. Plaza, and P. Gervás. 2010. A hybrid approach to emotional sentence polarity and intensity classification. pages 153–161.

Carrillo-De-Albornoz, J., L. Plaza, and P. Gervás. 2012. Sentisense: An easily scalable concept-based affective lexicon for sentiment analysis. In *Proceedings of the 8th Conference on Language Resources and Evaluation (LREC'12)*.

De Choudhury, M., Counts S. and M. Gamon. 2012. Not all moods are created equal! exploring human emotional states in social media. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*.

Hastings, J., W. Ceusters, B. Smith, and K. Mulligan. 2011. The emotion ontology: Enabling interdisciplinary research in the affective sciences. In *Proceedings of the 7th International and Interdisciplinary Conference on Modeling and Using Context (Context'11)*.

James, W. 1984. What is emotion? *Mind*, 9:188–205.

Meyers, O. C. 2007. A mood-based music classification and exploration system. Master's thesis, School of Architecture and Planning, MIT.

Picard, R. W. 1995. Affective computing. Technical Report 321, MIT Media Laboratory, Perceptual Computing Section.

Russell, J.A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Scherer, K. R., A. Shorr, and T. (Eds.) Johnstone. 2001. *Appraisal Processes in Emotion: Theory, Methods, Research.*

Winoto, P. and T. Ya Tang. 2010. The role of user mood in movie recommendations. *Expert Systems with Applications*, 37(8):6086–6092.

Zentner, M., D. Grandjean, and K. Scherer. 2008. Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion*, 8:494–521.

Demostraciones

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 205-208 recibido 30-04-2013 revisado 16-06-2013 aceptado 21-06-2013

DysWebxia: Textos más Accesibles para Personas con Dislexia*

DysWebxia: Making Texts More Accessible for People with Dyslexia

Luz Rello
NLP & Web Research Groups,
Universitat Pompeu Fabra,
Barcelona, Spain
luzrello@acm.org

Ricardo Baeza-Yates
Yahoo! Labs &
Web Research Group,
Universitat Pompeu Fabra,
Barcelona, Spain
rbaeza@acm.org

Horacio Saggion
NLP Research Group, DTIC,
Universitat Pompeu Fabra,
Barcelona, Spain
horacio.saggion@upf.edu

Resumen: Alrededor del 10% de la población mundial tiene dislexia, un trastorno de aprendizaje que afecta a las habilidades de lectoescritura. Aunque la dislexia tiene un origen neurológico, mediante ciertas modificaciones en los textos podemos conseguir que éstos sean más accesibles para este colectivo. En este trabajo presentamos *DysWebxia*, un modelo público que integra recomendaciones de diseño textual y técnicas de procesamiento del lenguaje natural. El modelo está fundamentado en los resultados de nuestras investigaciones llevadas a cabo con personas con dislexia usando metodologías de evaluación de interacción hombre-máquina como *eye-tracking*. Asimismo presentamos las integraciones actuales del modelo en diferentes aplicaciones de software de lectura.
Palabras clave: Dislexia, diseño textual, simplificación léxica, software de lectura, plug-in, servicio web, tableta, teléfono inteligente.

Abstract: About 10% of the world population has dyslexia, a learning disability affecting reading and writing. Even if dyslexia is neurological in origin, certain text modifications can make texts more accessible for people with dyslexia. We present *DysWebxia*, a public model that integrates our findings from research conducted with this target group by using natural language processing strategies and human computer interaction evaluation techniques such as eye-tracking. This model alters content and presentation of the text to make it more readable. We also present the current integrations of *DysWebxia* in different reading software applications.
Keywords: Dyslexia, text presentation, text explanation, readability, browser plug-in, web service, tablet, smartphone.

1 Introducción

En este trabajo presentamos *DysWebxia*: un modelo para crear textos más legibles para personas con dislexia. A continuación explicaremos:

- La motivación que nos llevó a desarrollar este modelo (Sección 2);
- En qué consiste *DysWebxia* (Sección 3);
- Los experimentos que fundamentan *DysWebxia* y las herramientas que integran el modelo hasta el momento (Sección 4).

* Esta investigación ha sido parcialmente financiada por la beca predoctoral FI de la Generalitat de Catalunya.

2 ¿Por Qué?

La **dislexia** es un trastorno de aprendizaje de origen neurológico que afecta a las habilidades lectoescritoras. Entre un 10 y un 17,5% de los hablantes nativos de inglés (Interagency Commission on Learning Disabilities, 1987) y entre un 7,5 y un 11% de los de español (Carrillo et al., 2011) tienen dislexia, dificultando su acceso a la información escrita como la que se encuentra en la Web.

Aunque en la Web hay un considerable porcentaje de errores de ortografía debidos únicamente a la dislexia (Baeza-Yates y Rello, 2011), en las directrices de accesibilidad Web, *Web Content Accessibility Guidelines (WCAG) 2.0* (Caldwell et al., 2008), la dislexia está incluida dentro de un grupo am-

ISSN 1135-5948

plio de discapacidades cognitivas, sin que se presenten recomendaciones específicas para este colectivo. Sin embargo, investigaciones con este grupo de personas han demostrado que ciertas alteraciones en el texto pueden ayudar a las personas con dislexia a leer mejor (Gregor y Newell, 2000).

A pesar de que se ha señalado extensamente que el uso de lenguaje complejo es una de las dificultades principales para este grupo de usuarios (McCarthy y Swierenga, 2010), todas las aplicaciones existentes sólo alteran el diseño del texto pero no su contenido, como por ejemplo la ex *Firefixia* (Santana et al., 2013), *SeeWord* (Gregor et al., 2003),[1] *Claro Screen Ruler suite*,[2] *Color Explorer*[3] o *Penfriend XL*,[4], siendo *SeeWord* y *Firefixia* las únicas aplicaciones diseñadas a partir de un estudio empírico con personas con dislexia.

Sin embargo, dado que la dislexia afecta al **lenguaje**, la accesibilidad del texto debería ser abordada no sólo desde el punto de vista de su **presentación** sino también en el **contenido** del texto. Por esta razón, *DysWebxia* integra modificaciones textuales de los dos tipos, tanto en la forma como en el contenido.

3 ¿Qué Es?

DysWebxia es un modelo para hacer que los textos sean más accesibles para las personas con dislexia. Está basado en las siguientes características originales:

- El modelo ha sido desarrollado a partir de estudios con personas con dislexia que miden el impacto de ciertas alteraciones textuales en la rapidez de lectura y la comprensión utilizando el seguimiento de la vista (*eye-tracking*).
- Se trata del primer modelo para personas con dislexia que presenta sinónimos de las palabras complejas del texto (Rello et al., 2013a) y que incluye cambios en el diseño de la presentación de texto fundamentados en estudios cuantitativos con personas con dislexia (Rello, Kanvinde, y Baeza-Yates, 2012).
- Aunque *DysWebxia* se implementó por primera vez para la lectura de textos en la Web (Rello, Kanvinde, y Baeza-Yates, 2012), el modelo está siendo adaptado para otras plataformas donde no existía software de lectura específico para personas con dislexia.

4 ¿Cómo?

A continuación presentamos los resultados que integra *DysWebxia* y las diferentes aplicaciones y prototipos que integran el modelo.

4.1 Fundamentación Científica

Los resultados de los experimentos a partir de los cuales se diseñó *DysWebxia* incumben tanto a la forma como al contenido del texto.

Forma: La presentación del texto tiene un efecto significativo en la lectura y en la comprensión de las personas con dislexia (Rello, Kanvinde, y Baeza-Yates, 2012). Por tanto, *DysWebxia* presenta los textos utilizando los parámetros de diseño textual en los que este grupo alcanza la mejor legibilidad y comprensión. Estos parámetros están fundamentados en tres experimentos usando *eye-tracking* con 36, 46 y 48 participantes con dislexia. En estos experimentos se estudiaron: el tipo de fuente (Rello y Baeza-Yates, 2013); la combinación del tamaño de la fuente y el interlineado en el contexto de la Web (Rello et al., 2013); y el ancho de columna, el espacio entre caracteres, el espacio entre párrafos, combinaciones de colores, contrastes de brillo en escalas de grises en la fuente y en el fondo (Rello, Kanvinde, y Baeza-Yates, 2012).

Contenido: Estudios previos han demostrados que las personas con dislexia leen con más dificultad textos que contengan palabras poco frecuentes o muy largas (Rello et al., 2013b), lo que implica que podrían beneficiarse de técnicas del procesamiento del lenguaje natural como la simplificación léxica (Bott et al., 2012).

Así mismo también se investigó el impacto en la lectura de la calidad léxica (Rello y Baeza-Yates, 2012), la paráfrasis (Rello, Baeza-Yates, y Saggion, 2013), la simplificación numérica (Rello et al., 2013) o el uso de mapas conceptuales (Rello et al., 2012) sin llegar a resultados tan diferenciados como con la simplificación léxica.

Sin embargo, un estudio con 47 personas con dislexia (Rello et al., 2013a) ha de-

[1] http://www.computing.dundee.ac.uk/projects/seeword/

[2] http://www.clarosoftware.com/index.php?cPath=348

[3] http://colour-explorer.software.informer.com/9.0/

[4] http://www.penfriend.biz/pf-xl.HTML

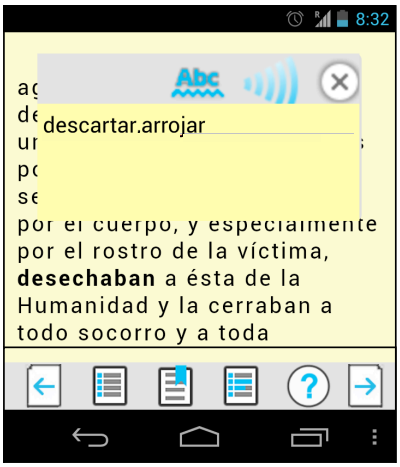

Figure 1: DysWebxia en el lector IDEAL eBook Reader para Android.

mostrado que el hecho de que la simplificación léxica pueda ser útil para este colectivo depende, en gran medida, no sólo de la calidad de los sinónimos generados, sino también de la interacción persona-ordenador, es decir, de cómo se efectúa la presentación de dichos sinónimos al usuario. En este estudio se utilizaron diferentes dispositivos (ordenador portátil, tableta y teléfono inteligente) y se compararon dos estrategias de presentación de simplificaciones léxicas, una sustituyendo los sinónimos y otras en la que sólo se presentaban los sinónimos si el usuario lo solicitaba mediante el uso de *pop-ups*. Los resultados demostraron que las personas con dislexia percibían significativamente como más legibles y comprensibles los textos en los que podían solicitar de manera interactiva los sinónimos. Por el contrario, los textos en los que se había efectuado la sustitución léxica automática no resultaron ser ni más legibles ni más comprensibles que los textos originales ni los textos en que manualmente se realizaron las mejores simplificaciones posibles (*gold standard*).

Por esta razón, en vez de sustitución léxica, *DysWebxia* integra una técnica auxiliar para presentar los sinónimos de las palabras complejas cuando lo demanda el usuario (Figura 3).

4.2 Prototipos

DysWebxia se ha integrado en los siguientes software de lectura:

- Un lector de libros electrónicos para Android[5] (Kanvinde, Rello, y Baeza-Yates, 2012) (Figura 1).
- Un servidor de personalización de textos de páginas Web *Text4all*[6] (Topac, 2012) (Figura 2).
- Como una herramienta de lectura para iOS (Figura 3).

Figure 2: DysWebxia en el servidor web Text4all.

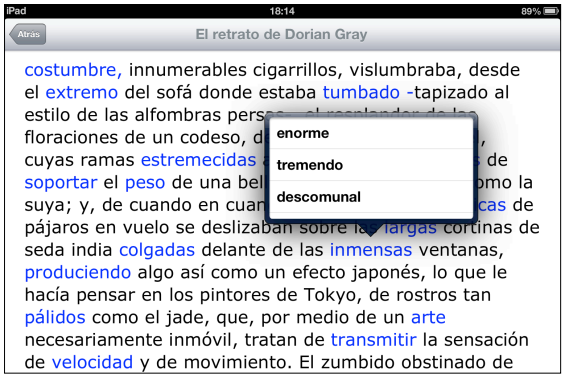

Figure 3: DysWebxia en un software de lectura para iPad.

Dado que no existe un perfil universal de un usuario con dislexia (Dickinson, Gregor, y Newell, 2002), en todas estas implementaciones las configuraciones iniciales se pueden personalizar según las preferencias personales de lectura.

Bibliografía

Baeza-Yates, R. y L. Rello. 2011. Estimating dyslexia in the Web. En *Proc. W4A 2011*, Hyderabad, India. ACM Press.

[5]Descargable en Google Play en: `https://play.google.com/store/apps/details?id=org.easyaccess.epubreader`

[6]El servidor se puede consultar en: `http://www.text4all.net/dyswebxia.html`

Bott, Stefan, Luz Rello, Biljana Drndarevic, y Horacio Saggion. 2012. Can Spanish be simpler? LexSiS: Lexical simplification for Spanish. En *Proc. Coling '12*, Mumbay, India, December.

Caldwell, B., M. Cooper, L. G. Reid, y G. Vanderheiden. 2008. Web content accessibility guidelines (WCAG) 2.0. *WWW Consortium (W3C)*.

Carrillo, M. S., J. Alegría, P. Miranda, y Sánchez Pérez. 2011. Evaluación de la dislexia en la escuela primaria: Prevalencia en español. *Escritos de Psicología*, 4(2):35–44.

Dickinson, A., P. Gregor, y A.F. Newell. 2002. Ongoing investigation of the ways in which some of the problems encountered by some dyslexics can be alleviated using computer techniques. En *Proc. ASSETS'02*, Edinburgh, Scotland. ACM Press.

Gregor, P., A. Dickinson, A. Macaffer, y P. Andreasen. 2003. Seeword: a personal word processing environment for dyslexic computer users. *British Journal of Educational Technology*, 34(3):341–355.

Gregor, Peter y Alan F. Newell. 2000. An empirical investigation of ways in which some of the problems encountered by some dyslexics may be alleviated using computer techniques. En *Proc. ASSETS'00*, ASSETS 2000, New York, NY, USA. ACM Press.

Interagency Commission on Learning Disabilities. 1987. *Learning Disabilities: A Report to the U.S. Congress*. Government Printing Office, Washington DC, U.S.

Kanvinde, G., L. Rello, y R. Baeza-Yates. 2012. IDEAL: a dyslexic-friendly e-book reader (poster). En *Proc. ASSETS'12*, Boulder, USA, October. ACM Press.

McCarthy, Jacob E. y Sarah J. Swierenga. 2010. What we know about dyslexia and web accessibility: a research review. *Universal Access in the Information Society*, 9:147–152.

Rello, L. y R. Baeza-Yates. 2012. Lexical quality as a proxy for web text understandability (poster). En *Proc. WWW '12*, Lyon, France. ACM Press.

Rello, L. y R. Baeza-Yates. 2013. Good fonts for dyslexia. En *Proc. ASSETS'13*, Bellevue, Washington, USA. ACM Press.

Rello, L., R. Baeza-Yates, S. Bott, y H. Saggion. 2013a. Simplify or help? Text simplification strategies for people with dyslexia. En *Proc. W4A '13*, Rio de Janeiro, Brazil. ACM Press.

Rello, L., R. Baeza-Yates, L. Dempere, y H. Saggion. 2013b. Frequent words improve readability and short words improve understandability for people with dyslexia. En *Proc. INTERACT '13*, Cape Town, South Africa.

Rello, L., R. Baeza-Yates, y H. Saggion. 2013. The impact of lexical simplification by verbal paraphrases for people with and without dyslexia. En *Proc. CICLing'13*. Springer Berlin Heidelberg.

Rello, L., R. Baeza-Yates, H. Saggion, y E. Graells. 2012. Graphical schemes may improve readability but not understandability for people with dyslexia. En *Proceedings of the NAACL HLT Workshop PITR'12*. Montreal, Canada.

Rello, L., S. Bautista, R. Baeza-Yates, P. Gervás, R. Hervás, y H. Saggion. 2013. One half or 50%? An eye-tracking study of number representation readability. En *Proc. INTERACT '13*, Cape Town, South Africa.

Rello, L., G. Kanvinde, y R. Baeza-Yates. 2012. Layout guidelines for web text and a web service to improve accessibility for dyslexics. En *Proc. W4A '12*, Lyon, France. ACM Press.

Rello, L., M. Pielot, M. C. Marcos, y R. Carlini. 2013. Size matters (spacing not): 18 points for a dyslexic-friendly Wikipedia. En *Proc. W4A '13*, Rio de Janeiro, Brazil. ACM Press.

Santana, V. F., R. Oliveira, L.D.A. Almeida, y M. Ito. 2013. Firefixia: An accessibility web browser customization toolbar for people with dyslexia. En *Proc. W4A '13*, Rio de Janeiro, Brazil. ACM Press.

Topac, V. 2012. The development of a text customization tool for existing web sites. En *Text Customization for Readability Symposium*, November.

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 209-212 recibido 30-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Bologna Translation Service: Improving Access To Educational Courses Via Machine Translation (system demonstration)

Bologna Translation Service: mejorando el acceso a los planes de estudios universitarios mediante la traducción automática (demostración del sistema)

Justyna Pietrzak, Elena García Berasategi, Amaia Jauregi Carrera
Eleka Ingeniaritza Linguistikoa S.L. (Elhuyar Group)
Zelai Haundi kalea 3, Osinalde industrialdea,
20170 Usurbil, Gipuzkoa, Spain
{j.pietrzak, e.garcia, a.jauregi}@elhuyar.com

Resumen: Se presenta una demostración del trabajo realizado en el proyecto Bologna Translation Service (BTS), un proyecto cofinanciado por la Unión Europea en el marco de "Information and Communications Technology Policy Support Programme" (ICT PSP) que ofrece la traducción automática de planes de estudios desde 7 idiomas diferentes (alemán, español, finés, francés, holandés, portugués y turco) al inglés, y desde inglés al chino mandarín. BTS ofrece traducción automática accesible on-line. La estructura completa del sistema incluye motores de traducción automática basada en reglas, basada en estadística, ambas combinadas y también un sistema de post-edición automática y manual.
Palabras clave: traducción automática, traducción automática estadística (SMT), traducción automática basada en reglas (RBMT), post-edición automática, post-edición manual

Abstract: This is a demonstration of the Bologna Translation Service (BTS), an EU-funded project (in the framework of Information and Communications Technology Policy Support Programme - ICT PSP) which specialises in the automatic translation of study programmes from 7 languages (German, Spanish, Finnish, French, Dutch, Portuguese, and Turkish) to English, and from English to Mandarin Chinese. At the core of the BTS framework there are several machine translation (MT) engines through which web-based translation services are offered. The fully integrated BTS architecture groups rule-based and statistical MT, their combination, and automatic and manual post-editing modules.
Keywords: machine translation, statistical machine translation (SMT), rule-based machine translation (RBMT), automatic post-editing, manual post-editing

1 Introduction

There is a continuing increasing need for universities and other higher educational institutions to provide course syllabi and other educational information in English. Access to this content plays a crucial role in the degree to which these institutions effectively attract students and, more importantly, has an impact on international profiling of universities, helping them to operate in a globalized environment.

The regulatory environments in the context of the Bologna Accords combined with budget constraints and limited human resources make it very difficult for higher educational institutions to deliver English documentation, which affects their capacity to promote their services internationally. Confronted with the European Credit Transfer System (ECTS) requirements, many of them now spend vast amounts of money and time providing traditional human translated documents.

2 Objectives

As European higher education and research are two pillars of the knowledge-based society,

ISSN 1135-5948

BTS[1] aims to provide a solution to this problem by offering a low-cost, web-based, high-quality machine translation (MT) service geared towards this specific use case. The project makes use of existing rule-based (RBMT) and statistical (SMT) technologies and tailors them in order to produce the best possible quality for syllabus translations.

The first phase of the project includes the automatic translation of syllabi from 7 languages (German, Spanish, Finnish, French, Dutch, Portuguese, and Turkish) to English and from English to Mandarin Chinese.

The BTS approach is to integrate existing MT components into a web-based collaboration framework, in which users with different roles (requester, reviewer, manager, post-editor...) participate on-line at different stages of the translation workflow. The basis is a SMT engine for all language pairs, which were further refined by adding in data from the educational domain and applying domain adaptations and automated and human post-editing. For a selected number of language pairs, systems were combined (SMT and RBMT combination) in order to further improve translation quality.

3 System description

BTS integrates different MT components in order to improve the quality of translation.

3.1 Statistical Machine Translation

For training and tuning the SMT systems, the freely available Moses (Koehn *et al.*, 2007) software tools and relevant wrapper scripts included in OpenMaTrEx (Dandapat *et al.*, 2010) were used. The phrase-based translation models were trained using an out of domain corpus of nearly 6 million segments and a 2 million segment in-domain one. GIZA++ (Och and Ney, 2003) was used for word-aligment with the default number of iterations for the implementations of IBM Models.

To build the language models (LM), we used the state-of-the-art open-source IRSTLM toolkit (Federico and Cettolo, 2007). The LMs were five- to seven-gram models, applying Kneser-Ney discounting and using word forms. Separate LMs were built for in-domain and out-of-domain corpora.

3.2 Rule-Based Machine Translation

Rule-based systems were developed for 4 language pairs. Only the Turkish to English system was developed entirely for the BTS project. The remainder were modifications and further development of machines that were already in use by Convertus and Eleka.

The state-of-the-art open-source platform for developing RBMT systems *Apertium* (Forcada *et al.*, 2011) was used for Spanish and Portuguese, and in-house RB system made by one of the project's partners, the Convertus company, was applied for Turkish to English and Finnish to English. RBMT models included the adaptation of linguistic data to the environment of the Bologna Translation Service: enrichment of dictionaries (by adding the terminology used in the academic and education fields), modification of bilingual dictionaries (in order to promote domain-specific translations), and adaptation of some transfer rules to syllabi writing style.

3.3 Automated post-editing

In order to keep on improving the quality of translation, the next developed modules were fine-tuned MT systems built upon the advanced SMT and RBMT systems.

SMT fine-tuning was achieved by means of rule-based automated post-editing (rbAPE); rbAPE modules included a pre-processing string-based layer focusing mostly on lexical and syntactic issues. The rbAPE rules were developed for every language pair, with an average of 200 transfer rules, and 1000 generation rules.

For those language pairs for which RBMT systems are used, statistical APE (sAPE) was developed, using the Moses tool. The translation in these cases was carried out from rule-based machine translation to the reference target translation, so to speak, from "not perfect SMT-produced English" into "better post-edited English". In every case, statistically significant improvements were achieved; nevertheless RBMT+sAPE systems

[1] "Bologna Translation Service" project has received funding from the European Community (ICT-PSP 4th Call) under Grant Agreement nº 270915. The official website of the project: http://www.bologna-translation.eu

obtained in general worse automatic evaluation than their respective SMT+rbAPE systems.

3.4 System combination

Automatic evaluation lead us to the conclusion that the output from the SMT systems is considerably better than that from the RBMT systems, but we expected further advances to be shown when such system types are combined. Using the CMU system combination toolkit (Hildebrand and Vogel, 2008), we conducted a series of experiments combining the RBMT and SMT engines for Spanish–, Portuguese–, Finnish– and Turkish–English language pairs. Experiments were conducted with different development corpora, LM and phrase-table combinations, searching for the best performing systems. On average, improvements of 2 BLEU points were achieved with statistically significant improvements.

3.5 Front-end, manual post-editing

All system elements were combined into a workflow, offering web-access to all translation models.

The platform was optimized following previous usability surveys conducted among potential users. In addition to previously mentioned components, the translation memory module was added (personalized and confidential for every user institution), as well as the manual post-editing module, to produce one centralized, fully-fledged MT and post-editing environment.

Service is available to be tested at demo.bologna-translation.eu.

4 Evaluation

For translation quality evaluation, we compiled results for the automatic MT evaluation metrics BLEU (Papineni *et al.*, 2002), METEOR (Banerjee and Lavie, 2005) and TER (Snover, 2006), using 2000 sentence test sets. Statistical significance testing was carried out using the tool 'Multeval' (Clark *et al.*, 2011), which implements approximate randomisation normalisation, which has been shown to be more reliable than bootstrap resampling (Koehn, 2004) to test the statistical significance of MT quality measurements.

Apart from automatic evaluation of the translation quality, three different types of human evaluation were carried out, each of which focused on a different aspect of translation. With the *quality evaluation*, the focus was on the linguistic quality of the translations. With these evaluations we try to answer the question 'how good is the translation?'. With the *productivity evaluation*, we investigate to what extent automated translation can speed up human translation. The question we are trying to answer here is 'is post-editing MT output quicker than translating from scratch?'. With the *usability evaluation*, the focus was on the utility of raw MT to end-users, trying to answer the question 'how useful is the MT output for gisting?'.

For the quality and productivity evaluations, the same test set was used. This test set was compiled by taking a random sample of 300 sentences from the test sets that were used for the automatic evaluations. The size of the set was reduced to keep the human evaluation within acceptable limits in terms of cost and time needed to complete them.

5 Results and further work

Good evaluation scores were obtained for every language pair, as reported in previously published paper (Van de Walle *et al.*, 2013), with an average of 20 BLEU point of improvements for MT systems, and 30% of productivity improvements, compared to human translation.

In order to develop the fully functional system, various universities were invited to take part in qualitative and quantitative surveys before the development of the platform, and their opinions were taken into account during and after prototype development.

MT quality is also constantly improving, due to the incorporation of translation memories and periodic retraining of translation models.

Another round of surveys and evaluations will be conducted in 2014 after the first year of working of BTS.

References

Banerjee, S., A. Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human

Judgments. In: *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*. University of Michigan, Ann Arbor, MI; pp. 65–72.

Clark, J., C. Dyer, A. Lavie, N. Smith. 2011. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In: *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Short papers*. Volume 2, Portland, OR; pp 176–181.

Dandapat, S., M. L. Forcada, D. Groves, S. Penkale, J. Tinsley, A. Way. 2010. OpenMaTrEx: A Free/Open-Source Marker-Driven Example-Based Machine Translation System. In: H. Loftsson (Red.), *Advances in Natural Language Processing: 7th International Conference on NLP. IceTAL 2010* (Reykjavík, 16-18 Aug. 2010) (Vol. Col. Lecture Notes in Artificial Intelligence, vol. 6233, pp. 121–126). Berlin/Heidelberg: Springer.

Federico, M., M. Cettolo. 2007. Efficient Handling of N-gram Language Models for Statistical Machine Translation. In: *Proc. of ACL Workshop on SMT*. Pp. 88–95, Prague, Czech Republic.

Forcada, M. L., M. Ginestí-Rosell, J. Nordfalk, J. O'Reagan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, F. M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. In: *Machine Translation*. 25:2(2011), pp 127–144.

Hildebrand A.S., S. Vogel. 2008. Combination of machine translation systems via hypothesis selection from combined n-best lists. In: *MT at work: Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas.* Pp. 254–261, Waikiki, Hawaii, October, 2008.

Koehn, P. 2004. Statistical significance tests for machine translation evaluation. In: *EMNLP-2004: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 25-26 July 2004, Barcelona, Spain; 8pp.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, Christine Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In: *ACL 2007: proceedings of demo and poster sessions*, Prague, Czech Republic; pp. 177–180.

Och, F. J., H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1): 19–51.

Och, F. J. 2003. Minimum error rate training in statistical machine translation. *ACL-2003: 41st Annual meeting of the Association for Computational Linguistics*, July 7-12, 2003, Sapporo, Japan; pp. 160–167.

Oflazer, K. 2008. Statistical Machine Translation into a Morphologically Complex Language. In: *Proceedings of CICLING 2008: Conference on Intelligent Text Processing and Computational Linguistics*, February 2008, Haifa, Israel; pp. 376–387.

Papineni, K., S. Roukos, T.d Ward & W-J. Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In: *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics*, Philadelphia, PA; pp. 311–318.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, L. and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In: *AMTA 2006: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, "Visions for the Future of Machine Translation"*, August 8-12, 2006, Cambridge, MA, USA; pp. 223–231.

Van de Walle, J., H. Depraetate, J. Pietrzak. 2013. Improving access to educational courses via machine translation: Bologna Translation Service evaluation. In: *Proceedings of the 5th international Conference on Education and New Learning Technologies*, Barcelona 1 – 3 July.

Proyectos

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 215-218 recibido 30-04-2013 revisado 16-06-2013 aceptado 21-06-2013

OpeNER: Open Polarity Enhanced Named Entity Recognition

OpeNER: Reconocimiento de entidades nombradas con polaridad

Rodrigo Agerri
IXA NLP Group
UPV/EHU
rodrigo.agerri@ehu.es

Montse Cuadros Seán Gaines
HSLT, IP department
Vicomtech-IK4
{mcuadros,sgaines}@vicomtech.org

German Rigau
IXA NLP Group
UPV/EHU
german.rigau@ehu.es

Resumen: Actualmente existe una gran cantidad de empresas ofreciendo servicios para el análisis de contenido y minería de de datos de las redes sociales con el objetivo de realizar análisis de opiniones y gestión de la reputación. Un alto porcentaje de pequeñas y medianas empresas (pymes) ofrecen soluciontes específicas a un sector o dominio industrial. Sin embargo, la adquisición de la necesaria tecnología básica para ofrecer tales servicios es demasiado compleja y constituye un sobrecoste demasiado alto para sus limitados recursos. El objetivo del proyecto europeo OpeNER es la reutilización y desarrollo de componentes y recursos para el procesamiento lingüístico que proporcione la tecnología necesaria para su uso industrial y/o académico.
Palabras clave: Reconocimiento y Desambiguación de Entidades Nombradas, Coreferencia, Análisis de Sentimiento

Abstract: Currently there are a many companies offering Content Analytics and Social Internet Mining services for the purposes of Opinion Mining and Reputation Management. A high percentage of Small and Medium Enterprises (SMEs) are active offering niche solutions to specific segments of the market and/or domains. However, acquiring or developing the base qualifying technologies required to enter the market is an expensive undertaking that redirects the already limited resources of SMEs away from offering products and services that the market demands.The main goal of the OpeNER european project is the reuse and repurposing of existing language resources and data sets to provide a set of underlying technologies to the broader industrial and academic community.
Keywords: Named Entity Recognition and Disambiguation, Coreference, Opinion Mining

1 Introduction

Customer reviews and ratings on the Internet are increasing importance in the evaluation of products and services by potential customers. In certain sectors, it is even becoming a fundamental variable in the purchase decision. A recent Forrester study showed more than 30% of Internet users have evaluated products online, and that 70% of those studied end user generated reviews[1]. Furthermore, another study by Complete Incorporated for the Tourist Domain showed that more than 80% of users preferred other users' opinions in order to make their buying decisions. In fact, it has been concluded that 97% of Internet users have read and been influenced by other users' opinions while planning a trip (Gretzel and Yoo, 2008). Obviously, this trend will continue with the growth of Social Media and access to Information and Communication Technologies (ICT). Consumers tend to trust the opinion of other consumers, especially those with prior experience of a product or service, rather than company marketing (see footnote 1). The role of user comments is of particular importance when there is little differentiation between the product and services on offer. Therefore, there is an objective necessity to manage and understand the knowledge conveyed by opinions. Opinion Mining consists of extracting and analysing, from unstructured text, opinions about products, people, events, institutions, etc. (Pang and Lee, 2008). In other words, the goal is to know "who" is speaking about "what", "when" and in "what sense" (Hu and Liu, 2004). More specifically, OpeNER will stress the importance of providing a good Name Entity Resolution system (Named Entity Recognition or NERC, Coreference and

[1] http://www.bazaarvoice.com/resources/stats

ISSN 1135-5948

Named Entity Disambiguation or NED) to feed the feature-based opinion mining systems with relevant information with respect to the entities about which the opinions are being expressed.

Currently there are a many companies offering Content Analytics and Social Internet Mining services for the purposes of Opinion Mining and Reputation Management. A high percentage of Small and Medium Enterprises (SMEs) are active offering niche solutions to specific segments of the market and/or domains. However, acquiring or developing the base qualifying technologies required to enter the market is an expensive undertaking that redirects the already limited resources of SMEs away from offering the products and services that the market demands.

The main goal of the OpeNER project[2] is the reuse and repurposing of existing language resources and data sets to provide a set of underlying technologies to the broader community. OpeNER will focus on the provision of a supplementary sentiment lexicon with culturally normalised and graduated values. NERC will also be addressed leveraging Linked Data with the aim of disambiguating the entity types recognised for the languages covered in the project: Spanish, English, French, German, Dutch and Italian. In the first year the project will be focused on a generic application domain, and later, adapted to the Tourism domain.

This will be achieved in conjunction with and End User Advisory Board (EUAB) composed of European Tourism Promotion Agencies, an online Tourism Portal and other interested parties. Furthermore, OpeNER will employ proven software from the Open Source community and develop an online development community thus ensuring long term viability beyond the project timeframe. In that way the benefits of the project will be adopted and extended to new domains and languages, OpeNER will strive to make the tools and techniques resulting from the project available under Open Source or Hybrid Licenses.

2 Objectives

OpeNER aims to provide enterprise and society with base technologies for Cross-lingual Named Entity Recognition and Classification and Sentiment Analysis through the reuse of existing resources and the open development of complementary technologies. The key objectives of the project are the following: (i) Repurposing and/or developing of existing language resources and generation of a reference generic multilingual sentiment lexicon with cultural normalisation and scales; (ii) An extension lexicon for the Tourist sector in several languages (Spanish, Dutch, German, Italian, English and French); (iii) Named Entity Resolution (NERC, NED and Coreference) in the same set of target languages as the Sentiment Lexicon which is extensible to other languages by leveraging multilingual resources such as Wikipedia and Linked Data[3] resources such as DBpedia[4], etc; (iv) Development and open availability of validated reference Sentiment and Opinion Mining techniques and tools based on the results of the project; (v) Evaluation and Application of the project results in the cloud, principally in the tourism sector, with leading SMEs in the sector and with the support of several stakeholders as part of the End User Advisory Board (EUAB); (vi)Research and trialling of models that will ensure that the project results are self-sustainable and economically viable in the long term; (vii) Achievement of the projects objectives by repurposing and leveraging existing state of the art and established language resources.

3 Work Plan

In order to optimise the value of OpeNER technology, all the requirements along the value chain for the development and the exploitation of the project's objectives are directly represented in the project's Consortium, formed by 6 institutions from Italy, The Netherlands and Spain, with Vicomtech-IK4 as coordinator. The OpeNER Work Plan is structured in 8 Work Packages (WP), and can be divided in three blocks. Although we first describe every WP we will henceforth focus on the most relevant aspects to SEPLN, namely, those related to work packages 4-7:

1. **Management, Dissemination and Exploitation**: WP1 and WP8 led by Vicomtech-IK4. As this is an SME oriented project, the Dissemination and Exploitation of results will go beyond

[2] http://www.opener-project.org

[3] http://linkeddata.org

[4] http://dbpedia.org

the publication of scientific articles. It shall include industrial dissemination and exploitation also.

2. **System Design and Deployment**: WP2 and WP7 led by Synthema and Olery respectively. In order to ensure the future exploitation of the project by SMEs, the system design and deployment is crucial. Both Synthema and Olery have experience in software integration for industry related applications and products.

3. **NLP and Web techniques**: WP3 (Universidad del País Vasco/Euskal Herriko Unibertsitatea, UPV/EHU), WP4 (Consejo Nacional de Investigación de Italia, CNR), WP5 (Universidad Libre de Amsterdam, VUA) and WP6 (Vicomtech-IK4). Focused on Opinion Mining (WP5) and Named Entity Resolution (WP3) and any other basic NLP (WP6) and Web tools (WP4) required to perform those tasks.

OpeNER will provide language analysis tool chains for several languages to help researchers and companies make sense out of unstructured text via Natural Language Processing. It will consist of easy to install, improve and configure components to: (i) Detect the language of a text; (ii) Determine polarity of texts (sentiment analysis) and analysis of feature-based opinions; and (iii) Detect and classify the entities named in the texts and link them together via Named Entity Recognition, Coreference and Named Entity Disambiguation (e.g. President Obama or The Hilton Hotel). Besides the individual components, guidelines exists on how to add languages and how to adjust components for specific situations and topics. The following section will describe the English and Spanish OpeNER toolchains.

4 OpeNER NLP Pipelines

An OpeNER tool chain or pipeline consists of a broad mix of technologies glued together using Ruby. The prerequisites for running an OpeNER tool chain are the following: A GNU/Linux or Unix type operativing system (including MAC OS), Ruby 1.9.3+, Python 2.7+, Java 1.7+, Perl 5+. Every part of the OpeNER tool chain has individual dependencies, most of which are included in the components themselves. The OpeNER architecture consists of several building blocks called *components*, which can be used to build a tool chain called a *configuration*,

A *component* consists of a *kernel* which can be for example a POS tagger implemented in Java and a *glue* in Ruby to connect with other components. Figure 1 represents a possible flow of information between several components. Each component is configured to take the information it requires to perform a specific analysis from the previous module. KAF(Bosma et al., 2009)[5] is used as inter-component representation between the components. Each of the tool chains built are then deployed via Cloud Computing services such as Amazon Elastic Computing Cloud[6] (Amazon EC2).

As described in section 3, the NLP focus of OpeNER is on Named Entity Resolution (NERC, Coreference and NED) and Opinion Mining. The overall objective of Named Entity Resolution is to be able to recognise, classify and link every mention of a specific named entity in a text. A named entity can be mentioned using a great variety of surface forms (Barack Obama, President Obama, Mr. Obama, B. Obama, etc.) and the same surface form can refer to a variety of named entities: for example, the form 'San Juan' can be used to ambiguously refer to many toponyms, persons, a saint, etc. (e.g. see http://en.wikipedia.org/wiki/San_Juan). Furthermore, it is possible to refer to a named entity by means of anaphoric pronouns and coreferent expressions such as 'he', 'her', 'their', 'I', 'the 35 year old', etc. Therefore, in order to provide an adequate comprehensive account of named-entities in text it is necessary to recognise the mention of a named-entity, to classify it as a type (e.g. person, location, etc.), to link it or disambiguate it to a specific entity, and to resolve every form of mentioning or co-referring to the same entity in a text.

The *Opinion Mining* approach in OpeNER consists of three levels: (i) generation of polarity lexicons from WordNets for each language (ii) development of polarity systems at document and sentence level based on those lexicons and (iii) feature-based opinion mining based on supervised classification and feature extraction. For

[5] http://www.kyoto-project.eu

[6] http://aws.amazon.com/ec2/

hotel reviews we will be looking at *features* such as room service, cleanliness, etc.

As we are working with 6 languages, it would be convenient, where possible, to use *one solution*, *one tool* and *one methodology* to provide most of the NLP annotation, including not only NERC, Coreference and NED, but also language identification, tokenisation, POS tagging, lemmatisation, and parsing. Otherwise, maintaining so many different tools for every annotation process would be far too cumbersome to provide easy-to-use integrated NLP pipelines in a virtual machine. Thus, every NLP component (except Opinion Mining) in the English and Spanish pipelines are being developed using the Apache OpenNLP API[7] for supervised Machine Learning based linguistic annotators: Sentence Segmentation, Tokenisation, Part of Speech tagging, NERC and Constituent Parsing. The Consortium is training new models for every component using the usual general domain datasets such as CoNLL and Evalita datasets for NERC, Penn Treebank WSJ for English POS and Constituent Parsing, Ancora[8] for Spanish POS and Constituent Parsing. Furthermore, lemmatisation is performed using word form and POS tags lookups in a dictionary for each language.

With respect to coreference, the Stanford multi-sieve pass system (Lee et al., 2013) is being re-implemented in such a manner that it facilitates its adaptation to other languages. The coreference module takes KAF containing POS and NERC annotated tokens and a constituent parse tree as input. The sieves are implemented in a way that the only requirements to adapt to another language is to change the POS and Parsing tagsets and a number of static dictionaries that contains information such as demonyms, gender, number, etc. Finally, the NED systems are being adopted from the English DBpedia Spotlight[9] (Mendes et al., 2011) which is based on DBpedia and Wikipedia for disambiguation of Named Entities.

5 Concluding Remarks

This paper presents OpeNER, a European project that will provide completely 'off-the-box' usable language analysis tool chains for six languages to make sense out of unstructured text via Natural Language Processing. These chains will easily be incorporated by SMEs in their workflow for applications such as Reputation Management and Information Access. Furthermore, on the second year of the project the toolchain will be adapted to process texts from the Tourist domain. To this purpose, the project will also investigate how the performance of the OpeNER toolchains can be improved by inter-relating with each other the various layers of linguistic annotation.

[7] http://opennlp.apache.org/

[8] http://clic.ub.edu/corpus/es/ancora

[9] https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki

Acknowledgments

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement n. 296451

References

Bosma, W., P. Vossen, A. Soroa, G. Rigau, M. Tesconi, A. Marchetti, M. Monachini, and C. Aliprandi. 2009. Kaf: a generic semantic annotation format. In *Proceedings of the Generative Lexicon (GL2009) Workshop on Semantic Annotation*, Pisa, Italy.

Gretzel, Ulrike and Kyung Hyan Yoo. 2008. Use and impact of online travel reviews. In *Information and communication technologies in tourism 2008*. Springer, page 3546.

Hu, M. and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, page 168177.

Lee, Heeyoung, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, pages 1–54, January.

Mendes, P. N., M. Jakob, A. Garca-Silva, and C. Bizer. 2011. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, page 18.

Pang, B. and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1135.

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 219-222 recibido 29-04-2013 revisado 16-06-2013 aceptado 21-06-2013

LEGOLANG: Técnicas de deconstrucción aplicadas a las Tecnologías del Lenguaje Humano

LEGOLANG: Deconstruction Techniques applied to Human Language Technologies

P. Martínez-Barco, A. Ferrández, D. Tomás, E. Lloret, E. Saquete, F. Llopis, J. Peral, M. Palomar, J.M. Gómez
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
{patricio,antonio,dtomas,elloret,stela,llopis,jperal,mpalomar,jmgomez}@dlsi.ua.es
M.T. Romá-Ferri
Departamento de Enfermería
Universidad de Alicante
mtr.ferri@ua.es

Resumen: El objetivo de este proyecto se basa en la necesidad de replantearse la filosofía clásica del TLH para adecuarse tanto a las fuentes disponibles actualmente (datos no estructurados con multi-modalidad, multi-lingualidad y diferentes grados de formalidad) como a las necesidades reales de los usuarios finales. Para conseguir este objetivo es necesario integrar tanto la comprensión como la generación del lenguaje humano en un modelo único (modelo LEGOLANG) basado en técnicas de deconstrucción de la lengua, independiente de su aplicación final y de la variante de lenguaje humano elegida para expresar el conocimiento.
Palabras clave: tecnologías del lenguaje humano (TLH), comprensión del lenguaje, generación del lenguaje, desconstrucción del lenguaje

Abstract: The main objective of this project is based on the need to reconsider the clasical HLT philosophy to adapt it, not only to the currently available resources (unstructured data with multimodality, multilinguality and different levels of formality) but also to the real needs of the final users. In order to reach this objective it is necessary to include the understanding as well as the generation of human language in a unique model (LEGOLANG model) based on language deconstruction techniques, independently of the final application and the human language variant chosen to express the knowledge.
Keywords: human language technologies (HTL), human language understanding, human language generation, language deconstruction

1 Introducción

Se conoce como generación del lenguaje natural (GLN) al proceso de construcción deliberada de texto en lenguaje natural con el fin de alcanzar capacidades comunicativas previamente especificadas (McDonald, 1987). Con este objetivo, la GLN se convierte en elemento indispensable para múltiples aplicaciones que derivan en fines más concretos como la construcción automática de informes estandarizados, la producción automática de resúmenes, la traducción automática, etc. De esta manera, y tomando como base la definición de Reiter & Dale (1997), podemos hablar de GLN como una línea de investigación en el ámbito de las Tecnologías del Lenguaje Humano (TLH), cuyo fin último es el de proporcionar un conjunto de herramientas y técnicas capaces de producir texto comprensible en lenguaje natural a partir de una representación no lingüística de la información, generalmente, desde bases de datos o fuentes de conocimiento.

Otra de las grandes líneas que emanan de las TLH es la compresión del lenguaje natural (CLN) que trata de extraer, de manera automática, el significado de un texto dado y obtener una representación estructurada del mismo para su uso posterior. Así, GLN y CLN podrían llegar a entenderse como grandes

ISSN 1135-5948

procesos de análisis simétricos. Sin embargo, la investigación tradicional de TLH ha disociado sus líneas en las dos grandes ramas citadas, GLN y CLN, dando lugar a un conjunto de aproximaciones diferentes, que si bien parten de hipótesis teóricas comunes, sus realizaciones finales distan en muchos casos de ser compatibles (Reiter & Dale, 2000).

Además, la nueva situación implica que los sistemas de GLN deben acometer la captura de información desde colecciones documentales no estructuradas, multilingües y multimodales, con escasas garantías de fiabilidad y diversos grados de formalidad, provenientes de fuentes tan dispersas y diversas como artículos periodísticos, informes técnicos, blogs, microblogs, wikis o redes sociales. Esta situación deriva en un problema aún sin resolver, y es que no existe un modelo único de comprensión y generación del lenguaje independiente de la aplicación.

2 *Propuesta*

Nuestra propuesta consiste en plantear una concepción del lenguaje humano totalmente novedosa en la que se descontextualizará el concepto de deconstrucción para redefinirlo en el marco de las Tecnologías del Lenguaje Humano, como un modelo que permitirá descomponer textos conocidos en un caos de unidades básicas de conocimiento (fase de comprensión del lenguaje) que, mediante la apropiada definición de nuevas estructuras, volverá a combinarse para proporcionar nuevos conocimientos (fase de generación del lenguaje). La deconstrucción, así entendida, nos permitirá modelar una nueva metodología para la generación de un lenguaje humano no tan centrado en la definición de estructuras gramaticales correctas sino de estructuras prácticas que muestren al receptor nuevos conocimientos ocultos en los documentos originales.

En consecuencia, este proyecto persigue tres metas fundamentales:

1. La definición de una unidad básica de conocimiento orientada al GLH a la que denominaremos L-Brick (Language Brick, o Ladrillo de Lenguaje).
2. El modelado del proceso de deconstrucción que, a partir de una colección documental, debe ser capaz de generar la representación del mismo en un sistema caótico de L-Bricks, definiendo el conjunto de recursos y técnicas útiles para dotar de contenido necesario a esas estructuras.
3. El rediseño de las tareas de los sistemas clásicos de GLN en función de los L-Bricks y de sus reglas de composición, de tal manera que permitan definir nuevas formas de comunicación del conocimiento, tomando como única base la información contenida en ellos.

En la Figura 1 se puede ver gráficamente la propuesta de este proyecto.

3 *Antecedentes*

Como se ha comentado previamente, la investigación clásica en Tecnologías del Lenguaje Humano ha tratado a las técnicas de generación de lenguaje natural (GLN) de manera aislada respecto a las de comprensión (CLN). En concreto, el estudio de la gran mayoría de trabajos iniciales de GLN (Bernardos, 2007) coincide en considerar una arquitectura común cuyo origen es siempre una representación computacional de la información, y un posterior procesamiento de la misma basado en dos grandes niveles (Hovy, 2000): a) determinación del qué decir y b) determinación del cómo decirlo. Para abordar estos dos niveles, Reiter E. (1994) propone tres fases diferentes: Macroplanning (para el qué decir), Microplanning (para el cómo decirlo), y Realización (para decirlo).

Además, en los últimos años ha surgido un

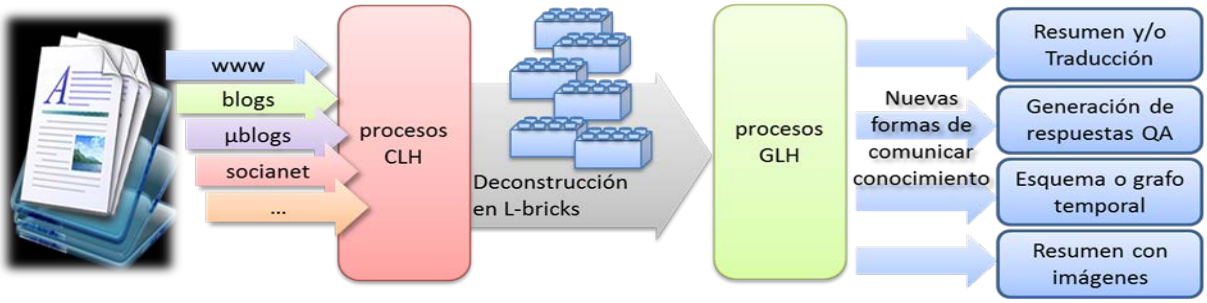

Figura 1: Propuesta del proyecto

creciente interés por abordar el problema de la generación del lenguaje no únicamente desde la componente "natural" (GLN, generación de lenguaje textual, formal y sintácticamente correcto), sino en general, desde la vertiente "humana" (GLH, generación de cualquier tipo de lenguaje para comunicación entre humanos). Cabe destacar, por poner algunos ejemplos, conferencias específicas como Generation Instructions in Virtual Environments (GIVE, 2011), Workshop on Multimodal Output Generation (MOG, 2011) y Generation Challenges (GC, 2011), donde se han presentado múltiples aproximaciones centradas en esta idea.

4 Metodología y plan de trabajo

La ejecución del proyecto se ha estructurado en las capas que se muestran en la Figura 2.

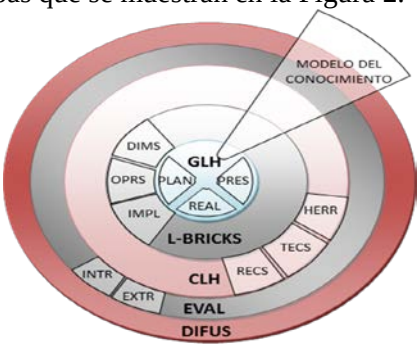

Figura 2. Estructuración modular del proyecto

La capa central, GLH (generación del lenguaje humano) es la motivación principal del proyecto, es el fin a alcanzar, conseguir generar lenguaje. Por encima de ella, la capa L-BRICKS (deconstrucción en unidades básicas de conocimiento) es una representación intermedia básica para poder construir la GLH. A su vez, la capa de CLH (comprensión del lenguaje humano) prestará sus servicios para dotar de contenido a la capa L-BRICKS. La capa EVAL (evaluación) recoge todas las actividades que nos permitirán conocer la validez de las propuestas de las capas internas. Finalmente, la capa más externa corresponde a las actividades de DIFUS (difusión) representando la piel del proyecto, es decir, lo que se va a dar a conocer de todas las actividades internas.

Por otra parte, el plano transversal (modelo del conocimiento) cruza las fronteras de todas las capas con el fin de identificar los hilos conductores comunes a todas aquellas tareas relacionadas con una misma motivación. Las actividades transversales son redundantes a las anteriores, pero sirven para dar coherencia al modelo desde una perspectiva diferente.

A continuación vamos a especificar un poco más en detalle las actividades y los hitos de cada una de estas capas.

Capa CLH: Comprensión del lenguaje humano

En esta capa se analizarán, recopilarán, adaptarán e integrarán todos los recursos, técnicas y herramientas necesarias para transformar la información obtenida desde diferentes fuentes en conocimiento útil que posteriormente se almacenará en las unidades básicas de conocimiento a través de tres actividades: a) *CLH.RECS: Recursos*, cuyo objetivo es la obtención y puesta a disposición del proyecto del conjunto de recursos necesarios para las tareas de comprensión del lenguaje, b) *CLH.TECS: Técnicas*, que se encargará de la recopilación e investigación del conjunto de técnicas necesarias para las tareas de comprensión del lenguaje, y c) *CLH.HERR: Herramientas*, cuyo hito es la obtención, implementación e integración del conjunto de herramientas necesarias para las tareas de comprensión del lenguaje.

Capa L-BRICKS: Deconstrucción del lenguaje en unidades básicas de conocimiento

En esta capa se tratarán las actividades relacionadas con la definición, estructuración e inserción de datos en estas unidades, denominadas L-Brick (language brick: ladrillo del lenguaje) por el paralelismo generado con las unidades de los juegos infantiles para construcción en bloques: a) *L-BRICKS.DIMS: Dimensiones*, en la que se definirá la estructura multidimensional del ladrillo, b) *L-BRICKS.OPRS: Operaciones*, cuyo objetivo es la planificación del conjunto de operaciones posibles en la unidad L-Brick, y c) *L-BRICKS.IMPL: Implementación*, cuya finalidad es la implementación computacional de la estructura, operaciones y almacenamiento del L-Brick.

Capa GLH: Generación del lenguaje humano

El objetivo de esta capa es la generación de lenguaje humano a partir de los L-Bricks, que son nuestras unidades básicas de conocimiento.

Por tanto, en esta capa, nos centraremos principalmente en el análisis de técnicas y herramientas para poder comunicar la información contenida en los L-Bricks, que se corresponde con la etapa de realización del proceso de GLH. Las actividades de esta capa son las siguientes: a) *GLH.PLAN: Planificación*, cuyo hito es la obtención de técnicas para planificar la presentación del conocimiento, b) *GLH.REAL: Realización*, cuya finalidad es la obtención de técnicas para la realización del conocimiento contenido en los L-Bricks, y c) *GLH.PRES: Presentación*, cuyo objetivo es la definición del modelo de presentación del conocimiento del L-Brick.

Capa EVAL: Evaluación

La capa EVAL contempla las actividades necesarias para la realización de la evaluación del proyecto en dos niveles: a*) EVAL.INTR: Evaluación intrínseca*, que pretende llevar a cabo el análisis y definición de una serie de métricas cualitativas y cuantitativas que permitan evaluar intrínsecamente el modelo de GLH definido, y b) *EVAL.EXTR: Evaluación extrínseca*, que plantea la definición de un escenario sobre el cual aplicar el modelo de GLH para realizar una evaluación extrínseca del mismo.

Capa DIFUS: Difusión de la investigación

El objetivo de esta actividad es la difusión de los resultados científicos y tecnológicos alcanzados durante el desarrollo del proyecto. Si bien la finalidad última es la definición del modelo basado en L-Brick y su explotación para la generación de lenguaje, existen numerosas tareas intermedias para llevar a cabo este objetivo que por sí mismas resultan de interés para la comunidad científica.

Plano transversal: Modelo de conocimiento

El plano transversal representa un conjunto de actividades relacionadas con la representación del modelo de conocimiento para cada uno de los niveles de análisis del lenguaje (léxico, sintáctico, semántico, pragmático), desde su comprensión en el origen hasta la generación en el destino, atravesando todas las capas de la arquitectura. Este plano actuará como mecanismo cruzado de cohesión entre las capas y de esta manera se garantiza que únicamente se realizarán esfuerzos en tareas de CLH cuando realmente tengan utilidad para tareas de GLH, y a su vez, todas las tareas de GLH obtengan el conocimiento necesario desde las tareas de CLH.

5 Agradecimientos

El proyecto LEGOLANG[i] está financiado por el Ministerio de Economía y Competitividad con número de referencia TIN2012-31224.

Bibliografía

Bernardos, S. 2007. ¿Qué es la generación de lenguaje natural? Una visión general sobre el proceso de generación. *Revista Iberoamericana de Inteligencia Artificial, 34*, 105-128.

GC. 2011. *Generation Challenges*. Obtenido de http://www.nltg.brighton.ac.uk/research/genchal10/

GIVE. 2011. *Generation Instructions in Virtual Environments*. Obtenido de http://www.give-challenge.org/research/

Hovy, E. 2000. Language Generation (article 86). En E. Reilly, A. Ralston, & D. Hemmendinger, *Encyclopedia of Computer Science*. London: McMillan.

McDonald, D. D. 1987. Natural Language Generation. En S. C. Shapiro, *Encyclopedia of Artificial Intelligence* (págs. 642-655). John Wiley and Sons.

MOG. 2011. *Workshop on Multimodal Output Generation* . Obtenido de http://www.mog-workshop.org/

Reiter, E. 1994. Has a Consensus NL Generation. *Proceedings of the 7th International Workshop on Natural Language Generation*, (págs. 163-170). Kennebunkport.

Reiter, E., & Dale, R. 1997. Building applied natural language generation systems. *Natural Language Engineering, 3*, 57-87.

Reiter, E., & Dale, R. 2000. *Building natural language generation systems*. Cambridge: Cambridge University Press.

[i] http://gplsi.dlsi.ua.es/proyectos/legolang/

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 223-226 recibido 05-04-2013 revisado 16-06-2013 aceptado 21-06-2013

DIANA: Análisis del discurso para la comprensión del conocimiento

DIANA: DIscourse ANAlysis for knowledge understanding

Paolo Rosso
Universitat Politècnica de València
Camino de Vera s/n. 46022
Valencia, España
prosso@dsic.upv.es

M. Antònia Martí, Mariona Taulé
Universitat de Barcelona
Gran Via 585, 08007
Barcelona, España
{amarti, mtaule}@ub.edu

Resumen: DIANA es un proyecto coordinado en el que participan el grupo de *Ingeniería del Lenguaje Natural y Reconocimiento de Formas* (ELiRF) de la Universitat Politècnica de València y el grupo *Centre de Llenguatge i Computació* (CLiC) de la Universitat de Barcelona. Se trata de un proyecto del programa de I+D (TIN2012-38603) financiado por el Ministerio de Economía y Competitividad. Paolo Rosso coordina el proyecto DIANA y lidera el subproyecto DIANA-Applications y M. Antònia Martí lidera el subproyecto DIANA-Constructions.
Palabras clave: Semántica distribucional, construcciones, detección de ironía y de fraude en medios sociales, detección de paráfrasis y plagio

Abstract: DIANA is a coordinated Project involving the research group of *Ingeniería del Lenguaje Natural y Reconocimiento de Formas* (ELiRF) of the Universitat Politècnica de València and the research group of *Centre de Llenguatge i Computació* (CliC) of the Universitat de Barcelona. This is an R&D project (TIN2012-38603) funded by the Spanish Ministry of Economy and Competitiveness. Paolo Rosso coordinates the DIANA project and leads the subproject DIANA-Applications and M. Antònia Marti leads the DIANA-Constructions subproject.
Keywords: Distributional semantics, constructions, irony and deception detection in social media, paraphrasing and plagiarism detection

1 Introducción

La finalidad de este proyecto es avanzar en el área de la Lingüística Computacional (LC) y el Procesamiento del Lenguaje Natural (PLN) con el fin de superar las actuales limitaciones de los sistemas en una doble línea de actuación: (i) Desde la LC proponiendo y evaluando empíricamente nuevos fundamentos teóricos en la concepción de la estructura del lenguaje humano; (ii) desde el PLN desarrollando nuevas técnicas y métodos aprovechando el estado actual de los conocimientos científico-técnicos. Estas nuevas aproximaciones se aplicarán, por un lado, en tareas que impliquen el tratamiento del lenguaje en el marco del discurso (resolución de la correferencia de entidades y eventos, tratamiento de los argumentos implícitos e identificación de paráfrasis en el plagio) y, por otro lado, en el análisis y comprensión de textos subjetivos. Nuestro propósito es dar respuesta a la pregunta que se plantea Wintner (2009) en su artículo 'What Science Underlies Natural Language Engineering?' en el que expone un problema fundamental: 'What branch of science, then, underlies Natural Language Engineering? What is the theoretical infrastructure on which we build our applications? And what kind of mathematics is necessary for reasoning about human languages? (...)'.

Desde la lingüística cognitiva (Croft y Cruse, 2000), la neurociencia (Hawkins, 2004) y la psicología del desarrollo (Tomasello, 2003) se proponen alternativas a lo que constituyen los fundamentos de la estructura lingüística. De esta conjunción de factores emerge un nuevo modelo de lenguaje que toma como base las producciones lingüísticas para inferir la estructura de la lengua. Este nuevo modelo tiene el concepto de construcción como unidad

ISSN 1135-5948

básica (Goldberg, 1995) y el discurso como marco del análisis. Nuestro objetivo es complementar el conocimiento lingüístico tradicional -análisis morfológico (PoS), sintáctico y semántico- con la identificación de construcciones, unidades complejas de tipo sintagmático que corresponden de manera aproximada a las unidades que en PLN se tratan bajo el nombre de expresiones multipalabra (EM). Se entiende por construcción, la conjunción de una forma y un significado que expresan una determinada función comunicativa. Sag et al. (2001) incluyen dentro de las construcciones o EM tanto expresiones fijas, semifijas (locuciones, compuestos nominales, nombres propios), como expresiones sintácticamente flexibles (construcciones verbo-partícula, locuciones descomponibles, construcciones con verbos 'light', expresiones estereotipadas). Se parte de la hipótesis de que las construcciones son unidades lingüísticas que almacenamos en memoria a las que se accede tanto en la producción como en la comprensión de lenguaje. Jackendoff (1999) estima que el 50% del léxico de los hablantes está constituido por este tipo de construcciones.

Esta concepción del lenguaje conlleva la necesidad de explorar nuevos caminos desde la LC (modelos formales computacionalmente tratables) y desde el PLN (desarrollo de técnicas y métodos) que superen el umbral de calidad al que han llegado las aproximaciones que se han desarrollado hasta este momento.

2 Objetivos

El objetivo general del proyecto DIANA es desarrollar nuevas herramientas de análisis lingüístico coherentes con las nuevas formulaciones teóricas y las necesidades aplicadas desde el PLN y ensayar nuevas aproximaciones en la hibridación de estas herramientas y nuevos recursos para conseguir unos mejores resultados en el análisis del lenguaje. La consecución de este objetivo general se concreta en los siguientes objetivos más específicos:

1. Desarrollo de sistemas híbridos de PLN que combinen conocimiento lingüístico (corpus anotados a diferentes niveles) con técnicas de aprendizaje automático incorporando medidas de similitud semántica que extiendan el conocimiento disponible a casos no tratados.
2. Aplicación de estas tecnologías para superar las actuales limitaciones de los sistemas de resolución de la correferencia, la identificación de paráfrasis y estructura argumental. En concreto, en resolución de la correferencia nos centraremos en el tratamiento de los sintagmas nominales no coincidentes, la identificación de eventos correferentes y argumentos implícitos (Peris, Taulé y Rodríguez, 2013). En cuanto a la paráfrasis y el plagio, nuestro objetivo es ampliar el campo de tratamiento abordando el problema de su detección a partir de estrategias de análisis del discurso (Barrón-Cedeño et al., 2013).
3. Aplicación de estas tecnologías para la identificación de construcciones lingüísticas con el objetivo de mejorar el análisis y comprensión de textos subjetivos, para la identificación de estados de ánimo (estrés, frustración, depresión, neurosis y agresividad), para la detección de pedofilia y acoso en los medios sociales de comunicación (Bogdanova et al., 2013), así como de engaño en los mismos (detección de opiniones fraudulentas artificialmente construidas (Hernández et al., 2013)). Se desarrollarán aplicaciones basadas en este tipo de conocimiento.
4. Inferencia de construcciones lingüísticas representativas en textos figurados de los medios sociales de comunicación para la interpretación de su verdadero sentido. En especial, el reconocimiento de humor y la identificación de ironía en opiniones (Reyes y Rosso, 2012).

3 Metodología

El proyecto dará como resultado un entorno de PLN para el tratamiento de textos a nivel discursivo que detectará construcciones de carácter general y específicas de determinados dominios semánticos, con el objetivo de avanzar en la resolución de la correferencia, detección de paráfrasis y plagio, y en la identificación de la actitud subjetiva en los textos. Las herramientas del entorno serán independientes de la lengua y dispondrán de una interfaz para su extensión a diferentes formatos. Este entorno de PLN tendrá asociadas diferentes estructuras de datos lingüísticos que se obtendrán durante el desarrollo del proyecto: léxicos de construcciones y grafos de palabras semánticamente relacionadas tanto de carácter

general como específicos de los dominios tratados (Franco-Salvador et al., 2013).

Se aplicarán diferentes técnicas de aprendizaje automático para el desarrollo de un *parser* semántico. Se aplicarán modelos geométricos (*Vector Space Models*) para la representación semántica a partir del contexto basándonos en las propuestas de Turney y Pantel (2010), Padó y Lapata (2007) y Baroni y Lenci (2010). Para la inferencia y generalización de construcciones se utilizarán técnicas de inducción, generalización y jerarquización de patrones (Zuidema, 2007; Gries 2003). Esta metodología se refleja analíticamente en los puntos que detallamos a continuación:

a) Recopilación de los corpus disponibles y de corpus de nueva creación.
b) Estandarización de los corpus: formato común en XML con etiquetas identificadoras.
c) Procesamiento de los corpus a nivel básico con las herramientas disponibles.
d) Aplicación de medidas de semántica distribucional para la extracción de relaciones de similitud entre palabras basados en diferentes tipos de modelización del contexto. Nos centraremos en los predicados nominales y verbales.
e) Extensión de la información contenida en los léxicos AnCora-Nom y AnCora-Verb (léxicos semilla) a aquellas palabras del castellano semánticamente relacionadas, dando lugar a los léxicos DIANA-Nom y DIANA-Verb. Los corpus de cada tarea tendrán su léxico específico.
f) A partir del recurso AnCora-Net, que vincula los léxicos nominal y verbal del castellano a los correspondientes del inglés y catalán, se extenderá al conocimiento de DIANA-Nom y DIANA-Verb al catalán y al inglés.
g) Se desarrollará un parser semántico, DIANA-Parser, que tomará como datos de entrada los léxicos DIANA y el corpus analizado automáticamente con dependencias. El resultado será el análisis parcial de corpus con argumentos y papeles temáticos. El parser se desarrollará para las tres lenguas implicadas en el proyecto.
h) Análisis de los corpus con DIANA-Parser.
i) Obtención de léxicos de contextos sintáctico-semánticos asociados a los ítems léxicos nominales y verbales a partir de los árboles de dependencias semánticas del corpus. Estos léxicos de contextos constituirán la base para la obtención de construcciones.
j) Obtención de construcciones de base léxica y de base no léxica a partir de los léxicos de contextos sintáctico-semánticos. Se aplicarán técnicas de generalización y jerarquización de árboles. Se elaborará una jerarquía de construcciones de manera que se puedan establecer equivalencias entre las mismas de cara a su explotación en las diferentes aplicaciones.
k) Aplicación de la tecnología DIANA al desarrollo de aplicaciones que implican la comprensión de textos subjetivos.

4 Resultados esperable del proyecto

Los resultados del proyecto DIANA podrán incidir favorablemente en el desarrollo de aplicaciones en el marco web. Los avances en la tecnología web han puesto a nuestro alcance contenidos generados por los propios usuarios en forma de blogs, opiniones y todo tipo de interacciones en los medios de comunicación social, que se encuentran en forma no estructurada y expresados en fragmentos de texto donde se combina la simple narración de hechos con la toma de decisiones, experiencias aleccionadoras y opiniones sobre todo tipo de eventos. Se trata de una fuente de información valiosa que, si se dispone de los medios necesarios, se puede utilizar para la resolución de problemas sobre la base de conocimiento compartido y reutilizado. Las tecnologías que se desarrollarán permitirán avanzar en la captación de contenidos basados en la experiencia de los propios usuarios y en su aprovechamiento colectivo. Destacamos el desarrollo de diferentes técnicas y métodos susceptibles de ser incorporados en aplicaciones para:

- La identificación de estados de ánimo en textos subjetivos (entusiasmo, fatiga, relajación, estrés, frustración, depresión, agresividad), que permitirán captar el grado de satisfacción de los usuarios respecto de los servicios y aplicaciones que se ofrecen on-line. En los sistemas de tutorización un aspecto clave es conocer los diferentes estados de ánimo del alumno (grado de satisfacción, desorientación, motivación, comprensión de la materia, o ciertos síntomas relevantes en el proceso de aprendizaje).

- La detección en los medios sociales de comunicación de usuarios con potenciales trastornos de la personalidad, a nivel de agresividad o de neurosis, que podrían alertar sobre posibles casos de acoso y pedofilia. En lugar de rastrear manualmente la red para encontrar potenciales acosadores, un sistema automático de alerta ayudaría a los expertos en la identificación de los acosadores potenciales.
- La detección de opiniones fraudulentas creadas por parte de personas específicamente contratadas para este fin. Nuestra contribución consistirá en el desarrollo de técnicas para la extracción de las construcciones más recurrentes en la expresión de opiniones fraudulentas sobre personas, organizaciones, productos y servicios.
- La incorporación de un detector de ironía que mejorara las prestaciones de los sistemas de análisis de opinión. En el lenguaje figurado, el sentido literal del texto no coincide con el sentido que se quiere comunicar, de ahí la importancia de poder identificar las construcciones prototípicas de la expresión de la ironía en los medios sociales de comunicación.

Los nuevos recursos y herramientas que se prevé desarrollar serán utilizables en diferentes entornos de PLN y, muy especialmente, en aplicaciones como las que acabamos de describir. En concreto, los grafos de similitud léxica basados en el contexto y el analizador semántico serán de utilidad para superar las actuales limitaciones en temas como la resolución de la correferencia, la detección de paráfrasis, la obtención de la estructura argumental y la caracterización de usos estereotipados del lenguaje en casos como los que acabamos de apuntar.

Bibliografía

Baroni M. y A. Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Barrón-Cedeño, A., M. Vila, M. A. Martí y P. Rosso. 2013. Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics.* [doi:10.1162/COLI_a_00153].

Bogdanova D., Rosso P., Solorio T. Exploring High-Level Features for Detecting Cyberpedophilia. *Computer Speech and Language* (aceptado).

Croft, W. y D. A. Cruse. 2004. *Cognitive linguistics*, Cambridge Textbooks in Linguistics, Cambridge University Press.

Franco-Salvador M., Gupta P., Rosso P. 2013. Cross-Language Plagiarism Detection Using Multilingual Semantic Network. *Proc. 35th, ECIR-2013*, Springer-Verlag, LNCS(7814).

Goldberg, A. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Gries, S. Th. 2003. Collostructions: Investigating the interaction of words and constructions, *International Journal of Corpus Linguistics*, 8:2.

Hawkins, J. 2004. *On Intelligence*, Henry Holt and Company, LLC, New York, USA.

Hernández D., Guzmán R., Montes-y-Gómez M., Rosso P. 2013. Using PU-Learning to Detect Deceptive Opinion Spam. *WASSA*-2013.

Jackendof, R. 1999. Possible stages in the evolution of the language capacity. *Trends in Cognitive Sciences*, 3(7): 272-279.

Padó, S. y M. Lapata. 2007. Dependency-Based Construction of Semantic Space Models, *Computational Linguistics*, 33(2).

Peris, A., M. Taulé, H. Rodríguez y M. Bertran. 2013. LIARc: Labeling Implicit ARguments in Spanish deverbal nominalizations, *CICLING-2013.* Springer.

Reyes A. y P. Rosso 2012. Making Objective Decisions from Subjective Data: Detecting Irony in Customers Reviews. *Journal on Decision Support Systems*, 53(4):754–760.

Sag I., T. Baldwin, F. Bond, A. Copestake y D. Flickinger. 2001. Multiword Expressions: A Pain in the Neck for NLP. *CICLING-2002*. Springer, LNCS (2276): 1-15.

Tomasello, M. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition.* Harvard University Press.

Turney, P. y P. Pantel. 2010. From Frequency to meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141-188.

Wintner, S. 2009. What science underlies natural language engineering? *Computational Linguistics*, 35 (4): 641-644.

Zuidema, W. 2007. Parsimonious Data-Oriented Parsing, Proc. *EMNLP-CoNLL.*

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 227-230 recibido 08-04-2013 revisado 16-06-2013 aceptado 21-06-2013

TIMPANO: Technology for complex Human-Machine conversational interaction with dynamic learning

TIMPANO: Tecnología para interacción conversacional hombre-máquina compleja con aprendizaje dinámico.

Emilio Sanchis
ELiRF-Universitat Politecnica Valencia
Camino de Vera s/n 46022 Valencia
esanchis@dsic.upv.es

Alfonso Ortega
Universidad de Zaragoza
C/ María de Luna 1. 50018 Zaragoza
ortega@unizar.es

M. Inés Torres
Universidad del País Vasco UPV/EHU
Campus de Leioa. 48940 Leioa. Bizcaia
manes@we.lc.ehu.es

Javier Ferreiros
Universidad Politécnica Madrid
Ciudad Universitaria s/n. Madrid
jfl@die.upm.es

Resumen: El proyecto TIMPANO tiene por objetivo profundizar en el desarrollo de sistemas de comunicación oral hombre-máquina atendiendo principalmente a la capacidad de dar respuesta a múltiples requerimientos de los usuarios, como pueden ser el acceso a información, la extracción de información, o el análisis de grandes repositorios de información en audio. En el proyecto se hace especial énfasis en la adaptación dinámica de los modelos a diversos contextos, tanto de tipo acústico, como semántico o de idioma.
Palabras clave: Tecnologías del Habla, Interacción oral.

Abstract: The goal of the TIMPANO project is to research about the development of speech-driven human-machine interaction systems, regarding mainly the ability of answering multiple requirements from the users, such as accessing and extracting information, and analyzing large repositories of audio information. This project is especially focused in the dynamic adaptation of the models to different acoustic and semantic contexts as well as to different languages.
Keywords: Speech Techonogies, Oral interaction.

1 Informations of the Project

This Project is founded by the "Ministerio de Economía y Competitividad" TIN2011-28169-C05 and there are four reseach groups involved in it: ELiRF (Universidad Politécnica de Valencia), ViVoLab (Universidad de Zaragoza), TR&ST (Universidad del Pais Vasco), GTH (Universidad Politécnica de Madrid).

2 Introduction

Significant advances have been achieved in speech technologies allowing researchers to face new challenges in different areas of human-machine interaction slightly exploited so far, improving and consolidating already developed technologies at the same time. The new generation of speech interaction systems requires a progressive increase of robustness against the variability of users, languages and acoustic conditions, as well as against the huge volume of information those systems provide access to. The international scientific community is devoting strong efforts to provide solutions allowing to face unconstrained vocabularies, different tasks, several languages and open domain information retrieval systems for which an exhaustive signal analysis helps to extract as much information as possible.

Speech technologies, like some other disciplines that have arisen based on the demands of the information society, should

ISSN 1135-5948

have a major role in the future of the management and use of the large amount of digital contents that are being generated. The increasing flow of information in different languages and through different media, the large audio repositories (conferences, parliaments, broadcasting, ..), interactive information services, etc. are some examples of the new challenges the scientific community is facing. But this requires broadening the concept of oral interaction; because it is not just a problem of obtaining or emitting a sequence of words, but also detecting voice activity, who spoke in a given moment, in which language, the semantic information contained, etc. to be able to seek certain segments, or cooperate with the machine to achieve the proposed goal. Human speech is not only characterized by segmental information and spectral correlates of each sound. Several cognitive, emotional and contextual factors affect speech variability which operates at higher levels than phonetics. This higher level variability affects several speech features, including sound intensity, intonation or stops timing, and bears consequences on the emotional state of speech. Human beings do make a successful intensive use of this prosodic information to favor our communication processes. Thus, incorporating prosodic processing capabilities to advanced spoken interaction systems is a key factor to improve interaction results.

As a consequence of facing this problems, where the volume of information as well as the variety of users, languages, or media is enormous, the relevance of machine learning methods to allow systems to dynamically adapt to the new features of the interaction, or to detect and model new events (new words, concepts, users,..) is really important. Since obtaining these kind of systems from a-priori information or with models learned from few samples is unfeasible, it is necessary for the system to have self-learning mechanisms based on the evaluation of its behavior or, at least, ways to generate information to be readily used by the designer for adapting the models dynamically, and with little effort (in the line to provide tools to help designers).

In this framework, the research project proposed here is oriented to the development of voice interaction systems able to address new and more complex tasks, both in the sense of working with larger amounts of data and increasing the complexity of the application domain. Obviously we bound the specific areas where we are going to focus our work. Moreover, we try to follow a realistic approach, encouraging the development of assistive tools that for certain applications can facilitate the fulfillment of specific tasks by using spoken interaction systems. Much of the efforts will be focused on developing and applying self-learning techniques at different levels of knowledge representation, maximizing the benefit of using the data generated through the interaction with the system. On the other hand, special attention will be paid to the development and improvement of methodologies to process large amounts of audio information.

As a summary, it is worth to highlight the following lines of work in this project:

- Development of conversational interactive systems along two different lines: increase the robustness of dialogue systems for restricted domains, and facilitate the access to large amounts of audio data, improving two-way human-machine communication process (Griol et al., 2008) (Ortega et al., 2010).
- Performance improvement of several technological components needed for human-machine interaction.
    - Processing, analysis and distillation of information from raw data: speaker and language identification, acoustic segmentation, event detection, speech recognition for open-vocabulary tasks, topic detection, content retrieval… (Guijarrubia y Torres, 2010) (Pardo, Ferreiros y Montero, 2011)
    - Propose new learning techniques to allow dynamic model adaptation at all levels of knowledge (acoustic, lexical, semantic and dialogue) (Buera et al., 2010).
    - Deal with other components of conversational interaction as prosody, emotion and multilingualism (Perez-Ramirez, Torres y Casacuberta, 2008) (Barra-Chicote et al., 2010).

- In certain tasks give higher priority to systems that assist the user versus the fully automatic ones. Tasks that would be impossible to deal with due to the large amount of information to be processed can become tractable thanks to the use of systems that do part of the work.
- Regarding the development of applications we propose two different goals.
    - On one hand develop tools where conversational interactive systems assist the access to both structured and unstructured information,
    - On the other hand to take advantage of speech technologies on the design of socially useful applications such as systems to help people with special needs (i.e., speech/language disorders, physical disabilities like visual or hearing impairments …) Examples of this kind of systems are tools for speech therapy, aid devices for deaf people or sign language translation, areas in which some groups of the consortium have proven experience (Saz et al., 2010).

3 Objetives of the Project

Strategic Objectives:

This project will focus on covering gaps identified in existing systems of spoken human-machine interaction in areas such as:

- **Self-assessment** systems that provides information about the quality of their results.
- **Self-learning**, which allows systems to improve their use and adapt easily to new situations.
- **Scalability** and extensive interaction domains, connected with the need to address open-vocabulary tasks and large and diverse information repositories.

- Its usefulness as assistive tools to allow the user the fulfillment of tasks whenever they become complex or in situations when the user has serious difficulties (certain disabilities or speech and language disorders).

Scientific-Technological objectives:

These objectives are divided into three fundamental aspects: the treatment of acoustic information, treatment of the interaction and the transition from the acoustic domain to the interaction domain and vice versa.

- In terms of the acoustic aspects, the researchers aim to develop modules such as the identification / classification of acoustic events, segmentation / diarization / clustering or automatic speaker or language recognition all of them being able to offer, along with its outputs, an indication of how reliable these are, ie the capacity of self-assessment, using the information obtained from all the available sources. Also, if the self-assessment is not positive, we propose to exploit information obtained through user interaction to allow the modules to evolve as independently as possible.

- With regard to the processing of the interaction, to advance the semantic modeling by using unsupervised learning with automatic detection of new semantic events. Progress in the dynamic learning interaction systems in restricted domains, autonomously adapting them to specific tasks and modeling the temporal evolution of personalized interaction. In open domains, to assist the user in accessing unstructured information. Wherever necessary the treatment of multilingualism will be addressed, and if necessary, translation tools to assist the interaction will be used. Progressing this kind of systems to incorporate the emotional component in the interaction with the detection of user's emotional state, combination with an internal emotional model and appropriate response generation (emotional speech synthesis).

- In the transition between both domains, we will mainly develop two aspects, acoustic modeling and language modeling. According to the first one, we propose improvements in statistical modeling of acoustic emissions to be included in automatic speech recognition systems (ASR), modeling of disorders and variations of speech, the use of language independent acoustic units or robust feature extraction and specific classifiers: tandem, or hybrid. The on-line learning to adapt to speakers (including to those unregistered), speech pathologies, language / dialect / accent and also advances in acoustic modeling, also

for personalized speech synthesis. Regarding language modeling, progress in large vocabulary ASR, in the detection of terms and words out of vocabulary and progress in the adaptability to unsupervised semantic domain of finite-state transducers.

Implementation Objectives

In this field, this project proposes the development of tools that enhance the technological capabilities of the consortium, progressing the software architecture which is already available from previous common projects and the building of prototypes and demonstrators which show the achievements in this project:

- Adaptable Interaction Systems in Restricted Domains.
- Assistive Interaction Systems Including People with Special Needs (e-inclusion).
- Support to the Processing of Multimedia and Unstructured Information.

4 Aknowledgements

This work is founded by the "Ministerio de Economía y Competitividad" TIN2011-28169-C05

References

Barra-Chicote Roberto, Yamagishi Junichi, King Simon, Montero Juan Manuel, Macías-Guarasa Javier . 2010 "Analysis of Statistical Parametric and Unit-Selection Speech Synthesis Systems Applied to Emotional Speech" *Speech Communication*, Volume 52 Issue 5, Pages 394-404.

Buera, Luis, A. Miguel, O. Saz, A. Ortega y E. Lleida Solano. 2010 "Unsupervised Data-Driven Feature Vector Normalization With Acoustic Model Adaptation for Robust Speech Recognition*" IEEE Transactions on Audio, Speech and Language Processing" . Vol.* 18, no. 2, pp. 296-309.

Griol David, Hurtado Lluís F.,Segarra Encarna, y Sanchis Emilio. 2008. Acquisition and Evaluation of a Dialog Corpus throughWOz and Dialog Simulation Techniques. *Speech Communication*, 50, pp. 666-682.

Guijarrubia, Víctor G. y Torres, M. Inés. 2010, "Text- and speech-based phonotactic models for spoken language identification of Basque and Spanish", *Pattern Recognition Letters*, 31, 6: 523 - 532,

Ortega Lucia, Galiano Isabel, Hurtado Lluís-F., Sanchis Emilio y Segarra Encarna. 2010. A Statistical Segment-Based Approach for Spoken Language Understanding. Proc. of INTERSPEECH'10, pp. 1836-1839.

Pardo, J.M., Ferreiros, J. y Montero, J.M. 2011 "Speaker diarization based on intensity channel contribution", *IEEE Transactions on Audio, Speech and Language Processing* 19 , Vol.19, Issue: 4 Pages: 754 – 761.

Pérez-Ramírez, A., Torres, M. Inés y Casacuberta, F. 2008, "Joining linguistic and statistical methods for Spanish-to-Basque speech translation", *Speech Communication*, 50: 1021-1033.

Saz O., Yin S.-C., Lleida E., Rose R., Vaquero C. y Rodriguez W. R.. 2009. "Tools and technologies for Computer-Aided Speech and Language Therapy". *Speech Communication*, Special Issue on Spoken Language Technology for Education.

Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 231-234 recibido 27-03-2013 revisado 16-06-2013 aceptado 21-06-2013

Tratamiento de textos para mejorar la comprensión lectora en alumnos con deficiencias auditivas

Handling text in order to improve reading comprehension for hearing-impared students

Estela Saquete, Sonia Vázquez, Elena Lloret, Fernando Llopis, Jose Manuel Gómez, Alejandro Mosquera
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
{stela,svazquez,elloret,llopis,jmgomez,amosquera}@dlsi.ua.es

Resumen: Proyecto emergente centrado en el tratamiento de textos educativos en castellano con la finalidad de reducir las barreras lingüísticas que dificultan la comprensión lectora a personas con deficiencias auditivas, o incluso a personas aprendiendo una lengua distinta a su lengua materna. Se describe la metodología aplicada para resolver los distintos problemas relacionados con el objetivo a conseguir, la hipótesis de trabajo y las tareas y los objetivos parciales alcanzados.
Palabras clave: tecnologías del lenguaje humano, comprensión lectora, simplificación textual

Abstract: This project is focused on textual treatment in Spanish in order to reduce language barriers that hinder hearing impaired people from reading comprehension, or even people learning a new language. This paper describes the methodology used to face the different problems related to the proposed objective, as well as the working hypothesis and partial tasks and objectives achieved.
Keywords: human language technologies, Reading comprehension, textual simplification

1 Datos del proyecto

Este proyecto está dirigido por Estela Saquete, miembro del Grupo de Procesamiento del Lenguaje y Sistemas de Información (GPLSI) de la Universidad de Alicante. Está financiado por la Universidad de Alicante (GRE11-21) dentro del programa de ayudas a proyectos emergentes.

Contacto
Email: stela@dlsi.ua.es
Teléfono: 965903400 ext. 2894
Dpto. de Lenguajes y Sistemas Informáticos, Universidad de Alicante,
Carretera San Vicente del Raspeig, s/n,
03690, Alicante, España

2 Introducción

La comprensión lectora se define en el informe de PISA 2000[1] como: *"la capacidad de comprender, emplear información y reflexionar a partir de textos escritos, con el fin de lograr las metas individuales y desarrollar el conocimiento y el potencial personal y así participar eficazmente en sociedad"*.

Es por ello que, actualmente, dicha comprensión lectora se ha convertido en uno de los principales temas de estudio en el ámbito de la psicología y la educación (Olivé, 2009). Dichas investigaciones han determinado que las habilidades y condiciones que se requieren para la comprensión lectora son complejas y múltiples, siendo fundamental dentro de este

[1] http://www.mec.es/multimedia/00005713.pdf

grupo las habilidades lingüísticas, y que son en las que nos centraremos para este proyecto.

La escasa capacidad de comprensión lectora es un problema que va en aumento en nuestra sociedad, como ya se constató en el último informe PISA (Programa para la Evaluación Internacional de Alumnos) 2006. Además, el problema de la comprensión lectora todavía es mayor para las personas con deficiencias auditivas. Esta problemática que relaciona las dificultades en la comprensión lectora con la sordera lleva tiempo siendo estudiada, tanto a nivel léxico como sintáctico (King & S.P., 1985) (LaSasso & Davey, 1987) (Paul & Gustafson, 1991) (Berent, 1996).

Estudios previos han detectados los siguientes obstáculos que las personas con deficiencias auditivas se encuentran a la hora de comprender un texto (Mies, 1992) (Herrera, 2003) (Stockseth, 2002):

a) *problemas de ambigüedad*. Existen numerosas palabras en castellano que pueden poseer más de un significado, y toman su sentido correcto en función del contexto en el que aparecen. Todo este tipo de palabras polisémicas pueden generar muchos problemas para la comprensión de un texto;

b) *vocabulario limitado.* Estos lectores se fijan más en palabras familiares y usan más sustantivos concretos y verbos familiares que palabras abstractas. Muchas veces también tienen problemas con el reconocimiento de entidades de nombre y su contextualización;

c) *dificultades en la interpretación de estructuras sintácticas complejas* cuya sintaxis se desvía de la estructura básica de constituyentes explícitos sustantivo-verbo-sustantivo y del orden Sujeto-Verbo-Objeto. Es decir, estructuras más complejas, como por ejemplo, oraciones activas simples transitivas, pasivas o coordinadas o subordinadas suponen un problema de compresión extra para personas con deficiencias auditivas;

d) *problemas en situar los eventos en la línea temporal* si en el texto existen saltos en el tiempo, hacia delante o hacia atrás, que implican una interpretación de las señales y expresiones temporales para su total comprensión.

Como suele ocurrir, muchos de estos estudios se centran en la problemática lingüística del inglés, pero afortunadamente también existen numerosos trabajos que estudian la comprensión lectora en castellano, tanto a nivel léxico (Mies, 1992) como a nivel sintáctico (Stockseth, 2002). Además, hay múltiples estudios en castellano especialmente aplicados al ámbito educativo (Alegría & Leybaert, 1985) (Asensio & Carretero, 1989) (Ferrández Mora, 1989).

Dentro del área de investigación que nos ocupa, las Tecnologías del Lenguaje Humano, existe una amplia variedad de trabajos relacionados con la obtención de manera automática o semi-automática del lenguaje de signos a partir de textos hablados o escritos (Parton, 2005) (Wu, Chiu, & Guo, 2004) (Duchnowsku, Lum, Krause, Sexton, Bratakos, & Braida, 2000). Además, debemos destacar el proyecto MÁS, financiado por la Unión Europea, y cuya finalidad es comprobar los efectos del uso de una herramienta multimedia para mejorar la comprensión lectora utilizando como apoyo el lenguaje de signos (Ferrer Manchón, 2001).

3 Objetivos del proyecto

El objetivo principal de este proyecto es diseñar, implementar y evaluar una herramienta tecnológica que sea capaz de convertir documentos en castellano en textos de lectura fácil para personas con deficiencias auditivas, y por tanto, pueda servir de ayuda para la comprensión lectora de dichas personas. El proceso de conversión implica detectar automáticamente características lingüísticas del documento de entrada que puedan impedir su fácil comprensión y automáticamente reducir y/o eliminar esos obstáculos pero preservando el significado general del documento en la medida de lo posible. Las tecnologías del lenguaje humano serán aplicadas para eliminar los obstáculos en la comprensión derivados de las estructuras complejas y de la ambigüedad en los significados en textos escritos, y generando un apoyo extra mediante imágenes o pictogramas, líneas temporales, definiciones de enciclopedias online simples y resúmenes del texto original.

Por tanto, los objetivos concretos del proyecto serán:

1. Analizar en profundidad las diferentes aproximaciones existentes en las Tecnologías del Lenguaje Humano para tratar cada uno de los obstáculos que dificultan la comprensión de los textos.

2. Desarrollar las herramientas necesarias para poder solucionar los problemas analizados y evaluar cada una de estas herramientas de manera independiente. Además, también serán evaluadas en las diferentes competiciones relacionadas con el fin de demostrar su funcionalidad y eficiencia.
3. Desarrollar una interfaz que, dado un texto de entrada, aplique las diferentes herramientas desarrolladas para obtener un texto de lectura fácil con los apoyos que hemos comentado y reduciendo y/o eliminando las posibles barreras del texto.
4. Analizar y evaluar la herramienta con los usuarios finales, con el fin de determinar su eficacia, su usabilidad y su impacto en la compresión lectora de estos usuarios.

Tal y como hemos presentado en el apartado anterior, las herramientas tecnológicas existentes hasta el momento para facilitar la comprensión de textos a las personas con sordera se basan principalmente en el uso del lenguaje de signos (Parton, 2005) (Wu, Chiu, & Guo, 2004) (Duchnowsku, Lum, Krause, Sexton, Bratakos, & Braida, 2000). Sin embargo, nuestra propuesta, pretende facilitar al lector el significado del texto, transformándolo a un texto de lectura sencilla, facilitando por un lado la comprensión léxica, y por otro la comprensión sintáctico-semántica. Para la comprensión léxica se detectarán aquellas palabras más complejas y se aportará, en dichos casos, de manera automática, un conjunto de posibles palabras sinónimas más comunes[2] y un pictograma relacionado con la palabra[3]. Además, para las entidades nombradas se facilitarán tanto imágenes relacionadas como una definición obtenida de una enciclopedia online simple[4]. Por otro lado, a nivel sintáctico-semántico, la herramienta detectaría automáticamente estructuras sintácticas complejas e intentaría traducirla a una o varias estructuras más sencillas, así como la detección y resolución de expresiones temporales y la generación de resúmenes que simplifiquen textos complejos.

Esta herramienta supone un interés muy importante para las personas con deficiencias auditivas, puesto que podría ayudarles a mejorar en su comprensión lectora y por tanto permitir a estas personas ampliar sus horizontes informativos y culturales. Además, puede ser una ayuda muy indicada en el caso de la educación de los niños con sordera, en quienes los mismos profesores suelen realizar manualmente esta tarea de conversión para facilitar una lectura fluida con menor esfuerzo y más provecho para estos niños.

4 Tareas a desarrollar

Para la consecución del proyecto será necesario completar el conjunto de tareas que se mencionan a continuación:

Análisis del problema

En esta tarea se analizarán las distintas aproximaciones existentes a la desambiguación de significados, al tratamiento de estructuras complejas, a la recuperación de definiciones e imágenes asociadas a un significado concreto, palabras, o frases y así como la asociación de significados a pictogramas. También analizaremos el problema de la resolución de información temporal y la generación de resúmenes sencillos de fácil comprensión. Sobre esta base teórica se investigarán nuevas técnicas para la mejora del sistema, basándonos en la adquisición de conocimiento general del mundo.

Desarrollo y evaluación

En esta tarea se llevará a cabo la implementación de las técnicas estudiadas en la tarea anterior, en módulos independientes, dando como resultado un sistema capaz de detectar entidades complejas en un texto, desambiguarlas, proporcionar pictogramas asociados y la línea temporal asociada, además de generar un pequeño y simple resumen del texto de forma automática. En este punto se realizará también la evaluación de nuestra investigación en las diferentes competiciones internacionales relacionadas.

Visualización de la información

En esta tarea se busca crear un interfaz sencillo y muy amigable que permita al usuario acceder a toda esta información extra proporcionada automáticamente a partir de un texto de entrada en castellano y que pretende hacer más sencillo su entendimiento. En concreto, esta interfaz será puesta a disposición del Centro de Apoyo al Estudiante (CAE) de la

[2] http://es.wiktionary.org
[3] Google imágenes
[4] http://simple.wikipedia.org

Universidad de Alicante para que pruebe la herramienta con este colectivo de la Universidad, realizando de esta forma una evaluación de la misma, que medirá su eficiencia, usabilidad y eficacia en el soporte a la comprensión de textos, recibiendo así mismo una retroalimentación para su mejora.

5 Situación actual del proyecto

Dentro de las tareas antes mencionadas, actualmente se ha realizado parte del desarrollo del sistema propuesto.

Hasta el momento, el sistema carga el texto en castellano y es capaz de reconocer tanto las entidades nombradas como las expresiones temporales, ambos elementos del texto son marcados con un enlace. Si el usuario final tiene problemas en el entendimiento de la entidad nombrada o de la expresión temporal elegiría el término y obtendría lo siguiente:

- *Para las entidades nombradas*: una definición del término en Simple Wikipedia y tres imágenes de Google imágenes.
- Para las expresiones temporales: la expresión temporal resuelta en una fecha o periodo de fechas concretas.

El sistema será completado con los avances que se vayan realizando para completar todas las tareas propuestas para el proyecto.

Bibliografía

Alegría, J., & Leybaert, J. (1985). Adquisición de la lectura en el niño sordo: un enfoque psicolingüístico. Investigación y Logopedia.

Asensio, M., & Carretero, M. (1989). La lectura en los niños sordos. Cuadernos de pedagogía 174.

Berent, G. (1996). The acquisition of English Syntax by Deaf Learners. Handbook of Second Language Acquisition, 469-506.

Duchnowsku, P., Lum, D., Krause, J., Sexton, M., Bratakos, M., & Braida, L. (2000). Development of speechreading supplements based on automatic speech recognition. IEEE Trans Biomed Eng. 47(4), 487-496.

Ferrández Mora, J. A. (1989). La lectura en el corriculum escolar del niño sordo. Rev. Logop Fon Audiol, IX.

Ferrer Manchón, A. (2001). La comprensión lectora en personas sordas adultas y el acceso a la Universidad. ISAAC 2001: Odisea de la Comunicación. Segundas Jornadas sobre comunicación Aumentativa y Alternativa.

Herrera, V. (2003). Desarrollo de habilidades lectoras en sujetos sordos signates, a partir del uso de código dactílicos. Tesis doctoral no publicada.

King, C., & S.P., Q. (1985). Reading and Deafness. Collegue Hill Press.

LaSasso, C., & Davey, B. (1987). The relationship between lexical knowledge and reading comprehension for prelingually, profoundly hearing impaired students. The Volta Review 89, 211-220.

Mies, B. (1992). El léxico en la comprensión lectora: Estudio de un grupo de alumnos sordos adolescentes. Rev. Logop., Fon., Audiol., Vol. XII.

Olivé, M. J. (2009). La lectura: una necesidad para la inclusión social y la democracia. Separata del Manifiesto PIAPAS. Confederación española de familias de personas sordas.

Parton, B. (2005). Sign language recognition and lanslation: a multidisciplined approach from the field of the artificial intelligence. J. Deaf Stud Deaf Educ.

Paul, P., & Gustafson, G. (1991). Hearing-impaired students' comprenhension of hig-frequency multi-meaning words. Remedial and special education, 12, 52-62.

Stockseth, D. R. (2002). Comprensión de la sintaxis española por lectores sordos chilenos. Revista Signos, v. 35.

Wu, C., Chiu, Y., & Guo, C. (2004). Text generation from Taiwanese Sign Language using PST-based language model for augmentative communication. IEEE Trans Neural Syst. Rehabil Eng., 12(4).

Información General

Información para los Autores

Formato de los Trabajos

- La longitud máxima admitida para las contribuciones será de 8 páginas DIN A4 (210 x 297 mm.), incluidas referencias y figuras.
- Los artículos pueden estar escritos en inglés o español. El título, resumen y palabras clave deben escribirse en ambas lenguas.
- El formato será en Word ó LaTeX.

Envío de los Trabajos

- El envío de los trabajos se realizará electrónicamente a través de la página web de la Sociedad Española para el Procesamiento del Lenguaje Natural (http://www.sepln.org)
- Para los trabajos con formato LaTeX se mandará el archivo PDF junto a todos los fuentes necesarios para compilación LaTex.
- Para los trabajos con formato Word se mandará el archivo PDF junto al DOC o RTF
- Para más información http://www.sepln.org/?page_id=358

Hoja de Inscripción para Instituciones

Datos Entidad/Empresa

Nombre : .....
NIF : ..... Teléfono : .....
E-mail : ..... Fax : .....
Domicilio : .....
Municipio : ..... Código Postal : ..... Provincia : .....
Áreas de investigación o interés: .....
.....

Datos de envío

Dirección : ..... Código Postal : .....
Municipio : ..... Provincia : .....
Teléfono : ..... Fax : ..... E-mail : .....

Datos Bancarios:

Nombre de la Entidad : .....
Domicilio : .....
Cód. Postal y Municipio : .....
Provincia : .....

Cód. Banco (4 dig.)	Cód. Suc. (4 dig.)	Dig. Control (2 Dig.)	Núm.cuenta (10 dig.)
.....	.....	.....	.....

Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN).

Sr. Director de:

Entidad : .....
Núm. Sucursal : .....
Domicilio : .....
Municipio : ..... Cód. Postal : .....
Provincia : .....
Tipo cuenta
(corriente/caja de ahorro) : .....
Núm Cuenta : .....

Ruego a Vds. que a partir de la fecha y hasta nueva orden se sirvan de abonar a la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) los recibos anuales correspondientes a las cuotas vigentes de dicha asociación.

Les saluda atentamente

Fdo: .....
(nombre y apellidos del firmante)

.....de .....de.....

.....

Cuotas de los socios institucionales*: 300 €.*

Nota: La parte inferior debe enviarse al banco o caja de ahorros del socio

Hoja de Inscripción para Socios

Datos Personales

Apellidos : .....
Nombre : .....
DNI : ..... Fecha de Nacimiento : .....
Teléfono : ..... E-mail : .....
Domicilio : .....
Municipio : ..... Código Postal : .....
Provincia : .....

Datos Profesionales

Centro de trabajo : .....
Domicilio : .....
Código Postal : ..... Municipio : .....
Provincia : .....
Teléfono : ..... Fax : ..... E-mail : .....
Áreas de investigación o interés: .....
.....

Preferencia para envío de correo:

[ ] Dirección personal [ ] Dirección Profesional

Datos Bancarios:

Nombre de la Entidad : .....
Domicilio : .....
Cód. Postal y Municipio : .....
Provincia : .....

Cód. Banco (4 dig.)	Cód. Suc. (4 dig.)	Dig. Control (2 Dig.)	Núm.cuenta (10 dig.)
.....	.....	.....	.....

En.....a.....de.....de.....
(firma)

Sociedad Española para el Procesamiento del Lenguaje Natural. SEPLN

Sr. Director de:
Entidad : .....
Núm. Sucursal : .....
Domicilio : .....
Municipio : ..... Cód. Postal : .....
Provincia : .....
Tipo cuenta
(corriente/caja de ahorro) : .....

Ruego a Vds. que a partir de la fecha y hasta nueva orden se sirvan de abonar a la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) los recibos anuales correspondientes a las cuotas vigentes de dicha asociación.

Les saluda atentamente

Fdo: .....
(nombre y apellidos del firmante)

.....de .....de.....

Cuotas de los socios*: 18 €* (residentes en España) o 2*4 €* (socios residentes en el extranjero).
Nota: La parte inferior debe enviarse al banco o caja de ahorros del socio

Información Adicional

Funciones del Consejo de Redacción

Las funciones del Consejo de Redacción o Editorial de la revista SEPLN son las siguientes:

- Controlar la selección y tomar las decisiones en la publicación de los contenidos que han de conformar cada número de la revista
- Política editorial
- Preparación de cada número
- Relación con los evaluadores y autores
- Relación con el comité científico

El consejo de redacción está formado por los siguientes miembros

L. Alfonso Ureña López (Director)
Universidad de Jaén
laurena@ujaen.es
Patricio Martínez Barco (Secretario)
Universidad de Alicante
patricio@dlsi.ua.es
Manuel Palomar Sanz
Universidad de Alicante
mpalomar@dlsi.ua.es
Felisa Verdejo
UNED
felisa@lsi.uned.es

Funciones del Consejo Asesor

Las funciones del Consejo Asesor o Científico de la revista SEPLN son las siguientes:

- Marcar, orientar y redireccionar la política científica de la revista y las líneas de investigación a potenciar
- Representación
- Impulso a la difusión internacional
- Capacidad de atracción de autores
- Evaluación
- Composición
- Prestigio
- Alta especialización
- Internacionalidad

El Consejo Asesor está formado por los siguientes miembros:

José Gabriel Amores	Universidad de Sevilla
Toni Badía	Universitat Pompeu Fabra
Manuel de Buenaga	Universidad Europea de Madrid
Irene Castellón	Universitad de Barcelona
Arantza Díaz de Ilarraza	Universidad del País Vasco
Antonio Ferrández	Universidad de Alicante
Mikel Forcada	Universidad de Alicante
Ana García-Serrano	UNED
Koldo Gojenola	Universidad del País Vasco
Xavier Gómez Guinovart	Universidad de Vigo
Julio Gonzalo	UNED

Cartas al director

Sociedad Española para el Procesamiento del Lenguaje Natural
Departamento de Informática. Universidad de Jaén
Campus Las Lagunillas, EdificioA3. Despacho 127. 23071 Jaén
secretaria.sepln@ujaen.es

Más información

Para más información sobre la Sociedad Española del Procesamiento del Lenguaje Natural puede consultar la página web http://www.sepln.org.
Los números anteriores de la revista se encuentran disponibles en la revista electrónica: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/issue/archive
Las funciones del Consejo de Redacción están disponibles en Internet a través de http://www.sepln.org/?cat=21
Las funciones del Consejo Asesor están disponibles Internet a través de la página http://www.sepln.org/?page_id=1061