Procesamiento del Lenguaje Natural, Revista nº **49**, septiembre de 2012

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 **49**, septiembre de 2012

ISSN: 1135-5948

Comité Editorial

Consejo de redacción

ISSN: 1135-5948
ISSN electrónico: 1989-7553
Depósito Legal: B:3941-91
Editado en: Universitat Jaume I de Castellón
Año de edición: 2012

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

Mariona Taulé	Universidad de Barcelona
L. Alfonso Ureña	Universidad de Jaén
Felisa Verdejo	UNED
Manuel Vilares	Universidad de A Coruña
Ruslan Mitkov	Universidad de Wolverhampton, UK
Sylviane Cardey-Greenfield	Centre de recherche en linguistique et traitement automatique des langues, France
Leonel Ruiz Miyares	Centro de Linguística 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
Nuno J. Mamede	Instituto de Engenharia de Sistemas e Computadores, Portugal
Bernardo Magnini	Fondazione Bruno Kessler, Italia

Revisores adicionales

María José Aramburu	Universitat Jaume I
Rafael Berlanga Llavori	Universitat Jaume I
Victor J. Díaz Madrigal	Universidad de Sevilla
Florentino Fernández Riverola	Universidade de Vigo
José María Gómez Hidalgo	Optenet
Antonio Jimeno Yepes	National Library of Medicine (Washington)
Paloma Martínez Fernández	Universidad Carlos III de Madrid
Manuel J. Maña López	Universidad de Huelva
Jacinto Mata Vázquez	Universidad de Huelva
Ramón López-Cózar Delgado	Universidad de Granada
Paolo Rosso	Universitat Politècnica de València
Ismael Sanz	Universitat Jaume I
Isabel Segura Bedmar	Universidad Carlos III de Madrid
José Antonio Troyano	Universidad de Sevilla

Procesamiento del Lenguaje Natural, Revista nº **49**, septiembre de 2012

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 49 de la revista de la Sociedad Española para el Procesamiento del Lenguaje Natural contiene trabajos correspondientes a tres apartados diferenciados: comunicaciones científicas, proyectos y demostraciones. 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 38 eran artículos científicos regulares, 4 correspondían a propuestas de proyectos y 4 a demostraciones. De entre los 38 artículos regulares 19 han sido finalmente seleccionados para su publicación, lo cual fija una tasa de aceptación del 50%.

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 hayan sido igual o superior a 5 sobre 7.

Septiembre de 2012
Los editores

Procesamiento del Lenguaje Natural, Revista nº **49**, septiembre de 2012

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

Análisis automático del Contenido Textual

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 13-20 recibido 04-05-12 revisado 01-06-12 aceptado 06-06-12

Reducing Text Complexity through Automatic Lexical Simplification: an Empirical Study for Spanish

Reducción de la complejidad de un texto a través de la simplificación léxica: un estudio para el español

Biljana Drndarević y **Horacio Saggion**
Universitat Pompeu Fabra
Departament of Communications and Information Technology
C/Tànger 122, Barcelona
{biljana.drndarevic, horacio.saggion}@upf.edu

Resumen: En este artículo presentamos los resultados de un estudio cuyo objetivo es sentar las bases para el desarrollo de un módulo de simplificación léxica para el español. Basándonos en estudios para otras lenguas analizamos, en primer lugar, la distribución de la frecuencia y la longitud de palabra en textos originales y sus simplificaciones manuales. En segundo lugar nos centramos en los casos de clarificación de información a través de la introducción de definiciones en textos simplificados. Finalmente estudiamos la reducción del contenido informativo del texto y proponemos un sistema para su tratamiento basado en técnicas de resumen. Nuestro estudio empírico sienta las bases para el desarrollo de un componente de tratamiento léxico en un sistema de simplificación de textos en desarrollo.
Palabras clave: simplificación léxica, frecuencia, longitud de palabra, reducción del contenido

Abstract: In this paper we present the results of a study directed towards developing a lexical simplification module of an automatic simplification system for Spanish, intended for readers with cognitive disabilities. We here observe the word length and frequency distribution of two sets of texts that make up our parallel corpus, and we focus on cases of information expansion (through the insertion of definitions)and content reduction (through summarisation). Our ultimate goal is computational implementation of lexical changes in the future.
Keywords: lexical simplification, word frequency, word lengh, information expansion, content reduction

1 Introduction

The digitalisation of information as an essential characteristic of our society has created an illusion of an ideal world where information is freely shared and equally accessible to everyone. Yet, the reality is disappointingly different, as shown by the results of a UN audit conducted with the aim of testing the state of accessibility of 100 leading websites around the world. Only three web pages achieved basic accessibility status. As a result, we have witnessed an increased interest in the issues of e-Accessibility, i.e. the ability for individuals with specific needs to access digital content. For that reason, in recent years NLP has seen a growing number of automatic text simplification systems developed for a wide range of end users. The need and interest for such systems arise from the fact that text is often so complex that it results incomprehensible.

Our project follows this line of research, centering on the development of a tool for automated simplification of newspaper articles in Spanish, meant as an aide for readers with cognitive disabilities. We are currently working on a lexical simplification module, more specifically detecting types of lexical change in a parallel corpus of original and manually simplified news articles, with the aim of preparing their computational implementation. The importance of lexical transformations in text simplification has already been underlined in previous work (Caseli et al., 2009; Specia, 2010). Our corpus analysis has also shown that lexical changes are the most common type of operations carried out by human editors. In broad terms, words and

ISSN 1135-5948

expressions perceived as complicated are substituted with simpler synonyms or rewritten using paraphrase. As shown in previous work (Carroll et al., 1998; Bautista, Gervás, and Madrid, 2009) and the simplification guidelines followed to obtain simplified texts for our corpus[1], "complicated" words tend to be longer and less frequently used ones. Hence, for example, *médico* (*doctor*) is preferred instead of its longer and less frequently used hyponym *psichiatra* (*psychiatrist*). We will, therefore, observe the distribution of word frequency and word length in the original and simplified texts in our corpus with the aim of testing how relevant the combination of these factors might be when conducting synonym substitution. In addition to that, we concentrate on cases of information expansion and content reduction. The former occurs through the insertion of definitions of difficult terms, where, for example, *Amnesty International* is defined as *an organisation that defends human rights worldwide*. On the other hand, content reduction is most often seen with numerical expressions and named entities.

The remainder of this paper is organised as follows: Section 2 addresses the related work in the field of automatic lexical simplification; in Section 3, we describe our methodology, followed by Section 4 where we discuss the results of our study. We conclude and outline our future work in Section 5.

2 Related Work

Previous work in the field of automatic text simplification already established the importance of lexical change. Carroll et al. (1998) presented a project for simplification of news articles in English, in which they used a combination of synonym look-up and word frequency count to carry out lexical substitution. For every content word in the input text, a set of synonyms is extracted from WordNet and Kucera-Francis frequencies are searched in the Oxford Psycholinguistic Database (Quinlan, 1992), upon which the most frequent synonym from the set is chosen for the simplified version of the text. Similar approach has been used in a number of other works. Lal and Ruger (2002) borrowed this method to deal with the lexical component of their automatic text summarizer. Burstein et al. (2007) focused on vocabulary changes when offering ATA V.1.0 as a text adaptation tool for L2 teachers and language learners. Bautista, Gervás, and Madrid (2009) also refer to a thesaurus when extracting candidates for lexical substitution, but their choice is guided by word-length rather than frequency. Caseli et al. (2009) build a parallel corpus for Brazilian Portuguese and extract lexical simplification operations applied by a human annotator, using a list of simple words and a list of discourse markers as resources for synonym substitution.

Acknowledging the fact that many words are polysemic and that, therefore, simple synonym substitution does not always produce a felicitous output, De Belder, Deschacht, and Moens (2010) suggest the use of word sense disambiguation techniques in order to account for contextual information. For every given word they create two sets of "alternative words" – one based on synonyms from WordNet or a similar dictionary, and another one generated by means of the Latent Words Language Model. Once the intersection of these two sets is found, the probability which determines whether it is a suitable replacement is calculated for every word of the intersection. To measure the probability they take into account the difficulty of the word, based on Kucera-Francis frequency, the average number of syllables and unigram probability extracted from a corpus of easy-to-read texts, such as Simple English Wikipedia.

3 Methodology

In order to conduct data analysis we have gathered a corpus of 200 news articles in Spanish. Subsequently, 40 articles have been manually simplified by trained human editors following easy-to-read guidelines proposed by Anula (2009). The most relevant for our current work are preserving the essential information and eliminating any superfluous content; using higher frequency words and avoiding technical terms; and avoiding long words and substituting them with their shorter synonyms with the same frequency index. However, we are interested to see how human editors deal with cases not envisioned by the guidelines, as well as those not described in sufficient detail, such as, for example, what terms are to be explained by means of a definition.

[1]The guidelines are currently in the form of internal project documentation and are to be published at a later date.

Both original and simplified texts have been automatically annotated with FreeLing (Padró et al., 2010). Additionally, sentence alignments have been produced automatically and any errors have been manually corrected through an alignment plug-in in GATE (Cunningham et al., 2002), a graphical editing tool for text processing. We have observed cases where one original (O) sentence corresponds to one or more simplified (S) sentences, cases where the relation is reversed as well as cases where there is no correlation between O and S sentences, due to information elimination or expansion.

We analyse thus aligned pairs of O and S texts in order to detect simplification operations applied at the lexical level, upon which we ponder the possibility of applying these operations computationally. In addition to that, we conduct text processing at the word level in order to gather data relative to word frequency and word length. Previous work having mainly concentrated on word frequency when applying synonym substitution, our intention is to test on our parallel corpus how this factor combines with that of word length, a traditional readability metric. Frequencies are extracted from a dictionary based on the Referential Corpus of Contemporary Spanish[2]. Every word in the dictionary is assigned a relative frequency index (FI) from 1 to 6, where 1 represents the lowest frequency and 6 the highest. The words that do not appear in the dictionary are assigned FI 0. We placed these words in three different categories: named entities, numerical expressions and what we call rare words. Among rare words we encounter multi-word expressions, such as complex function words, like *a través de* (*by means of*). This is due to the fact that multi-word expressions are recognized as such by FreeLing, whereas the current version of the frequency dictionary does not contain such words. However, the ratio of these words with respect to the total is fairly small so as to significantly influence overall results (1.08% in O and 0.59% in S).

4 Data Analysis

The corpus analysis has provided us with an insight into what lexical elements are treated and in what manner. We here concentrate on the insertion of definitions of difficult terms and concepts, and the treatment of named entities (NE) and numerical expressions (NumExp). In general, these can be seen as cases of information expansion on the one hand (insertion of definitions), and information elimination on the other, since a significant number of NE and NumExp are eliminated. In addition to that, we processed all original and simplified texts, placed into two separate sets (O and S), with the aim of obtaining a quantitative description of these sets, as reflected in the following:

- average sentence length;
- average number of sentences per text;
- average word length (in characters);
- the distribution of n-character words;
- the distribution of n-frequency words.

Sections that follow summarise the results of the quantitative analysis of O and S text sets, the treatment of definitions, named entities and numerical expressions.

4.1 Word length and frequency

Table 1 summarizes the data relative to text, sentence and word length, where s/t stands for "sentence per text" while w/s represents "word per sentence".

	Original	**Simple**
Total words	6595	3912
Total sentences	246	324
Average s/t	6.64	8.75
Average w/s	26.8	12.07
Average word length	5.44	5.07

Table 1: Average text, sentence and word length in original and simplified texts

As can be appreciated, S texts tend to be quite shorter on the whole, containing around 40% fewer words than O texts. However, they contain 24% more sentences than O texts, and their sentences are more than 50% shorter. The tendency is clear – long O sentences are generally split into shorter ones, and a considerable amount of O content is eliminated. We will explore the latter in more detail in Section 4.3.

One curious observation is that relative to average word length – contrary to our expectations, S words are only slightly shorter than O ones. We therefore focused on all

[2]http://corpus.rae.es/creanet.html

words with 1, 2, ... 20 characters, while longer words have been placed in categories of words with 21-30 characters, words with 31-40 characters and words with more than 40 characters.[3] The data analysis revealed that the most prolific words in both O and S texts are two-character words, accounting for as much as 27% of the texts. The vast majority of these are function words (97.61% in O and 88.97% in S). We have also observed that three to seven-character words are more abundant in S texts, whereas longer words are slightly more common in O texts. However, it is interesting to note that S texts contained on average slightly more eighteen-character words and words containing between 21 and 30 characters. These are all cases of named entities, often repeated through the insertion of definitions (discussed more in depth in Section 4.4).

On the whole, we can conclude that in S texts there is a tendency towards using shorter words of up to ten characters, with one to five-character words taking up 59.81% of the set and one to ten-character words accounting for 94.04% of the content. Longer words are almost exclusively reserved for named entities, often repeated when a definition of the terms in question is inserted.

Apart from word length, we explored how frequency acts as a factor in distinguishing between original and simplified, or difficult and easy words. We have detected words with frequency index 3, 4, 5 and 6, as well as words absent from the dictionary and were therefore assigned FI 0. The latter include numerical expressions (NumExp), named entities (NE), and what we here call *rare words*, i.e. all words not found in the dictionary that are neither NE nor NumExp. A small number of these are multi-word expressions, but the majority are indeed words not used very often, such as *intransigencia* (*intransigence*) or foreign words, like *e-book*. Table 2 contains data relative to average number of n-frequency words in O and S texts, where zero frequency words have been additionally separated into the categories of NumExp, NE and rare words and are printed in bold.

We can observe that the frequency distribution is fairly equal in both sets of texts, with the greatest divergence in the category of rare words, which are more than 50% less abundant in S texts than O texts. NE are slightly more common in S texts, due to the fact that these are often repeated - we have observed a preference for using NE instead of referring expressions like pronouns or definite noun phrases in S texts (see Section 4.2). We should also acknowledge that low frequency words (FI 3) are used around half as much in S as in O texts, while the former is somewhat more saturated in words with the highest frequency rate, in line with our predictions.

Frequency index	Original	Simple
Rare words	9.49%	4.19%
NE	7.08%	8.77%
NumExp	2.81%	2.02%
Freq. 0 total	19.38%	14.98%
Freq. 3	1.23%	0.66%
Freq. 4	1.21%	0.89%
Freq. 5	6.02%	5.06%
Freq. 6	72.16%	78.40%

Table 2: The distribution of n-frequency words in original and simplified texts

If, additionally, we analyse the word length of rare words, we notice that the majority of these (72.44% in O and 77.44% in S) are words made up of seven to nine characters, followed in percentage by longer words of up to twenty characters in O texts (39.42%) and fourteen characters in S texts (29.88%). We could, therefore, draw a general conclusion that longer words tend to be used more sparingly in S texts and that the combination of factors such as frequency and word length might be the one to be taken into account when carrying out lexical substitution based on synonymy.

4.2 Named Entities and Numerical Expressions

Examining the parallel corpus, we have observed that NumExp and NE are given special attention when simplifying texts for people with cognitive disabilities. We have documented numerous cases of such expressions, as well as changes applied to them. One common operation is the substitution of a definite noun phrase with a NE it refers to. For example, *the Andalusian town* is substituted with *Granada*. As for NumExp, a good example of common simplification operations are

[3] Words treated here are the result of processing with FreeLing, where multi-word expressions, among them named entities or numerical expressions, are treated as single words – hence words of more than 20 characters in our corpus.

rounding of big numbers, eliminating NumExp from parenthesis and the use of numerical modifiers, such as *almost* or *more than*, all three illustrated in the following pair of original (1) and simplified (2) sentences:

1. *The Secretary General of the UN, Ban Kimoon, asked for major funding for humanitarian actions in 2011, with a petition of* ***almost 7,400 million dollars*** *(****around 5,400 million euros****).*
2. *The Secretary General of the UN asked for* ***more than 7,000 million dollars*** *for humanitarian actions.*

However, our data shows that by far the most common operation applied to NumExp and NE is elimination. Almost 60% of the original NumExp have been eliminated as a result of simplification. In the case of NE, the average number of NE in simplified texts is slightly higher than in original ones (8.77% in S and 7.08% in O). This, however, is due to the fact that NE seen as essential for the core message of the text are often repeated, both through the introduction of definitions of such terms and the use of NE instead of a definite noun phrase, as already seen. When the number of *different* NE are counted in each set, we perceive a strong tendency towards elimination – S texts contain half as many NE types as O texts. The following sentences illustrate a case of NE elimination (the eliminated expressions are printed in bold).

1. *Today the Mayor of* ***Madrid****,* ***Alberto Ruiz-Gallardón****, inaugurated the new library, situated in the* ***Cultural Centre Eduardo Úrculo*** *and dedicated to the philosopher* ***María Zambrano****; the library caters for six neighbourhoods in the district of Tetuán.*
2. *The new library is in the Tetuán district.*

4.3 Sentence Elimination

Content reduction in text simplification is not only observed in the elimination of certain phrases such as NumExp or NE, as already suggested, but also in the deletion of full sentences. As our corpus study indicates, 20% of all sentences in O texts are deleted to create S texts. Even though one could argue that this percentage is too small to justify implementing a deletion operation, it is a striking fact that 72% of O texts in our corpus contain at least one case of sentence elimination. Therefore, we argue that a sentence deletion procedure should be a key element in making texts simpler, since it is indeed a very productive operation. The module to simplify content through sentence deletion is implemented as a sentence classification mechanism: it decides which sentences from O texts to delete, the data for training the classifier being the set of all O sentences annotated with a feature indicating whether the sentence should be deleted or kept. Every sentence in the corpus is represented as a set of features, some of them borrowed from text summarisation and others specific to our problem. For example, we consider that the position of the sentence in the text may be a factor when deciding whether to delete it or not. In fact, in the informative discourse we are treating, less "topical" sentences would likely appear towards the end of the document, being therefore good candidates for deletion. Other features we are considering are the number of NE and NumExp in the sentence (justified by our corpus study), the number of content words in the sentence, and the number of punctuation tokens. Various cohesion features are computed as the number of shared content words units between neighbouring sentences: this is done to implement topic shifts. Word frequency distribution is also used as a feature. Average word frequency is calculated for every sentence and this information is used as one of the features for the classifier. The classification system is based on a Support Vector Machines implementation (Li et al., 2002) that can be used for training, testing, and cross-validation experiments. We have considered two simple baseline (non-trainable) procedures which delete the last or last two sentences of each document (See Table 3). One of the baselines already provides a very reasonable performance with an F-score (F1)of 0.73. However, our more informed classifier, trained with our designed features, reaches an improved F-score of 0.79 in cross-validation experiments, improving on both precision and recall of the two baselines. The classifier performance needs to be improved, especially for recognising delete cases.

4.4 Insertion of Definitions

In 57.5% of all texts we found cases of S sentences with no correlation to O sen-

	Delete			Keep			Overall
Condition	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**F1**
Delete last	0.27	0.20	0.23	0.81	0.86	0.84	0.73
Delete 2 last	0.31	0.46	0.37	0.84	0.74	0.79	0.68
Classifier	0.42	0.26	0.30	0.86	0.89	0.87	0.79

Table 3: Results of Cross-validation Sentence Deletion Experiments: Baselines and Classifier

tences. These are all cases of definitions of difficult terms and concepts, inserted in the text as additional information. The majority of these are definitions of named entities, such as personal names (*El Greco*), organizations (*the United Nations*), geographical terms (*Guantanamo*) and alike. A certain number of lexical units are also explained by means of a definition, 80% of which are zero frequency words. Hence, for example, *molecules* are defined as *very small parts of the universe*.

As already shown in previous sections, both named entities and rare words are perceived as complicated. In the majority of cases, such terms are either eliminated or replaced by their synonyms with higher FI. However, when these terms are central to the core theme of the text, they cannot be eliminated. Synonym substitution is not always possible either, since NE do not have synonyms and nor do extremely rare and technical terms (like *molecules* from the above example). This is where definitions are inserted, as a means of simplifying complicated but essential elements of information.

In an attempt to investigate the issue of definition insertion as a possible component of our lexical simplification module, we catalogued all such expressions from the corpus[4] and juxtaposed them to definitions extracted from web sources. We then conducted quantitative analysis of both sets of definitions, termed *long* and *short* to avoid confusion with original and simplified texts. The following pair of sentences are an example of a long definition (1), taken from the web, and a short definition (2) inserted by human editors:

1. *Alhambra is a monumental complex created over the period of more than six hundred years by such diverse cultures as the Muslim, the Renaissance and the Romance culture.*
2. *Alhambra is an Arabic monument in Granada.*

[4]Definitions from the corpus were created from scratch by trained human editors.

We can see that not only does the long definition contain a lot more words, but some of the words are among less frequently used ones, such as *monumental complex* or *the Renaissance*. It is, therefore, clear that simple insertion of definitions found on the web or in encyclopaedias does not necessarily contribute to creating a "simple" text - further simplification of inserted sentences is necessary. In order to test that hypothesis, we analysed average word length and frequency distribution in both sets of definitions. Table 4 provides the obtained figures.

	Long	**Short**
Word length	5.80	2.17
Sentence length	27.74	11.37

Table 4: Word length and sentence length in long and short definitions

As can be appreciated, there is a significant discrepancy in both word and sentence length between short definitions and the ones found on the web. Sentences in short definitions are more than half as long as the ones found in long definitions, and a strong preference for the use of short words is observed in short definitions.

As for frequency distribution, presented in Table 5, we notice a similar pattern as when comparing O and S texts: the majority of the words are high frequency words, whereas the rate of low frequency words is rather negligible. Where the two sets do differ more significantly is the distribution of zero frequency words. The percentage of actual rare words (i.e. not NumExp and NE) is significantly less common in short definitions than in long ones - the latter contain four times as many rare words. NumExp are fairly rare in both sets, with the short set containing only one such example. What we mean by *defined terms* are those terms for which the definition is being inserted, like *Amnesty International* or *molecules*. Since the vast majority of the defined terms are NE, we placed them to-

FI	Long	Short
NumExp	1.61%	0.46%
Defined terms and other NE	6.66%	12.04%
Rare words	9.87%	2.31%
Freq. 0 total	18.14 %	14.81%
freq. 3	0.89%	1.39%
freq. 4	1.61%	0.46%
freq. 5	6.60%	3.70%
freq. 6	72.77%	79.63%

Table 5: The distribution of n-frequency words in long and short definitions

gether in the category of "defined terms and other NE". A somewhat striking initial observation is that such terms are twice as common in the short definition set as in the long one. In order to further analyse the distribution of this category of words in both sets, we divided the set into "defined terms proper", which include both NE and lexical units, and "other NE", which include NE other than the ones being defined. Subsequently, we calculated the number of different NE among the category "other NE", in order to see how repetition influences the number of these words in the definitions. Table 6 summarises the percentages of *defined terms*, *other NE* and *different NE* against the total number of *NE and defined terms*.

Word category	Long	Short
Defined terms	46.67%	80.77%
Other NE	53.33%	19.23%
Different NE	38.33%	19.23%

Table 6: Percentage of defined terms and other named entities in long and short definitions

As can appreciated, there is a stark difference between the two sets – there is more repetition of defined terms in short definitions, reducing the introduction of new named entities to the minimum (only five NE in total, with four different NE). In long definitions, however, the two categories of words are balanced out. In addition to that, the total of 65 long definitions contains 46 different named entities, which is in terms of percentages almost double the number of different NE in the short definition set. The following pair of sentences, introducing the definition of "Congo", are an illustration:

1. ***Democratic Republic of Congo**, previously known as **Zaire** and in the colonial period as **Belgian Congo**, is a country in **Africa**, with the capital in **Kinshasa**.*
2. ***Congo** is a country in **Africa**.*

As can be observed, the long definition (1) uses five different NE, introducing four new ones and changing the form of the NE being defined. On the other hand, the short definition (2) only introduces one extra NE and leaves the defined term unchanged.

Based on the figures analysed above we could draw the following conclusions:

- Definitions should employ short words with higher frequency index.
- Rare words (FI 3 or below) should be avoided.
- The introduction of NE other than those being defined should be avoided.
- A NE or a term being defined may be repeated in order to underline it and allow the reader to memorise it.

As part of our future work we intend to further explore the issue of deciding which terms to define, which to eliminate and where to apply synonym substitution.

5 Conclusions and Future Work

In this paper we presented the results of a quantitative analysis of a parallel corpus of original and manually simplified texts in Spanish, conducted with the aim of observing how word length and frequency act as factors to determine word difficulty and influence the choice of synonyms selection when applying lexical substitution. We have found that almost 95% of the words in simplified texts consist of up to ten characters, whereas original texts contain a larger number of longer words. As for frequency, simplified texts tend to contain a slightly greater number of high frequency words, whereas what we call *rare words* are almost 50% more common in original texts. Additional analysis shows that these words tend to be up to 20 characters long in original and 14 characters long in simplified texts. We could, therefore, conclude that the combination of the factors of word length and frequency could be indicative of word difficulty and could be chosen to guide the process of synonym substitution.

We additionally analysed cases of content expansion through the introduction of definitions of complicated terms, mostly named entities. Our results show that definitions are to be composed of short words with higher frequency index and that introduction of new named entities is to be avoided. Therefore, further simplification of definitions found on the web is to be applied for a truly simplified output. On the other hand, we observed the cases of information elimination, with special attention to numerical expressions and named entities, and found that entire sentences can be safely removed in order to produce a more readable text.

As part of our future work, we intend to further investigate the insertion of definitions, focusing on problems such as what terms to define, what to eliminate and where to apply synonyms substitution. Similarly, the performance of the classifier to recognise delete cases is also to be improved. Our final aim is computational implementation of a lexical simplification module as part of a system for automatic text simplification in Spanish, aimed at readers with cognitive disabilities.

Acknowledgements

We present this work as part of a project entitled Simplext: An automatic system for text simplification, with the file number TSI-020302-2010-84 (`http://www.simplext.es`). We are also grateful to the fellowship RYC-2009-04291 from Programa Ramón y Cajal 2009, Ministerio de Economía y Competitividad, Secretaría de Estado de Investigación, Desarrollo e Innovación, Spain.

References

Anula, A. 2009. Tipos de textos, complejidad lingüística y facilicitación lectora. In *Actas del Sexto Congreso de Hispanistas de Asia*, pages 45–61.

Bautista, S., P. Gervás, and R.I. Madrid. 2009. Feasibility analysis for semiautomatic conversion of text to improve readability. In *The Second International Conference on Information and Communication Technologies and Accessibility*.

Burstein, J., J. Shore, J. Sabatini, Yong-Won Lee, and M. Ventura. 2007. The automated text adaptation tool. In *HLT-NAACL (Demonstrations)*, pages 3–4.

Carroll, J., G. Minnen, Y. Canning, S. Devlin, and J. Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10.

Caseli, H. M., T. F. Pereira, L. Specia, Thiago 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)*.

Cunningham, H., D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.

De Belder, J., K. Deschacht, and Marie-Francine Moens. 2010. Lexical simplification. In *Proceedings of Itec2010 : 1st International Conference on Interdisciplinary Research on Technology, Education and Communication*.

Lal, P. and S. Ruger. 2002. Extract-based summarization with simplification. In *Proceedings of the ACL 2002 Automatic Summarization / DUC 2002 Workshop*.

Li, Y., H. Zaragoza, R. Herbrich, J. Shawe-Taylor, and J. Kandola. 2002. The Perceptron Algorithm with Uneven Margins. In *Proceedings of the 9th International Conference on Machine Learning (ICML-2002)*, pages 379–386.

Padró, Ll., M. Collado, S. Reese, M. Lloberes, and I. Castellón. 2010. Freeling 2.1: Five years of open-source language processing tools. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Quinlan, P. 1992. *The Oxford Psycholinguistic Database*. Oxford University Press.

Specia, Lucia. 2010. Translating from complex to simplified sentences. In *Proceedings of the 9th international conference on Computational Processing of the Portuguese Language*, pages 30–39, Berlin, Heidelberg.

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 21-28 recibido 04-05-12 revisado 04-06-12 aceptado 07-06-12

A Framework for Obtaining Structurally Complex Condensed Representations of Document Sets in the Biomedical Domain

Un marco para la obtención de representaciones condensadas estructuralmente complejas de conjuntos de documentos en el dominio biomédico

Yunior Ramírez-Cruz
Center for Pattern Recognition and Data Mining
Universidad de Oriente
Santiago de Cuba, Cuba
yunior@cerpamid.co.cu

Rafael Berlanga-Llavori
Department of Languages and Computer Systems
Universitat Jaume I
Castellón de la Plana, Spain
berlanga@lsi.uji.es

Reynaldo Gil-García
Center for Pattern Recognition and Data Mining
Universidad de Oriente
Santiago de Cuba, Cuba
gil@cerpamid.co.cu

Resumen: En este artículo presentamos un marco para la obtención de representaciones condensadas estructuralmente complejas de conjuntos de documentos, el cual servirá de base para la construcción de resúmenes, la obtención de respuestas para preguntas complejas, etc. Este marco incluye un método para extraer una lista ordenada de hechos, triplos de la forma entidad - relación - entidad, el cual usa patrones de extracción basados en análisis de dependencias y modelos de lenguajes; y métodos para construir un grafo bipartito que codifique la información contenida en el conjunto de hechos y determinar un orden de recorrido apropiado sobre dicha estructura. Evaluamos los componentes de nuestro marco sobre una subcolección extraída de MEDLINE. Los resultados obtenidos son prometedores.
Palabras clave: minería de textos, recuperación y extracción de información, aplicaciones biomédicas.

Abstract: In this paper, we present a framework for obtaining structurally complex condensed representations of documents sets, which will be used as a base for summarization, answering complex questions, etc. This framework includes a method for extracting a ranked list of facts, triples of the form entity - relation - entity, which relies on dependency parsing-based extraction patterns and language modeling; and methods for constructing a bipartite graph encoding the information contained in the set of facts and determining an appropriate traversing order on that structure. We evaluate the components of our framework on a subcollection extracted from MEDLINE, obtaining promising results.
Keywords: text mining, information retrieval and extraction, biomedical applications.

1 Introduction

Given the exponential growth of the amount of biomedical literature available, clinicians and researchers are forced to use automatic tools to find evidences to support to their tasks and experiments. In biomedicine, PubMed[1] is the main entry point for either users and text-mining applications. Starting from a free-text query, PubMed efficiently returns a list of titles or abstracts in XML format. Unfortunatelly, PubMed relies on boolean queries and results are just ordered by publication date (alternatively by journal, authors and title), which makes it difficult for users to explore the resulting document set.

One of the main retrieval goals of these users is to find relational information about the main entities they handle in their research tasks (e.g. gene, proteins, disease, etc.). Thus, there has been a great interest in developing tools aimed at extracting entity-based relations from the abstracts returned by PubMed.

These efforts may be divided into several classes. First, a number of systems obtain predefined relations between a given

[1] www.pubmed.org

ISSN 1135-5948

type of entities, for example, PubGene[2] for gene-gene relations. Other systems focus on finding co-occurrences between several types of entities, for example, iHOP[3] for co-occurrences between genes and other chemical compounds and EBIMed[4] for co-occurrences between genes, proteins, cellular components, biological processes, molecular functions, drugs and species, which are semantically annotated using ontologies and dictionaries.

These approaches are limited either by the restrictions on the types of entities they handle or the difficulty at extracting the semantics behind relations inferred by co-occurrence statistics, as this kind of information requires a deeper analysis of the sentences where the identified entities participate.

A group of systems apply deeper analysis techniques. For example, MEDIE[5] applies a deep parsing to the abstracts and performs a semantic annotation, which allows users to pose queries on either the subject, the verb and/or the object.

In a previous paper, we presented a first approximation to the extraction of relevant biomedical information in a document set treating a specific focus concept, which consisted on the obtention of a ranked list of *facts*, triples of the form entity - relation - entity, relevant with respect to that focus concept in a document collection conceptually annotated using terminology from the Unified Medical Language System (UMLS) (Bodenreider, 2006). In this paper, we build on that initial approach to propose a more general framework which includes the generation of this ranked fact list and, additionally, includes methods for building a graph-based structure representing the information contained in the entire document set and determining an appropriate navigation strategy within this structure through link analysis algorithms.

Our fact extraction method differs from related approaches in the nature of the information units used for constructing facts, the way they are extracted, and the way they are used. For example, Filatova and Hatzivassiloglou (2003) consider unnormalized named entities (e.g. persons, organizations, etc.) and a few very frequent nouns, whereas we focus mainly on UMLS concepts. For relations, we only consider verbs, whereas they also consider action nouns, as defined by WordNet (Miller, 1995). Finally, we use dependency parsing-based patterns to extract facts, while they use a named-entity tagger and a position-based event extraction heuristics. Filatova and Hatzivassiloglou (2004) extract triples in order to use them as features for other tasks (e.g. calculating a global score in a sentence extraction method), whereas we treat the graph containing the aggregation of the most relevant and distinctive triples as the information-conveying structure on which further tasks will rely.

The rest of the paper is organized as follows. In Section 2 we describe our framework in detail, whereas in Section 3 we experimentally evaluate its components. Finally, we expose our conclusions in Section 4.

2 Framework description

Given a document collection and a focus concept representing an information need, our framework obtains a representation of the set of documents where this concept is mentioned.

In order to obtain this representation, we first obtain a ranked list of facts, triples of the form entity - relation - entity, which describe events that are distinctive of this document set with respect to the collection and relevant with respect to the focus concept. Every fact conveys a very concise piece of information, e.g. *children - develop- uveitis*. Once the ranking has been obtained, a subset of the best ranked facts is selected for constructing the graph structure that represents the most important information extracted from the document set.

In Figure 1, we depict the overall workflow of our framework. As an offline previous step, we construct a document collection C, which is the result of a topic-based query on MEDLINE (e.g. a specific disease). This collection is conceptually indexed using the concepts from UMLS. The result of this step is a conceptual inverted file where each UMLS concept is mapped to the positions in documents where it is mentioned.

Our framework works on the collection C and uses this conceptual inverted file. For

[2] www.pubgene.org

[3] http://www.ihop-net.org/UniPub/iHOP/

[4] http://www.ebi.ac.uk/Rebholz/srv/ebimed/

[5] http://www-tsujii.is.s.u-tokyo.ac.jp/medie/search.cgi

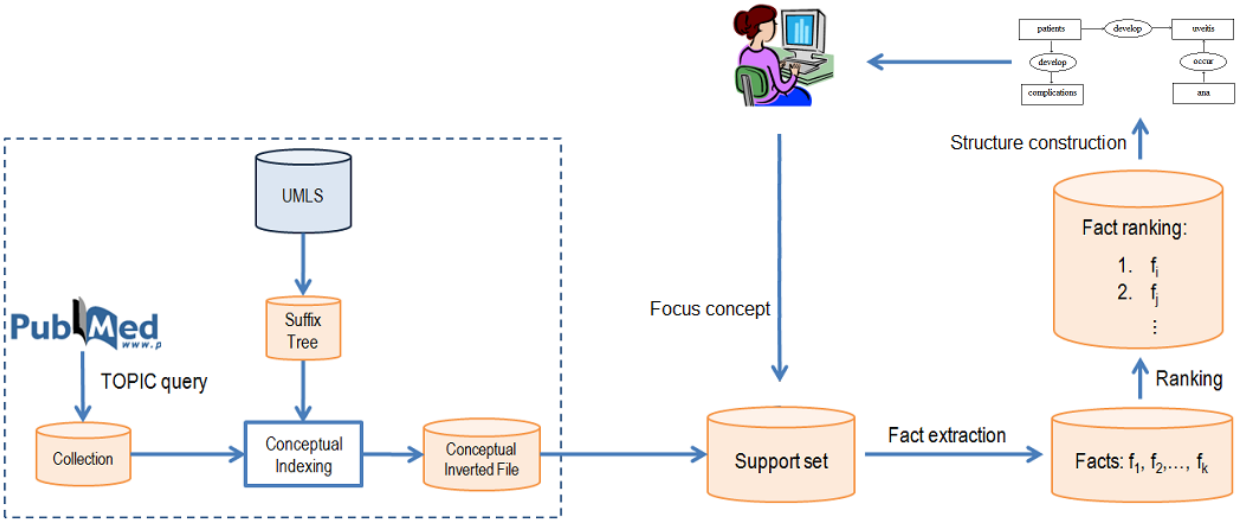


Figure 1: General architecture of our proposal.

a given focus concept, we retrieve the set S of documents from C where it is mentioned, which we call *support set*. The previously described steps are then performed on this support set.

2.1 Building a Conceptually Indexed Collection

Conceptually indexing a collection consists on determining the set of concepts that must be used to describe its contents. In the context of this paper, conceptual indexing allows us to homogenize the terminology used in the medical documents. Additionally, the conceptual index guides the selection of the document sets to be described and semantic relationships may allow to build concept hierarchies to enhance fact extraction with extra knowledge which is not explicitly stated in texts.

As we mentioned before, in this work we use UMLS, specifically version UMLS2008AC, as our knowledge source. The UMLS Metathesaurus is one of the three components of the UMLS Project and comprises many different controlled and well-known vocabularies[6]. Each UMLS concept is linked to a set of synonyms available in the associated vocabularies. In addition, UMLS provides taxonomic relations between concepts.

In order to avoid tagging the entire collection for the occurence of lexical variants of all concepts, the initial, non conceptual inverted index is merged with the lexicon containing the terminology. First, a suffix tree is created containing the entire lexicon. Then, the phrases defined by its paths are used as queries on the collection's single-term inverted index. Thus, a new inverted index is constructed, where entries are Concept Unique Identifiers (CUIs) associated to all documents retrieved by the constructed queries, i.e. those documents that contain some lexical variant of the concept represented by that CUI.

2.2 Fact extraction

As we mentioned previously, facts are simplified representations of the events described in the document set. We consider a fact as a relation between two entities, which is characterized by a verb. Thus, a fact is a triple of the form entity - relation - entity. Here, by *entity* we mean either a lexical variant of a UMLS concept or a non-stopword noun.

Documents are POS-tagged and lemmatized in order to identify verbs and nouns and normalize words into their canonical forms. All occurrences of lexical variants of UMLS concepts are also normalized into the corresponding CUI. For example, *uveitis* and *intraocular inflammation* are both lexical variants of CUI C0042164, so all occurrences of any of them are treated uniformly. No semantic disambiguation is performed on lexical variants, so if a phrase turns out to be a lexical variant of several concepts, it is treated simultaneously as an instance of every concept. In further studies, we will assess

[6] UMLS Source Vocabularies: http://www.nlm.nih.gov/research/umls/metaa1.html

the convenience of applying semantic disambiguation.

Fact extraction is performed in a sentence-by-sentence basis, using the dependencies obtained by a dependency parser in combination with a set of pattern-based extraction heuristics. In this work, we used the Stanford Parser (Klein and Manning, 2003) dependency analysis module (de Marneffe et al., 2006) to obtain dependencies. The following patterns are used:

- **subject - verb - direct complement**
- **subject - verb - indirect complement**
- **subject - verb - prepositional complement**
- **agent complement - verb - passive voice subject**

Since the use of passive voice is very common in scientific literature, the simultaneous use of patterns **subject - verb - direct complement** and **agent complement - verb - passive voice subject** implicitly introduces a simple, partial semantic role labeling-based heuristics allowing to extract facts that follow a general pattern of the form **agent - action - patient**.

For example, the triple *patients - tolerate - etanercept* may be extracted from the sentence *All patients tolerated etanercept with no side effects* as well as from the sentence *Etanercept is well tolerated by pediatric patients.*

When applying the extraction patterns, multi-word lexical variants of UMLS concepts are considered to be good matches for a member of the triple if any of its constituent words is labeled with one of the syntactic dependency tags used in the pattern.

If the subject or any of the used complements is a coordination of several noun phrases linked by the conjunctions *and* or *or*, as many facts are extracted as members of the coordination.

For example, the triples *etanercept - demonstrate - safety* and *etanercept - demonstrate - efficacy* are both extracted from the sentence *Etanercept has demonstrated excellent safety and efficacy in large scale randomised double blind placebo controlled trials.*

Finally, a verb and its negation are treated as a different relation so different facts will be extracted from dependencies where they occur, even if the same entities are involved.

2.3 Initial fact ranking

Two criteria are to be considered in creating a ranked fact list. They must be both relevant to the focus concept according to which the support set was constructed and distinctive with respect to the collection. In order to create a ranking where both criteria are simultaneously considered, we follow a language modeling approach. We construct the unigram models of the set of terms (entities and verbs) in both the support set S and the collection C, as well as the language models of the facts in the support set and the collection.

The unigram model of the collection, M_C, is estimated by maximum likelihood (ML). Thus, for a term t:

$$P(t \mid M_C) = \frac{count(t)}{\sum_{t' \in V} count(t')} \qquad (1)$$

where V is the vocabulary of the collection and $count(t)$ indicates the number of occurrences of t in the collection.

Since the support set being described is focused on a concept, we take this into account for estimating its unigram model in such a way that it is *biased* towards the focus concept. We express the biased unigram model of the support set, $M_{S_{biased}}$, as a mixture of three components: the ML unigram model of the set of sentences in S where some lexical variant of the focus concept occurs, M_{focus}, the ML unigram model of the set of sentences in S where some lexical variant of either the focus concept or its immediate hyponyms in the UMLS concept hierarchy occur, M_{exp}, and the ML unigram model of the support set S itself, M_S.

Unlike common language modeling approaches, where mixture models are used for smoothing or modeling the presence of several underlying topics in the documents, in our approach the mixture is used as a mechanism to favor the selection of terms coocurring with lexical variants of the focus concept and/or its related concepts. Notice that the sentence set from which M_{focus} is estimated is a subset of the sentence set from which M_{exp} is estimated, which is in turn a subset of S.

For instance, if the focus concept is C0042164, the sentences containing *uveitis* or *intraocular inflammation* will be considered for estimating M_{focus}, whereas the sentences

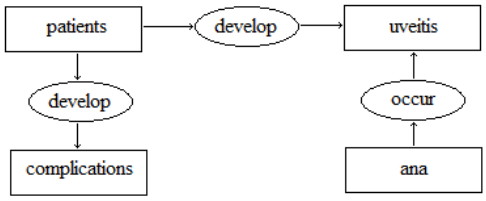


Figure 2: Example of the graph generation method.

containing *uveitis*, *intraocular inflammation*, *anterior uveitis*, *intermediate uveitis*, *posterior uveitis*, *panuveitis* or *diffuse uveitis* will be considered for estimating M_{exp}. The latter may be seen as a form of concept hierarchy-based query expansion.

Finally, the occurrences of terms in sentences not containing lexical variants of neither the focus concept nor any of its immediate hyponyms will only be accounted for when estimating M_S. Since the three components contribute to the focus concept-biased model $M_{S_{biased}}$, the estimated probability of terms in the context of the focus concept and/or its immediate hyponyms will be increased at expense of the estimated probabilities of non coocurring terms.

Thus, the probability of a term t in $M_{S_{biased}}$ is calculated as:

$$\begin{aligned} P(t|M_{S_{biased}}) &= \lambda_0 P(t|M_{focus}) + \\ &+ \lambda_1 P(t|M_{exp}) + \\ &+ \lambda_2 P(t|M_S) \end{aligned} \quad (2)$$

where $\lambda_0 + \lambda_1 + \lambda_2 = 1$.

The language models of the set of facts in the collection and the support set, M'_C and $M'_{S_{biased}}$, are estimated in a similar way.

Two criteria are considered when ranking facts: first, the triple representing the fact must be distinctive as a whole; second, the three terms composing the triple must be distinctive as well.

For a term, or a triple representing a fact, we use its contribution to the Kullback-Leibler (KL) divergence between the language model of the support set and that of the collection as a measure of how distinctive the term or triple is. The contribution of a term t to the KL divergence between $M_{S_{biased}}$ and M_C is defined as:

$$KLC(t) = P(t|M_{S_{biased}}) \log \frac{P(t|M_{S_{biased}})}{P(t|M_C)} \quad (3)$$

Notice that KLC values above zero characterize terms that are more frequent according to $M_{S_{biased}}$ than according to M_C, thus being distinctive terms of the support set. Also notice that as KLC values grow, terms may be considered more distinctive.

The contribution of a fact $f = (e_1, r, e_2)$ to the KL divergence between $M'_{S_{biased}}$ and M'_C is calculated similarly.

Since we intend to rank facts according to the distinctiveness of both the triples by which they are represented and that of the terms conforming these triples, we calculate the score of a fact $f = (e_1, r, e_2)$ as

$$\begin{aligned} score(f) &= KLC(f) * KLC(e_1) * \\ &* KLC(r) * KLC(e_2) \end{aligned} \quad (4)$$

2.4 Constructing and traversing a global structure

In order to make the extracted information navigable, as well as facilitating further tasks, such as summarization, complex question answering, etc., we construct a structure where all relevant information is aggregated, thus allowing to consider global scale interactions between entities and relations.

Graphs have been widely used for representing entity-relation information in a structured way. Following this line of thought, we aggregate all information in the ranked fact list into a bipartite graph. In this graph, a first set of nodes represents the entities and a second set represents the relations.

Every entity occurring in the ranked fact list is represented by one node in the graph. Relations are not treated in the same way. For every fact a relation is involved in, a new node representing this occurrence of the relation is added. Finally, for every fact (e_1, r, e_2), edges are included linking the node representing e_1 to the node representing the corresponding occurrence of r and this node to the node representing e_2. Both edges are

weighted by the score of the fact. Notice that no edge links e_1 and e_2 directly. Adding a different node to represent every occurrence of a relation prevents the structure from encoding inconsistent information. For example, if a single node is used to represent every relation, the subgraph obtained by adding the facts (e_1, r, e_2) and (e_3, r, e_4) would be the same as the one obtained by adding the facts (e_1, r, e_4) and (e_3, r, e_2), which is not desirable. To better illustrate the graph construction process, Figure 2 shows an example of the graph obtained for the set of facts

patients - develop - uveitis
ana - occur - (in) uveitis
patients - develop - complications

In order to determine a convenient presentation order, we take into account both the scores obtained at creating the original ranked fact list and the sctructural importance of nodes in the graph.

It is important to notice that, while the first ranking aims at obtaining the most relevant and distinctive facts, i.e. determining which information to include in the representation; this second ranking aims at determining a convenient order for presenting and navigating the information conveyed by these facts. Since fact scores are used for weighting the edges in the graph, the second ranking is not unaware of informational relevance when determining structural importance.

Structural importance of entities and relations is assessed via a link analysis algorithm on the graph. In our framework, we use a variation of PageRank for weighted graphs, which is defined as follows (Mihalcea, 2004):

$$\begin{aligned} PR(v_i) &= (1-\alpha) + \\ &+ \alpha \sum_{v_j \in In(v_i)} w_{ji} \frac{PR(v_j)}{\sum_{v_k \in Out(v_j)} w_{kj}} \end{aligned} \quad (5)$$

where $In(v)$ represents the set of nodes v_i such that there exists and edge (v_i, v), $Out(v)$ represents the set of nodes v_i such that there exists and edge (v, v_i), w_{ij} represents the weight of the edge linking node v_i to v_j, and parameter α expresses how much importance is given to the graph structure and is normally set to 0.85.

Adding nodes to represent both entities and relations allows us to obtain scores for all of them, not only for entities. Although a single relation may be represented by several nodes in the graph, the final measure we use for determining its structural importance is the sum of PageRank values over all the nodes representing it, which we will refer to as *aggregated* PageRank.

Once final scores have been obtained, the presentation order to be used is determined by a breadth-first traversal of the graph in the following manner. First, the entity-representing node having the greatest PageRank value is selected as the starting point v_{e_s}. Let $v_{r_1}, v_{r_2}, \ldots, v_{r_k}$ be the relation-representing nodes linked to v_{e_s} and $v_{e_1}, v_{e_2}, \ldots, v_{e_k}$ the corresponding entity-representing nodes linked to them. For every pair v_{r_i}, v_{e_i}, if $AggrPR(v_{r_i})$ or $PR(v_{e_i})$ are below given thresholds, the fact (e_s, r_i, e_i) is discarded for presentation. For every remaining fact (e_s, r_i, e_i) to be considered, a new score $score_{gr}$ is calculated as follows:

$$score_{gr}(f) = AggrPR(v_{r_i}) * PR(v_{e_i}) \quad (6)$$

Facts are ordered for presentation in descending order of this score. Notice that $PR(v_{e_s})$ is not considered since it does not affect the ordering. Following this order, every newly reached entity-representing node is then taken as starting point and the process is repeated recursively until all includible facts have been added to the final ordering.

3 Evaluation

There are different considerations to take into account at evaluating the components of our framework. In the case of the initial ranked fact list, traditional Information Retrieval quality measures may be used as good indicators of the performance of the method. On the other hand, evaluating navigability or appropriateness of a given presentation order is not trivial since these notions are not well specified and are difficult to quantify.

In our experimental setting, we constructed a conceptually indexed collection by retrieving from MEDLINE documents that satisfy the query *juvenile idiopathic arthritis* (*JIA*). This collection is composed by 7654 documents (45672 sentences), which are described by 32350 terms, out of which 12572 represent lexical variants of UMLS concepts found during conceptual indexing.

Three support sets were retrieved according to the focus concepts C0177758 (*etaner-*

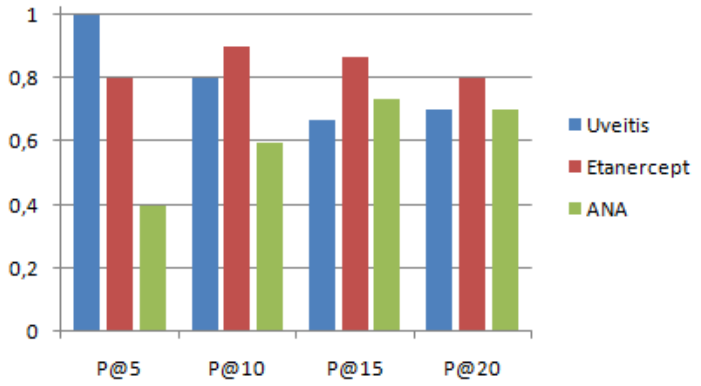


Figure 3: Precision at top elements for the three support sets.

cept), a drug used for treating JIA; C0042164 (*uveitis*, *intraocular inflamation*), a complication of JIA; and C0003243 (*antinuclear antibody*), an indicator of the presence of the disease.

The parameters in Equation 2 were empirically set to $\lambda_0 = 0.7$, $\lambda_1 = 0.2$ and $\lambda_2 = 0.1$. After fact rankings for each support set were constructed according to the proposed method, the 20 top-ranking facts in each case were manually evaluated, labeling them as relevant or not relevant.

The quality of the rankings was measured in terms of precision at k top ranking elements ($P@k$), a typical IR measure, which is defined as:

$$P@k = \frac{\#\text{ of relevant facts in top } k}{k} \quad (7)$$

The nature of the problem makes it impossible to define the entire set of relevant facts, which prevents us from using metrics depending on it, such as recall or average precision.

Figure 3 shows the results obtained for the three support sets for $k \in \{5, 10, 15, 20\}$.

A manual inspection of the rankings allowed us to determine that the main cause for the extraction of incorrect facts was the effect of dependency parsing errors, which mislead the extraction rules. The error that most commonly affected fact extraction was the incorrect attachment of prepositional phrases modifying a noun phrase, which were instead attached to the clause main verb as a prepositional complement. Some of these erroneous facts reached a high position in the ranking because of two reasons. First, their noisy nature, which makes them extremely unfrequent in the entire collection, thus obtaining high KLC values. Second, the occurrence of high KLC-valued entities in the fact. The combination of both circumstances is likely to make these facts obtain high scores. Although we observed cases of this situation for all three focus concepts, it occurred particularly often in the fact ranking obtained for focus concept *antinuclear antibody*.

In our opinion, the values at which $P@k$ appears to stabilize, around 0.7, are reasonably good, although there is still room for improvement in the fact extraction patterns and the ranking score formula.

As we mentioned previously, it is hard to define measures of how navigable or purpose-fit a particular fact presentation order is. In order to illustrate the performance of the graph-based structuration method, in Figure 4 we show a fragment of the ordering obtained for focus concept C0717758 (*etanercept*), setting the fact discarding thresholds at 0.01. As it may be observed, the presentation order first provides all facts having as subject the entity that emerges as the most important in the graph structure (which does not necessarilly mean that it is the most relevant with respect to the focus), then provides this information for the entities that are introduced by facts linking them to the chosen entity, and so on. We consider this behavior to be useful, as it may provide a good paragraph structure for future text generation methods.

4 Conclusions

In this paper, we have presented a framework for obtaining structurally complex condensed representations of document sets in the biomedical domain. Facts, concise information units conveying information about

```
patients (C0030705) -- tolerate -- etanercept (C0717758)
patients (C0030705) -- tolerate -- treatment (C0087111)
patients (C0030705) -- tolerate -- (with no) side effects (C0001688) (incorrect fact)
etanercept (C0717758) -- demonstrate -- efficacy (C1707887)
etanercept (C0717758) -- demonstrate -- safety (C1705187)
etanercept (C0717758) -- demonstrate -- beneficial activity (C0600075)
etanercept (C0717758) -- approved -- (in) europe (C0015176)
etanercept (C0717758) -- approved -- (in) united states (C0041703)
[...]
```

Figure 4: Fragment of the ordering obtained on the graph generated for focus concept *etanercept*.

relations held between entities, are the base from which a graph structure representing the document set is constructed.

Facts are extracted by simple dependency parsing-based patterns and ranked by their relevance and distinctiveness in the document set using a Language Modeling approach. Link analysis algorithms are used in order to determine a presentation order over this graph, which arguably facilitates tasks such as summarization, complex question answering, etc.

Despite the simplicity of the fact extraction procedure, experimental results, obtained over three different document sets from a subcollection of MEDLINE, are encouraging. We have presented a case study of the graph construction method and the obtained ordering, which we intuitively consider to be sound and useful. However, a principled evaluation criterion for the quality of the proposed presentation order is still required.

While our method has been initially proposed for the biomedical domain, we consider that it may be ported to other domains for which rich knowledge resources are available.

In addition to the previously mentioned need for an evaluation criterion for the presentation order, other attractive directions for future work include improving fact extraction mechanisms, mainly by taking into account the semantic nuances introduced by the use of different syntactic patterns, leading prepositions in prepositional phrases, etc. Besides, we intend to use semantic relations contained in the concept hierarchies to constrain and/or generalize the initial set of facts to be considered and/or enrich it with non explicit information.

References

Bodenreider, O.: 2006. Lexical, Terminological, and Ontological Resources for Biological Text Mining. In *Text Mining for Biology and Biomedicine*. Artech House.

Filatova, E. and V. Hatzivassiloglou: 2003. Domain-Independent Detection, Extraction, and Labeling of Atomic Events. In *Proceedings of RANLP 2003*, pages 145–152, Borovets, Bulgaria.

Filatova, E. and V. Hatzivassiloglou: 2004. Event-Based Extractive Summarization. In *Proceedings of the ACL 2004 Workshop "Text Summarization Branches Out"*, pages 104–111, Barcelona, Spain.

Klein, D. and C. D. Manning: 2003. Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423–430.

de Marneffe, M. C., B. MacCartney and C. D. Manning: 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of LREC 2006*.

Mihalcea, R.: 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. *Proceedings of the ACL 2004 Interactive Poster and Demonstration Sessions*.

Miller G. A.: 1995. WordNet: a Lexical Database for English. *Communications of the ACM*, 38(11): 39–41.

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 29-38 recibido 05-05-12 revisado 28-05-12 aceptado 04-06-12

Sistema de Acceso a la Información basado en conceptos utilizando Freebase en Español-Inglés sobre el dominio Médico y Turístico

Information Access System based on concepts using Freebase in Spanish-English over the domain Medical and Tourist

Rafael Muñoz
Universidad Europea de Madrid
C/ Tajo s/n 28670
Villaviciosa de Odón
Madrid

Fernando Aparicio
Universidad Europea de Madrid
C/ Tajo s/n 28670
Villaviciosa de Odón
Madrid

Manuel de Buenaga
Universidad Europea de Madrid
C/ Tajo s/n 28670
Villaviciosa de Odón
Madrid

{rafael.munoz, fernando.aparicio, buenaga@uem.es}

Resumen: En este artículo presentamos una herramienta de acceso a la información, basado en los conceptos, enfocada tanto a textos médicos como turísticos. Usando técnicas para el marcado de entidades reconocidas, el sistema permite extraer conceptos relevantes para aportar más información sobre ellos utilizando bases de conocimiento colaborativas y ontologías.

Componentes especialmente interesantes para el desarrollo del sistema son Freebase, una gran base de conocimiento colaborativa, además de recursos formales como MedlinePlus y PubMed. La arquitectura del sistema ha sido construida pensando en términos de escalabilidad, para constituir una gran plataforma de integración de información, con los siguientes objetivos: permitir la integración de diferentes técnicas de procesamiento de lenguaje natural y ampliar las fuentes desde las que se extrae información, así como facilitar la integración de nuevas interfaces de usuario.

Palabras clave: Extracción de Información, Integración de Información, Bases de datos Colaborativas, Procesado de Textos, Arquitectura Escalar, Freebase

Abstract: In this paper we present a tool for access to information, based on semantic, focused both medical texts and tourists. Using marking techniques for recognized entities, the system can extract relevant concepts to provide more information about them, using collaborative databases and ontologies.

Particularly relevant components to its the development are Freebase, a large collaborative base of knowledge and formal resources such as MedlinePlus and PubMed. The platform architecture has been built thinking in terms of scalability, in order to constitute a great platform for information integration, with the following objectives: to allow the integration of different natural language processing techniques, to expand the sources from which information extraction can be performed and to ease integration of new user interfaces.

Keywords: Information Extraction, Information Integration, Collaborative Databases, Text Processing, Scalable Architecture, Freebase.

1 Introducción

Los sistemas de acceso a la información se nutren en la actualidad de un número creciente de diversas fuentes externas como son las ontologías, almacenamientos heterogéneos o incluso redes de conocimiento colaborativas (Gutiérrez V. Y. et al, 2010). En los últimos años se ha incrementado notablemente la cantidad de sitios donde poder acceder a información (Chang C. H. et al, 2006). Esto hace que la cantidad de información sea cada vez mayor y más heterogénea, que se encuentre estructurada o desestructurada, haciendo

ISSN 1135-5948

cada vez más necesarios los sistemas de búsqueda y recuperación de información efectivos (Allan J. et al, 2005). Estos evolucionan rápidamente y ya no solo tienen como objetivo que muestren una enorme cantidad de información, sino que la información sea lo más útil posible (Egozi O. et al, 2011). Es por ello que lo que se impone son los sistemas de integración de información (Tuchinda R. et al, 2011), donde se aúnan diversas técnicas de anotación semántica, análisis de la información, recomendación y personalización. Todo esto para facilitar el acceso más que a una gran cantidad de información no relacionada, a información precisa y que le permita ampliarla profundizando en la cadena de búsqueda.

En este artículo presentamos un sistema de acceso a la información basado en la semántica, que sienta las bases del desarrollo de un sistema mayor de integración de información. Los dominios de aplicación escogidos son el biomédico y el turístico. El primero de ellos con el objetivo de conseguir un sistema destinado a introducirse en el campo de la Medicina Personalizada (Hamburg M et al, 2010) y el segundo, seleccionado por la heterogeneidad del vocabulario y la diversidad de formatos de la información almacenada, generalmente multimedia (Lew M. S. et al, 2006) (texto, fotos, video…).

El sistema propuesto en este artículo permite el acceso a la información a través de la identificación de conceptos relevantes y la integración del conocimiento sobre dichos conceptos, almacenados en la base de conocimiento colaborativa Freebase, que son procesados con técnicas lingüísticas computacionales a través del sistema GATE (Cunningham H. 2002). El resultado final es un sistema que ayuda a optimizar el tiempo dedicado a comprender el texto y ampliar fácilmente la información relevante contenida en él.

El resto del artículo está organizado como sigue. La sección 2 trata sobre los asuntos relacionados con la extracción de información de la base de datos Freebase, las ontologías y la indexación de conceptos. La sección 3 está dedicada a la arquitectura del sistema. La Sección 4 describe el procesamiento de información, se ilustra el interfaz Web a través de dos ejemplos de los dominios seleccionados. Finalmente se exponen las conclusiones, así como futuros trabajos.

2 Uso de Freebase en el acceso semántico

Freebase (Bollacker K. et al, 2008) es una gran base de datos de conocimiento colaborativa, publicada en 2007 por Metaweb y recientemente adquirida por Google.

En nuestra propuesta, Freebase es usada para recuperar listas de conceptos médicos o turísticos para el reconocimiento de las entidades mencionadas (Nadeau D. and Sekine S., 2007) en textos y conectarlos con contenidos semánticamente relacionados. Podríamos categorizar a Freebase como una ontología, ya que los términos se encuentran perfectamente clasificados.

La importancia de las ontologías referentes a dominios particulares es ampliamente reconocida. El objetivo que tenemos al utilizar una ontología es reducir al máximo la confusión entre conceptos. En el ámbito médico, por ejemplo, entre profesionales de la medicina, o en el dominio del turismo, entre operadores turísticos (Navigil R. and Velardi P. 2004).

En nuestro sistema combinamos Freebase con ontologías propias de los dominios. Presentamos una propuesta de método de acceso en dos idiomas, inglés y español gracias a la característica bilingüe de Freebase. Esto nos permite introducir los textos en cualquiera de los idiomas, devolviéndonos los resultados en el idioma origen de la fuente.

2.1 Escenarios de aplicación Cross-Lingüe

La recuperación de información cross-lingüe (CLIR) está relacionada con la recuperación de datos en idiomas diferentes al utilizado por el usuario, facilitando el acceso a recursos por otros criterios como la similitud o la calidad. Este tipo de tareas de recuperación es uno de los objetivos destacados en la iniciativa CLEF[1] y en otras de más reciente creación, como las mencionadas en (Mayfield et al., 2011).

[1] http://www.clef-initiative.eu/

Uno de los recursos más utilizados para la recuperación de información cross-lingüe son los contenidos de la Wikipedia, u otros más recientes como el Wiktionary (Müller y Gurevych, 2009). Las ontologías permiten el almacenamiento de datos en diferentes idiomas, conectando los conceptos a través de meta-información. Esto las convierte en una herramienta muy útil en la construcción de sistemas cross-lingüe (Carrero et al., 2007; Knoth et al., 2010). Freebase es un sistema que relaciona ambos ámbitos, es decir, es una ontología construida a partir de la extracción multilingüe de información desde la Wikipedia y otras fuentes (De Melo y Weikum, 2010).

El sistema desarrollado parte de un caso clínico en inglés para, a través del procesamiento del texto, extraer un conjunto de conceptos (en inglés) sobre los que es posible ampliar información (también en inglés). Sin embargo, este es sólo uno de los escenarios posibles. Cada uno de estos elementos puede ser llevado a cabo en inglés o en otros idiomas, como por ejemplo el español.

La generación de listas de conceptos procedentes de Freebase o de Medlineplus en ambos idiomas (en el formato apropiado para su procesado con el elemento Gazetteer de GATE), proveen de información detallada de los conceptos detectados ofrecida en inglés, mientras que los textos de entrada y los conceptos pueden estar en ambos idiomas. Además, dado que Medlineplus también proporciona los contenidos en español a sus usuarios, y que la información detallada de Freebase está fundamentalmente basada en la Wikipedia, es posible acceder a información de detalle mostrada en español.

Estos escenarios bilingües pueden contribuir al enriquecimiento de la metodología de aprendizaje publicada en (Aparicio et al., 2011a), tal y como ya se ha hecho en otras experiencias educativas. Un ejemplo próximo se puede encontrar en (Clark et al., 2012), donde se utiliza este tipo de sistema bilingüe para la realización de una experiencia con alumnos de un centro de educación secundaria en Estados Unidos, obteniéndose resultados muy interesantes en cuanto a la utilidad de la experiencia (ofrecer los recursos en ambos idiomas sirve tanto para facilitar la comprensión de la materia como para la mejora del idioma no nativo).

2.2 Ontologías turísticas y biomédicas

Turísticas: La proliferación de ontologías turísticas ha desembocado en el desarrollo y profundización de las mismas, llevándonos a realizar ontologías de ámbito regional.

Seguidamente se enumeran las ontologías turísticas más representativas a día de hoy:

Harmonise, IMHO (Interoperable Minimum Harmonization Ontology), Hi-Touch, Tourist Ontology por AIFB, Mondeca, Qall-Me. Existen un par de Ontologías Españolas: Cruzar, y ANOTA.

Estas ontologías contienen conceptos del dominio turístico y actividades de ocio, abarcando entidades turísticas, culturales, paquetes turísticos e incluso contenido multimedia.

En el ámbito de la biomedicina existen muchos ejemplos de ontologías desarrolladas a lo largo de los últimos años. Algunos ejemplos destacados, de este amplio conjunto de recursos biomédicos, son: GO (Gene Ontology), UMLS (Unified Medical Language System), SNOMED-CT (Systematized Nomenclature of Medicine - Clinical Terms) o FMA (Foundational Model of Anatomy).

2.3 Freebase

El uso de Freebase en este sistema se debe a que la información almacenada se encuentra estructurada. Esta se agrupa en "dominios" (como medicina, viajes, deportes, etc). Dentro de los dominios existen los "tópicos", que almacenan toda la información de ese concepto y tiene una relación semántica con tipos y propiedades de la forma "es un" tipo o "tiene una" propiedad. Los dominios, tipos y propiedades tienen un identificador único, consistente en la concatenación del tipo o el tópico, como se muestra en los dos siguientes ejemplos. (A) El tópico "autism" pertenece al dominio *medicine* y tipo *disease* de ahí su identificador es /medicine/disease. (B) El tópico "Taj Mahal" pertenece al tipo *tourist attraction* que se encuentra englobado en el dominio *travel*, por lo tanto su identificador es /travel/tourist attraction. Freebase ha sido usado en otras áreas de investigación, como piezas de trabajo relacionado a software en el contexto de la Web 2.0 y 3.0, herramientas para la desambigüación de

nombres (Han X. and Zhao J., 2009), clasificación de consultas (Brenes D. J. et al, 2009).

El sistema consigue listas de *diseases*, *symptoms* y *treatments* del dominio *medicine* y *tourist attraction*, *accommodation*, *travel destination* del dominio *travel* extraído de Freebase para el posterior reconocimiento usando la herramienta GATE (Aparicio F. 2011b).

2.4 Indexación Conceptual

Dependiendo del tipo de documentos a analizar, la indexación conceptual requiere solventar problemas como: desambiguación semántica, relaciones semánticas, traducción automática, resolución de polisemia, etc. (Verdejo F. et al 1999)

Nuestro sistema realiza un proceso de indexación conceptual en el que se procesan los textos y se identifican los tópicos (conceptos) de Freebase que aparecen en ellos, marcándolos para que a través de hipervínculos se pueda profundizar en el conocimiento (Voss A. et al, 1999).

Para el objetivo propuesto para este sistema, existen algunos de aspectos que obtienen mayor relevancia. Así, contar con un número de términos representativo (14.000 médicos y más de 4.000 turísticos) en los que profundizar en el conocimiento, y que sea una aplicación plenamente funcional utilizada por usuarios reales permitiéndoles el reconocimiento de entidades nombradas es el objetivo propuesto y conseguido.

Para este proceso de textos orientado a la indexación conceptual, utilizamos como elemento principal el sistema GATE (Generic Architecture for Text Engineering) (Cunningham H. et al, 2002).

Entre las diferentes opciones de programación para integrar en este software, hemos seleccionado una que admite el desarrollo y prueba con la GUI y reutiliza la lógica del módulo NLP. GATE es distribuido incluyendo un sistema de extracción de información llamado ANNIE (A Nearly-New Information Extraction System), incorporando un amplio rango de recursos que realizan tareas del análisis del lenguaje a diferentes niveles.

Gazetteer es uno de sus componentes, al que se le ha dado una importancia especial en nuestro diseño del sistema. Este componente, basado en listas predefinidas, permite el reconocimiento de entidades previamente mencionadas. Estas listas, a su vez, permiten la inclusión de detalles de cada entidad, que en nuestro caso son principalmente usados para almacenar los identificadores Freebase.

2.5 Arquitectura

Con respecto a la arquitectura del sistema, se ha puesto un particular énfasis en la creación de una arquitectura de software habilitando el almacenamiento y la presentación online de nuevas fuentes de información. Esto enriquecerá la interface del usuario, así como la incorporación de nuevas técnicas en el procesamiento de palabras y la generación de interfaces de usuario para diferentes tipos de clientes tales como dispositivos móviles u otros que requieran acceso a través de servicios web. El principal objetivo de este diseño del sistema es hacer la integración de diferentes componentes fácilmente agrupándolos en diferentes módulos. La modularización de componentes tiene los siguientes objetivos específicos:

- Acceso: recopila los diferentes mecanismos para acceder al sistema a través del protocolo HTTP, tal como el producido por interactuar desde la interfaz web del usuario o el ofrecido a través de servicios web. Es el responsable de la comunicación entre aquellos módulos dedicados a procesamiento del lenguaje natural y los de búsqueda de información asociada a un concepto particular.
- El módulo de procesamiento de lenguaje natural habilita al sistema a usar otras herramientas o librerías, manteniendo la misma estructura y haciendo posible la interacción entre diferentes componentes. Igualmente, hace uso de la lista de conceptos obtenidos a procesar texto usando librerías GATE.
- El módulo de recuperación de información extrae la información en tiempo de ejecución del sistema, optimizando el tiempo de respuesta online.
- Búsqueda: provee al sistema con un interfaz para que diferentes fuentes de búsqueda de información, asociadas a conceptos que aparecen en textos de entrada, puedan ser añadidos.

La Figura 1 muestra la arquitectura y el flujo de información generado cuando el usuario interactúa a través del interfaz Web.

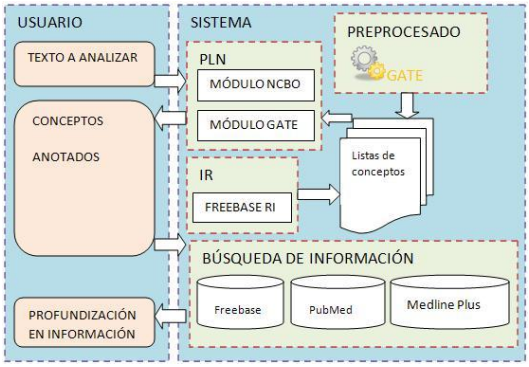


Figura 1: Arquitectura del sistema

3 Casos de Uso

Hemos escogido dos dominios diferentes, cada uno con ciertas características propias, en los que aplicar nuestro sistema, ambos ampliamente utilizados, el dominio biomédico y el dominio turístico.

Los textos biomédicos tienen ciertas características que los diferencian de los textos de otras disciplinas. Las peculiaridades de terminología y estilo de escritura usado por los profesionales médicos hacen la detección de conceptos y el análisis de los datos una tarea compleja y ambiciosa.

En el caso del turismo, el léxico utiliza una amplia terminología de varios campos (geografía, economía, arte, historia, etc.). Es, por lo tanto, un lenguaje poco especializado, que usa un vocabulario amplio y habitual.

Se han creado dos mecanismos de acceso a nuestro sistema: (1) A través de un interfaz web, permitiendo al usuario operar los textos desde su navegador. (2) A través de servicios web, permitiendo la recuperación de resultados de otros sistemas.

3.1 Aplicación al Dominio Biomédico

La búsqueda de información sobre salud usando Internet puede tener múltiples focos de interés, tales como la utilidad para varios perfiles de usuario: población general, personal médico o investigadores. Para dar soporte a los tres tipos de usuarios, los desarrollos emprendidos por el National Institutes of Health (NIH[2]), parte del U.S. Department of Health & Human Services, son dignos de mención. Dos ejemplos de los recursos disponibles por la población son: Healthfinder y MedlinePlus.

En nuestro caso algunos de estos recursos son tomados como referencias, particularmente los siguientes: (1) PubMed, que es usado para obtener casos científicos. Es una de las fuentes más utilizadas para buscar literatura biomédica (1300 millones de búsquedas en 2009). (2) La base de datos MIMIC II (Clifford G. D. et al, 2010) y los historiales clínicos de FMOD han sido seleccionados para probar la lógica lingüística en textos médicos, del mismo modo que han sido usados en otros trabajos relacionados (Luo G. and Tang C., 2008).

Para realizar la valoración previa al sistema de evaluación de la herramienta, hemos utilizado historiales médicos de MIMIC-II junto con casos facilitados por Pathology Department de Pittsburgh University[3] (elegidos al azar los casos: 223, 410, 474, 564, 565 y 616). Para mostrar la aplicación de la herramienta disponemos de uno de estos informes de casos recientes.

Después de introducir el texto en la herramienta y procesarlo, el reconocimiento de entidades medicas y las fuentes se muestra en una tabla que permite organizar y mostrar ambos. Los resultados se pueden ver en la Figura 2. El mismo proceso se puede aplicar al texto completo del caso.

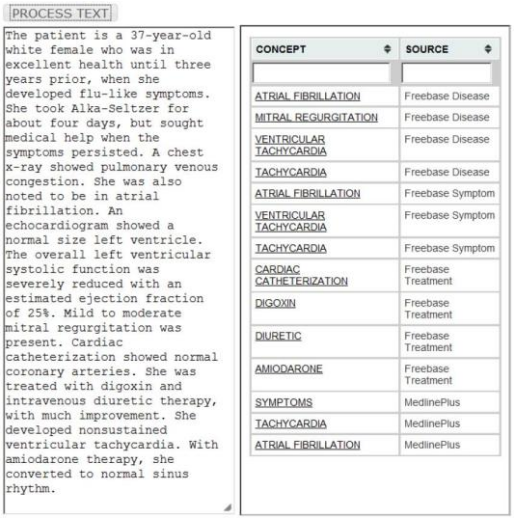


Figura 2: Resultados de la herramienta después del procesado de texto

Los conceptos resultantes están conectados a una página que muestra la información obtenida a través de los servicios ofrecidos por las fuentes. Por

[2] www.nih.gov

[3] path.upmc.edu/cases.html

ejemplo, el concepto "fibrilación auricular" se ha considerado como condición o síntoma en Freebase y también en Medlineplus (lo que indica que una entrada se realiza a través de la ontología NCBO llamada Medlineplus Health Topics).

Profundizando en los tópicos de enfermedades en Freebase, además de la descripción, podemos obtener un conjunto de síntomas, factores de riesgo y tratamientos asociados a una enfermedad en particular. Cada uno de estos términos relacionados semánticamente, están conectados a un tópico de la fuente. Si es necesario profundizar en los síntomas, la información semántica puede obtenerse a través de propiedades como "efectos secundarios" o "síntomas de". En el caso de buscar tratamientos, la información se obtiene a través de sus contraindicaciones, efectos secundarios y pruebas.

Si se selecciona Medlineplus como fuente, además de la descripción, se pueden obtener otros resultados relacionados con la investigación, como *Arrhythmia* o *Blood Thinners* entre otros, así como MeSH[4] (Medical Subject Heading) muestra sinónimos (los términos en MeSH para Arrhythmia son Arrhythmias y Cardiac).

Por último, tenemos la posibilidad de capturar una lista de publicaciones científicas en PubMed[4] utilizando la información obtenida en la búsqueda en Freebase y Medlineplus. En el primero, relacionamos los conceptos de la búsqueda a publicaciones relacionadas, en la segunda, relacionamos con los sinónimos de MeSH.

3.2 Aplicación al Dominio Turístico

El tipo de dominio al que nos enfrentamos tiene ciertas características que lo hacen especialmente interesante como caso de estudio para el sistema automático de extracción de información.

Lo primero, el léxico turístico utiliza una amplia terminología, tomada de varias áreas (geografía, economía, historia del arte, etc.). Es, por tanto, un uso del lenguaje poco especializado, un vocabulario amplio y común aunque hay un núcleo más específico de vocabulario, incluyendo términos técnicos relacionados con organizaciones turísticas, servicios, etc (Muñoz G. R. 2011). El tipo de información específica que se soporta (videos de monumentos o destinos turísticos) se caracteriza principalmente por la presentación de una amplia variedad de información. El sistema procesa texto turístico-cultural extraído de estos videos, identificando términos relevantes contenidos en Freebase. El contexto del turismo está consagrado por la meta de recuperar y extraer información, con particular énfasis en el sistema de recuperación de información personalizada, enmarcado en el consorcio Mavir, donde el plan se ha llamado: "Mejorando el acceso, el análisis y la visibilidad de la información y los contenidos multilingüe y multimedia en red para la Comunidad de Madrid". Mavir (http://www.mavir.net)

Los videos han sido desarrollados por el Instituto Español de Turismo (TURESPAÑA) que es la organización dependiente de la Administración General para la promoción de España como destino turístico. (http://www.tourspain.es).

En el caso del turismo, para ilustrar el uso, podemos seleccionar videos en la Web[5] y ver los resultados del sistema después de procesar el texto englobados en las siguientes categorías:

- Atracción turística: Una atracción turística es un lugar o rasgo característico que se visitaría como turista. Los ejemplos incluyen monumentos, parques, museos y similares. P ej.: Museo de Historia Natural de Beijing, Disneyland Park…
- Alojamiento: Este tipo está pensado para hoteles, bed and breakfast, hostales o cualquier otro lugar donde puedes quedarte al viajar, ej.: Hotel Palace San Francisco.
- Destino de viaje: un destino de viaje es un lugar donde se va en vacaciones, ej.: Paris o Bali.

El procesado del texto junto con los resultados de los conceptos identificándose muestra junto con otras dos maneras de conseguir información: (a) relacionándola directamente con la web de Freebase o (b) recuperando información desde el sistema directamente. (Ver Fig 3).

[5] www.ncbi.nlm.nih.gov/pubmed

[6] http://orion.esi.uem.es:8080/TouristFace/

Figura 3: Resultados del Sistema después de procesar el texto del video turístico

4 *Conclusiones y trabajo futuro*

En este artículo, hemos presentado un sistema que se caracteriza por la incorporación de las entidades anteriormente mencionadas procedentes de Freebase, para así reconocer conceptos médicos y turísticos en sistemas de entrada de videos y textos, de este modo ayudar a los usuarios a aumentar la cantidad de información obtenida. Asimismo se ha mostrado la manera en que se relacionan los términos identificados con Freebase. Además, hemos descrito la arquitectura de la plataforma y el desarrollo sistemático de un método para optimizar la integración de nuevas funcionalidades basadas en GATE. Por último, se ha ejemplificado el manejo del sistema.

En futuros trabajos incluiremos una evaluación sistemática por varios grupos de usuarios así como nuevas estrategias de lógica computacional, la incorporación de otras fuentes y nuevas interfaces de usuario, representación de conceptos mediante webs semánticas y la integración de otros sistemas en la arquitectura común escalable desarrollada.

Además, dentro de la plataforma de integración de información, se incluirán funciones de recomendación y personalización y las herramientas necesarias para el paso de medicina traslacional a medicina personalizada.

En el caso del dominio Turismo, se incluirá un módulo para la extracción automática de textos desde video ya desarrollado en el consorcio Mavir y la incorporación de alguna de las ontologías mencionadas.

Hasta donde alcanza la evaluación, se ha desarrollado un test de comprensión de casos clínicos a estudiantes de segundo año de Medicina, para medir la utilidad de una forma objetiva y subjetiva. El resultado del test será publicado próximamente.

En relación a los recursos, estamos estudiando la posibilidad de integrar nuevos contenidos procedentes de fuentes ya en uso (nuevas listas de Freebase y otras ontologías o recursos NCBO) junto con nuevas relaciones entre los resultados. Finalmente, con respecto a las nuevas interfaces, estamos estudiando el acceso a la herramienta desde dispositivos móviles.

Agradecimientos

Esta investigación ha sido financiada por el Ministerio de Ciencia y Tecnología Español MEDICAL-MINER (TIN-2009-14057-C03-01) y por la Comunidad de Madrid bajo el auspicio de la red de investigación MA2VICMR (S2009/TIC-1542)

Referencias

Allan J., B. Carterette y J. Lewis. 2005. When will information retrieval be “good enough”? SIGIR Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval

Aparicio, F., M. De Buenaga, M. Rubio, y A. Hernando. 2011a. An Intelligent Information Access system assisting a Case Based Learning methodology evaluated in higher education with medical students. *Computers & Education.*

Aparicio, F., R. Muñoz, M. Buenaga, y E. Puertas. 2011b. MDFaces: An intelligent system to recognize significant terms in texts from different domains using Freebase. En *Procesamiento de Lenguaje Natural*, 47, pp. 317-318.

Bollacker, K., C. Evans, P. Paritosh, T. Sturge y J. Taylor. 2008. Freebase: a

collaboratively created graph database for structuring human knowledge. In SIGMOD '08: *Proceedings of the 2008 ACM SIGMOD international conference on Management of data held* in Vancouver, Canada, 1247-1250. ACM.

Brenes, D. J., D. G. Avello y K. P. González. 2009. Survey and evaluation of query intent detection methods. In WSCD '09: Proceedings of the 2009 workshop on Web Search Click Data held in Barcelona, Spain, 1-7. ACM.

Carrero, F., J. M. Gómez., M. de Buenaga, J. Mata, y M. Maña. 2007. Acceso a la información bilingüe utilizando ontologías específicas del dominio biomédico. *Procesamiento de Lenguaje Natural*, Vol. 38, pp. 107-117.

Chang, C., H. Kayed, M. Girgis y R. Shaalan. 2006. A Survey of Web Information Extraction Systems. IEEE Transactions on knowledge and data engineering. Volume: 18, Issue: 10

Clark, D. B., S. Touchman, M. Martinez-Garza, F. Ramirez-Marin, y T. Skjerping Drews. 2012. Bilingual language supports in online science inquiry environments. *Computers & Education*, 58(4), pp. 1207-1224.

Clifford, G., D. J. Scott y M. Villarroel. 2010. User Guide and Documentation for the MIMIC II Database, Rev: 259. Cambridge, MA, USA.

Cunningham H. et al, 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, Philadelphia.

De Melo, G., y G. Weikum. 2010. MENTA: inducing multilingual taxonomies from wikipedia. En *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pp. 1099–1108, New York (USA)

Egozi, O., Markovitch S. y Gabrilovich E. 2011. Concept-Based Information Retrieval Using Explicit Semantic Analysis.ACM Transactions on Information systems. Vol. 29 Issue 2

Gutiérrez, Y., A. Fernández, A. Montoyo y S. Vázquez. 2010. Integración de recursos semánticos basados en WordNet. Sociedad Española para el Procesamiento del Lenguaje Natural. Revista 45.

Hamburg, M. A. y Collins F. S. The Path to Personalized Medicine. New England Journal Med 2010; 363:301-304

Han, X. y J. Zhao. 2009. CASIANED: Web Personal Name Disambiguation Based on Professional Categorization. En Proceedings of 2nd Web People Search Evaluation Workshop (WePS2), Madrid, Spain.

Knoth, P., T. Collins, E. Sklavounou, y Z. Zdrahal. 2010. Facilitating cross-language retrieval and machine translation by multilingual domain ontologies. En *Workshop on Supporting eLearning with Language Resources and Semantic Data* (at LREC 2010), Valletta (Malta).

Lew, M. S., N. Sebe, C. Djeraba y R. Jain Content-based multimedia information retrieval: state of the art and challenges. 2006. Journal ACM Transactions on Multimedia computing, Communications and Applications. Vol.2 Issue 1

Luo, G. y C. Tang. 2008. On iterative intelligent medical search. In SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval held in Singapore, Singapore, 3-10. ACM.

Mayfield, J., D. Lawrie, P. McNamee, y D. W. Oard. 2011. Building a Cross-Language Entity Linking Collection in Twenty-One Languages. En *P. Forner, J. Gonzalo, J. Kekäläinen, M. Lalmas, & M. Rijke (Eds.), Multilingual and Multimodal Information Access Evaluation*, Vol. 6941, pp. 3-13, Berlin (Heidelberg).

Müller, C., y I. Gurevych. 2009. Using Wikipedia and Wiktionary in domain-specific information retrieval. En

Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access, CLEF'08, pp. 219–226, Berlin (Heidelberg)

Muñoz, R., F. Aparicio, M. De Buenaga. 2011. Tourist Face: A contents system base on concepts of freebase for Access to the cultural-tourist information. NLDB.

Nadeau, D. y Sekine S., 2007. A survey of named entity recognition and classification. En *Linguisticae Investigationes*, Vol. 30, pp. 3–26.

Navigil R. y P. Velardi. 2004. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites.

Tuchinda, R., C. Knoblock, A. y P. Szekely. Building Mashups by Demonstration. 2011. *ACM Transactions on the Web* (TWEB)

Voss, A., K. Nakata y M. Juhnke. Concept indexing. 1999. *Proceedings of the internation ACM SIGGROUP conference on Supporting group work.*

Verdejo, F., J. Gonzalo, D. Fernández, A. Peñas y F. López. 2000. ITEM: un motor de búsqueda multilingüe basado en indexación semántica. *Proceedings JBIDI.*

Extracción y Recuperación de Información

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 41-48 recibido 04-05-12 revisado 03-06-12 aceptado 06-06-12

Análisis de técnicas PLN de expansión de consulta aplicadas a la tarea de la recuperación de información geográfica*

Analysis of NLP techniques of query expansion applied to the Geographical Information Retrieval task

José M. Perea-Ortega **Miguel Á. García-Cumbreras**
L. Alfonso Ureña-López **Arturo Montejo-Ráez**

Departamento de Informática, Escuela Politécnica Superior
Universidad de Jaén, E-23071 - Jaén
{jmperea,magc,laurena,amontejo}@ujaen.es

Resumen: En este trabajo, proponemos diferentes técnicas relacionadas con el Procesamiento del Lenguaje Natural (PLN) para reformular las consultas geográficas lanzadas a un sistema GIR. Estás técnicas consistirán en la modificación y/o expansión de las dos partes normalmente reconocidas en una consulta geográfica: la parte temática y la parte geográfica. Hemos evaluado cada una de las reformulaciones propuestas utilizando un marco de experimentación para evaluar sistemas GIR como GeoCLEF. Los resultados obtenidos demuestran que todas las reformulaciones de consulta propuestas recuperaron documentos relevantes que no fueron recuperados utilizando la consulta original, por lo que estas estrategias se pueden considerar de utilidad a la hora de trabajar con sistemas GIR.
Palabras clave: Reformulación de consulta geográfica, recuperación de información geográfica, GeoCLEF

Abstract: In this paper, we propose different Natural Language Processing (NLP) techniques of query reformulation related to the modification and/or expansion of the thematic and geographic parts that are usually identified in a geographic query. We have evaluated each of the reformulations proposed using GeoCLEF as an evaluation framework for GIR systems. The results obtained show that all proposed query reformulations retrieved new relevant documents that were not retrieved using the original query.
Keywords: Geographic query reformulation, Geographical Information Retrieval, GeoCLEF

1. Introducción

En el campo de la Recuperación de Información (*Information Retrieval*, IR) (Baeza-Yates y Ribeiro-Neto, 1999), al enfoque basado en la modificación de la consulta del usuario para mejorar la calidad de los resultados de la búsqueda se le conoce como reformulación de consulta. El objetivo de dicho proceso es satisfacer la necesidad de información de los usuarios, normalmente mejorando la calidad y la cobertura de los resultados obtenidos mediante la consulta original. Esta característica la soportan algunos motores de búsqueda de forma explícita, sugiriendo búsquedas similares a la consulta inicial del usuario. Por otra parte, algunos motores de búsqueda también consiguen reformular la consulta de forma implícita, es decir, mediante la expansión de la consulta original con términos relacionados con sus palabras clave, por ejemplo.

La recuperación de información geográfica (*Geographical Information Retrieval*, GIR) es un área de investigación activa y en crecimiento que se centra en la recuperación de documentos textuales de acuerdo a un criterio geográfico de relevancia. Por esta razón, la GIR se puede considerar una extensión de la IR. En concreto, la GIR se ocupa de mejorar

* Este trabajo ha sido cofinanciado por la Comisión Europea bajo el Séptimo Programa Marco (FP7-2007-2013) a través del proyecto FIRST (FP7-287607), por el Fondo Europeo de Desarrollo Regional (FEDER) con el proyecto TEXT-COOL 2.0 (TIN2009-13391-C04-02) y por el proyecto local Geocaching Urbano (RFC/IEG2010)

la calidad de la recuperación específica de la información geográfica, centrándose en el acceso a documentos no estructurados (Jones y Purves, 2008; Larson, 1996). La comunidad IR ha sido principalmente responsable de la investigación en el campo de la GIR, en lugar de la comunidad relacionada con los Sistemas de Información Geográfica (SIG). El tipo de consulta en un motor IR se basa generalmente en lenguaje natural, en contraste con el enfoque común más formal de los SIG, en los que cada objeto geográfico se recupera de una base de datos estructurada. En un sistema GIR, una consulta geográfica puede ser estructurada como una tripleta <*tema*><*relación espacial*><*localización*>, donde <*tema*> es el foco principal de la consulta, <*localización*> representa el ámbito geográfico de la consulta y <*relación espacial*> determina la relación entre el foco o tema y el ámbito geográfico. Por ejemplo, la tripleta para la consulta geográfica "*airplane crashes close to Russian cities*" sería de <*airplane crashes*> <*close to*> <*Russian cities*>. En definitiva, para una búsqueda como "*castles in Spain*", un buen sistema GIR debería devolver no sólo los documentos que contienen la palabra "*castles*", sino también aquellos que tengan alguna entidad geográfica relacionada con España.

Dado que un sistema GIR puede ser visto o tratado como un motor de búsqueda tradicional (los resultados para una consulta se muestran como una lista de documentos ordenada de mayor a menor relevancia), es importante prestar atención a la búsqueda de métodos eficaces para reformular la consulta de usuario. Estos métodos pueden tener en cuenta tanto características léxico-sintácticas como aspectos geográficos. De esta manera, los resultados de la búsqueda mejorarán su calidad y su cobertura. El objetivo de este trabajo es evaluar y analizar el comportamiento de varias reformulaciones de consulta geográfica propuestas para la tarea GIR, teniendo en cuenta que un sistema GIR puede funcionar como un sistema IR. Para llevar a cabo esta evaluación, se ha utilizado el marco de trabajo más importante en este contexto: GeoCLEF[1] (Gey et al., 2005; Mandl et al., 2008).

El resto del artículo se estructura de la siguiente manera: en la Sección 2 se exponen los trabajos más importantes relacionados con la reformulación de consulta geográfica en general; en la Sección 3 se describe el sistema GIR utilizado para los experimentos; en la Sección 4 se describe brevemente el marco de evaluación; en la Sección 5 se presentan los experimentos realizados y un análisis de los resultados obtenidos; por último, en la Sección 6, se exponen las conclusiones y el trabajo futuro.

2. Trabajo relacionado

Jansen, Booth, y Spink (2009) definen el concepto de reformulación de consulta como el proceso de alteración de la consulta para mejorar el funcionamiento de la búsqueda o recuperación de información. En algunas ocasiones, los motores de búsqueda aplican este método utilizando la técnica conocida como "retroalimentación por relevancia" (*relevance feedback*). Esta técnica se puede aplicar, por un lado, permitiendo que los usuarios decidan si un documento recuperado es relevante o no, generando, a partir de su elección, reformulaciones de la consulta original automáticamente. Por otro lado, esta reformulación también se puede llevar a cabo de manera automática analizando los n primeros documentos recuperados sin la intervención del usuario, teniendo en cuenta estadísticas relacionadas con cada término del documento. No obstante, se ha estudiado que los usuarios pocas veces utilizan la retroalimentación por relevancia (Spink, Jansen, y Ozmultu, 2000) y normalmente reescriben su consulta de forma manual (Anick, 2003).

Este trabajo está relacionado específicamente con consultas de tipo geográfico. Según Gravano, Hatzivassiloglou, y Lichtenstein (2003), los motores de búsqueda actuales son criticados debido a su ignorancia por no considerar restricciones geográficas en las consultas de los usuarios y, por tanto, eso provoca que recuperen menos documentos relevantes. Este problema podría ser atribuido a la manera en la que los motores de búsqueda tradicionales manejan las consultas en general, ya que normalmente adoptan un enfoque basado en la coincidencia con las palabras clave de la consulta (*keywords matching*), sin tener en cuenta el ámbito espacial de los términos geográficos.

Varios autores han realizado estudios sobre las búsquedas realizadas por usuarios cuando utilizan consultas geográficas en mo-

[1] `http://ir.shef.ac.uk/geoclef/`

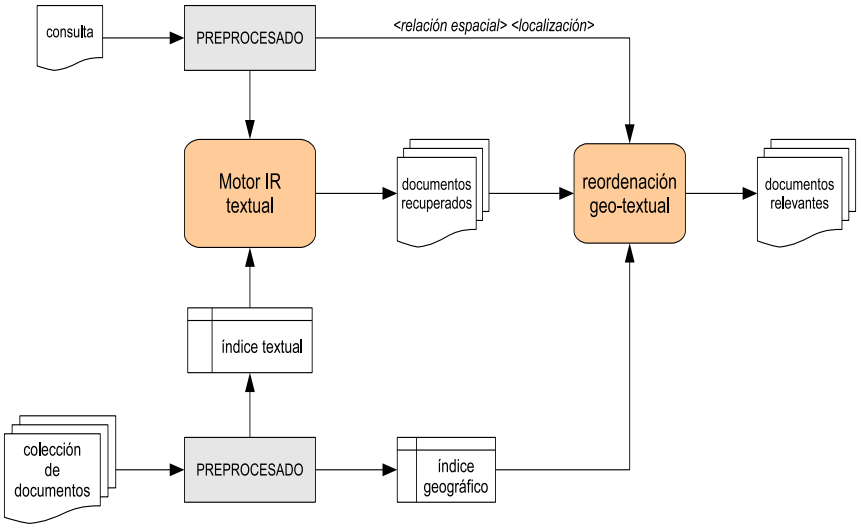


Figura 1: Arquitectura general del sistema SINAI-GIR

tores de búsqueda (Sanderson y Kohler, 2004; Gan et al., 2008; Jones et al., 2008). Una de las principales conclusiones de estos estudios es que la estructura de dichas consultas consistía normalmente en dos partes, una temática y otra geográfica, conteniendo esta última términos espaciales o direccionales. Desde un punto de vista geográfico, Kohler (2003) proporciona un estudio sobre geo-reformulación de consultas. Concluye que la adición de más términos geográficos en la consulta es una técnica utilizada comúnmente para diferenciar lugares que comparten el mismo nombre. A esta técnica se le conoce como expansión de consulta utilizando entidades geográficas.

En la literatura podemos encontrar diferentes trabajos que han abordado la expansión de consulta geográfica. Cardoso y Silva (2007) presentan un enfoque basado en el uso de tipos de característica (isla, ciudad, montaña), reajustando la estrategia de expansión de acuerdo a la semántica de la consulta. Fu, Jones, y Abdelmoty (2005) proponen un método de expansión basado en ontología que soporta la recuperación de documentos que son considerados geográficamente relevantes. Consiguen mejorar los resultados de búsqueda cuando la consulta contiene una relación espacial difusa, mostrando que el método propuesto funciona eficientemente utilizando ontologías realistas en un entorno de búsqueda espacial distribuido. Buscaldi, Rosso, y Sanchis Arnal (2005) utilizan WordNet[2] durante la fase de indexación, añadiendo los sinónimos y los holónimos de las entidades geográficas detectadas en el índice de términos de cada documento, demostrando la efectividad de dicho método. Por último, Stokes et al. (2008) concluyen que se obtiene una ganancia significativa en un sistema GIR cuando todos los conceptos (no sólo los geográficos) son expandidos.

3. *La arquitectura SINAI-GIR*

En esta sección se describe un ejemplo de sistema GIR. Específicamente, hemos utilizado nuestro propio sistema GIR llamado SINAI-GIR (Perea-Ortega et al., 2008b). Al igual que ocurre en los sistemas IR tradicionales, en un sistema GIR podemos diferenciar tres etapas: procesamiento de la colección de documentos y las consultas, indexación y búsqueda textual-geográfica y, finalmente, reordenación de los documentos recuperados utilizando una fórmula de relevancia particular que combina la similitud geográfica y textual entre la consulta y el documento recuperado. El sistema GIR utilizado en este trabajo sigue un enfoque similar, como se puede observar en la Figura 1.

Durante el proceso de funcionamiento de un sistema GIR, cada consulta es preprocesada y analizada, identificando el ámbito geográfico y la relación espacial que pudiera contener. Por otro lado, la colección de

[2] `http://wordnet.princeton.edu/`

documentos también es preprocesada, detectando todas las entidades geográficas y generando un índice geográfico con ellas. Durante esa fase, cada consulta preprocesada (incluyendo sus entidades geográficas) es lanzada contra el motor de búsqueda. Finalmente, los documentos recuperados son filtrados y reordenados, colocando en las últimas posiciones aquellos documentos cuyo ámbito geográfico no coincide con el detectado en la consulta. Por el contrario, aquellos documentos cuya similitud con el ámbito geográfico de la consulta es mayor, son colocados en las primeras posiciones de la lista resultado.

Con respecto al procesamiento de la consulta, éste se basa principalmente en el reconocimiento de las entidades geográficas. Además, también implica la especificación de la tripleta explicada en la Sección 1, que será utilizada más tarde durante el proceso de filtrado y reordenación. Para detectar dicha tripleta, hemos utilizado un etiquetador de la categoría léxica de las palabras (*Part Of Speech*, *POS tagger*) como TreeTagger[3], teniendo en cuenta además algunas reglas sintácticas como *preposición + nombre propio*, por ejemplo. Las palabras vacías fueron eliminadas y se aplicó un extractor de raíces como Snowball[4] a cada palabra, excepto para las entidades geográficas, que mantuvieron su forma original.

Durante el proceso de recuperación, se obtuvieron 1.000 documentos para cada consulta, utilizando Terrier[5] como motor de recuperación. Según un estudio previo realizado por Perea-Ortega et al. (2008a), se demostró que Terrier es una de las herramientas más utilizadas para aplicaciones IR en general y en sistemas GIR en particular, obteniendo resultados prometedores. El esquema de pesado utilizado ha sido *inL2*, el cúal está implementado por defecto en Terrier. Este esquema aplica el modelo de la frecuencia inversa de documento (*Inverse Document Frequency*, IDF) para la aleatoriedad, la sucesión de Laplace para la normalización en primer lugar y normalización 2 para la normalización de la frecuencia de cada término (Amati, 2003). Por último, cabe señalar que, aunque los sistemas GIR normalmente aplican un proceso de geo-reordenación después del módulo IR, en este trabajo no es necesario utilizarlo porque estamos interesados en evaluar la precisión y cobertura de cada reformulación de consulta propuesta desde el punto de vista de la recuperación de información.

4. GeoCLEF como marco de evaluación

Para evaluar las reformulaciones propuestas hemos utilizado el marco de experimentación GeoCLEF (Gey et al., 2005; Mandl et al., 2008), un foro de evaluación para sistemas GIR celebrado entre los años 2005 y 2008 bajo el marco de las conferencias CLEF[6]. GeoCLEF proporciona una colección de 169.477 documentos que consisten en noticias extraídas del periódico británico *Glasgow Herald* (1995) y del periódico américano *Los Angeles Times* (1994), representando una amplia variedad de regiones geográficas y lugares. Por otro lado, se proporcionaron un total de 100 consultas textuales o *topics* (25 consultas por año). Las consultas están compuestas por tres campos principales: título (T), descripción (D) y narrativa (N). Para los experimentos llevados a cabo en este trabajo, sólo hemos tenido en cuenta el campo título, ya que representa de forma similar la manera en la que un usuario lanzaría una consulta geográfica a un motor de búsqueda. Algunos ejemplos de consultas GeoCLEF son: "*vegetable exporters of Europe*", "*forest fires in north of Portugal*", "*airplane crashes close to Russian cities*" or "*natural disasters in the Western USA*".

Con respecto a las medidas de evaluación utilizadas, los resultados han sido evaluados haciendo uso de los juicios de relevancia proporcionados por los organizadores de GeoCLEF y del método de evaluación TREC[7]. La evaluación se ha realizado utilizando las medidas típicas de evaluación en recuperación de información: precisión media (*Mean Average Precision*, MAP), cobertura (*Recall*, R) y precisión en n (P@n).

5. Experimentos y resultados

Tal y como se ha comentado previamente, en este trabajo se analizan varias reformulacio-

[3]TreeTagger v.3.2 para Linux. Disponible en `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`

[4]Disponible en `http://snowball.tartarus.org`

[5]Versión 2.2.1, disponible en `http://terrier.org`

[6]Cross Language Evaluation Forum `http://www.clef-initiative.eu/`

[7]*trec_eval* es un programa para evaluar resultados TREC utilizando los procedimientos estándar del NIST `http://trec.nist.gov/trec_eval/`

Reformulación	Texto de la consulta
original	`visit American presid Germany`
QR1	`visit American presid`
QR2	`visit American presid visit American presid Germany`
QR3	`#and(#or(visit meet stay ) American presid Germany)`
QR4	`#and(visit American presid #or(Germany #3(Federal Republic of Germany) Deutschland FRG ) )`
QR5	`#and(visit American presid #or(Germany Berlin Hamburg Muenchen Koeln #2(Frankfurt am Main) Essen ) )`
QR6	`#and(#or(visit meet stay ) of the American presid ) #or(Germany Berlin Hamburg Muenchen Koeln #2(Frankfurt am Main) Essen )`

Tabla 1: Ejemplo de reformulaciones generadas para la consulta "*Visits of the American president to Germany*"

nes de consultas para la tarea GIR. Para ello se han utilizado las dos partes principales de una consulta geográfica: la parte temática y el ámbito geográfico detectado. El objetivo de estas reformulaciones es mejorar el proceso de recuperación, tratando de encontrar documentos relevantes que no han sido recuperados utilizando la consulta original. A partir de la consulta original preprocesada, hemos generado los siguientes tipos de reformulación:

- QR1: el ámbito geográfico es eliminado, dejando únicamente la parte temática de la consulta original.
- QR2: la parte temática es expandida repitiendo sus palabras clave. De esta forma, tratamos de dar más importancia a la parte temática con respecto a la parte geográfica.
- QR3: la parte temática es expandida utilizando únicamente los sinónimos de las palabras clave detectadas en dicha parte. Hemos considerado como palabras clave los sustantivos, por ser aquellos que más carga semántica poseen. Se ha utilizado WordNet como recurso léxico para la obtención de los sinónimos, consultando todos los sentidos del sustantivo.
- QR4: la parte geográfica es expandida utilizando únicamente sinónimos del ámbito geográfico detectado en la consulta. Se ha utilizado GeoNames[8] como base de conocimiento geográfico.
- QR5: la parte geográfica es expandida utilizando localizaciones o lugares que coinciden con el ámbito geográfico y la relación espacial detectados en la consulta.
- QR6: tanto la parte temática como la geográfica son expandidas, combinando las reformulaciones QR3 y QR5.

La Tabla 1 muestra un ejemplo de las diferentes reformulaciones generadas para la consulta "*Visits of the American president to Germany*". Como se puede observar, en las reformulaciones QR2 y QR3 se expande únicamente la parte temática, mientras que en las reformulaciones QR4 y QR5 se expande solamente la parte geográfica. Finalmente, la reformulación QR6 puede ser considerada una combinación de expansiones utilizando ambas partes.

Los diferentes resultados obtenidos utilizando cada reformulación de consulta (*Query Reformulation*, QR), junto con el obtenido usando la consulta original, se pueden observar en la Tabla 2. En dicha tabla, se muestra la precisión media en los 5, 10 y 100 primeros documentos recuperados, la cobertura y el valor MAP para cada reformulación. Aunque ninguna de ellas consigue mejorar el MAP obtenido con la consulta original, es interesante señalar que la QR2 (la parte temática es expandida repitiendo sus palabras clave) alcanza el mejor resultado P@10 en tres de los cuatro grupos de consultas.

Llegados a este punto, nos preguntamos si las reformulaciones propuestas estaban realmente recuperando documentos relevantes que la consulta original no era capaz de recuperar. Para resolver esta duda, utilizamos los juicios de relevancia proporcionados por los

[8] `http://www.geonames.org`

Grupo de consultas	QR	P@5	P@10	P@100	R	MAP
2005	original	**0,5520**	0,4560	**0,1904**	**0,8364**	**0,3514**
	QR1	0,2640	0,2560	0,1260	0,6748	0,1638
	QR2	0,5200	**0,4920**	0,1840	0,8276	0,3353
	QR3	0,3680	0,3160	0,1400	0,7596	0,2035
	QR4	0,3120	0,2800	0,1212	0,6552	0,2242
	QR5	0,1440	0,1240	0,0772	0,5624	0,0952
	QR6	0,1600	0,1480	0,0780	0,5692	0,0942
2006	original	**0,2400**	0,1920	**0,0716**	**0,7288**	**0,2396**
	QR1	0,0560	0,0640	0,0252	0,4604	0,0615
	QR2	0,2320	**0,2040**	0,0664	0,6796	0,2314
	QR3	0,1440	0,1400	0,0604	0,7356	0,1419
	QR4	0,1920	0,1720	0,0636	0,6984	0,2064
	QR5	0,2240	0,1840	0,0612	0,6524	0,1811
	QR6	0,1840	0,1760	0,0580	0,6772	0,1486
2007	original	**0,3040**	**0,2560**	**0,1188**	**0,7156**	**0,2311**
	QR1	0,1600	0,1320	0,0796	0,4452	0,1255
	QR2	0,2640	0,2120	0,1072	0,6656	0,1871
	QR3	0,2000	0,1800	0,0884	0,6284	0,1774
	QR4	0,2160	0,2000	0,1020	0,6608	0,1687
	QR5	0,2240	0,2000	0,0928	0,6720	0,1874
	QR6	0,2240	0,2040	0,0836	0,6344	0,1763
2008	original	**0,3760**	**0,2680**	0,1104	**0,7368**	**0,2484**
	QR1	0,1760	0,1400	0,0928	0,5996	0,1301
	QR2	0,3440	**0,2680**	**0,1124**	0,7196	0,2381
	QR3	0,2960	0,2320	0,1024	0,6884	0,1972
	QR4	0,2640	0,1960	0,0924	0,6404	0,1619
	QR5	0,2720	0,2040	0,0964	0,6984	0,1906
	QR6	0,2720	0,2280	0,0948	0,7028	0,2028

Tabla 2: Resultados obtenidos para cada reformulación de consulta propuesta

Grupo de consultas	Nº total docs relev.	ORIG	QR1	QR2	QR3	QR4	QR5	QR6
2005	1028	88,33 %	2,33 %	2,04 %	1,85 %	0,68 %	1,36 %	1,85 %
2006	378	75,13 %	2,65 %	1,32 %	4,76 %	1,59 %	3,97 %	5,56 %
2007	650	83,54 %	1,08 %	1,08 %	2,77 %	5,38 %	4,62 %	5,23 %
2008	747	78,71 %	4,82 %	4,28 %	2,95 %	0,80 %	9,10 %	8,97 %
Media		81,43 %	2,72 %	2,18 %	3,08 %	2,11 %	4,76 %	5,40 %

Tabla 3: Porcentaje de documentos relevantes recuperados por cada reformulación propuesta comparados con la consulta original

organizadores de GeoCLEF y obtuvimos los resultados que se muestran en la Figura 2. El número total de documentos recuperados fue siempre 1.000. Por otro lado, según los juicios de relevancia, el número total de documentos relevantes para cada conjunto de consultas (2005, 2006, 2007 y 2008) fue 1.028, 378, 650 y 747, respectivamente. Además, el número de documentos relevantes recuperados por la consulta original fue 908, 284, 543 y 588 para el conjunto de topics 2005, 2006, 2007 y 2008, respectivamente. La Tabla 3 muestra el porcentaje de documentos relevantes recuperados por cada reformulación propuesta comparados con la consulta original.

Analizando estos resultados, podemos observar que todas las reformulaciones propuestas siempre recuperan documentos relevantes que no fueron recuperados por la consulta original. Cabe destacar el buen comportamiento en general de las reformulaciones basadas en la expansión de la parte geográfica (QR4 y

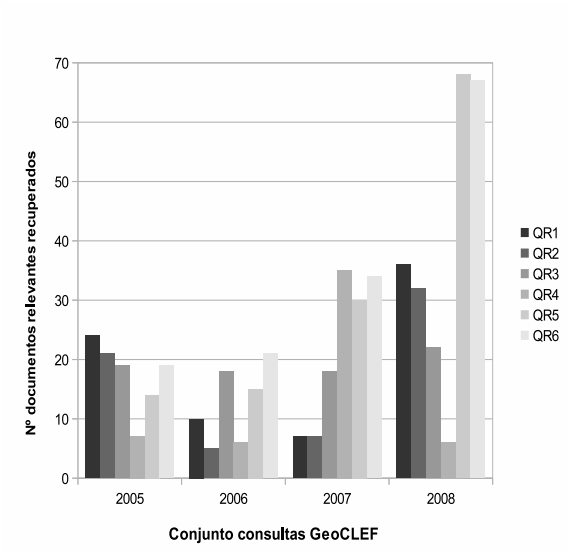


Figura 2: Comparativa del número de documentos relevantes recuperados para cada reformulación que no fueron recuperados por la consulta original

QR5). Específicamente, la QR5 alcanza una importante diferencia utilizando los topics de 2008, con un total de 68 documentos relevantes que no fueron recuperados por la consulta original, es decir, un 9,10 % del total de documentos relevantes para esos topics (ver Tabla 3). Esto significa que de los 159 (747-588) documentos relevantes no recuperados por la consulta original para esos topics, el 42,77 % de ellos fueron recuperados utilizando dicha reformulación. Otro ejemplo parecido ocurre con la QR4, que obtiene el valor más alto para los topics de 2007 consiguiendo el 5,38 % de los documentos relevantes para esas consultas. Este dato representa el 32,71 % de los documentos relevantes no recuperados por la consulta original.

Con respecto a las reformulaciones relacionadas con la expansión de la parte temática (QR2 y QR3) éstas también obtuvieron buenos resultados en general, aunque especialmente para los topics de 2005 y 2006. La QR3 consiguió un 4,76 % de los documentos relevantes para los topics de 2006. Todo esto hace que la reformulación QR6 que combina las QR3 y QR5 obtenga muy buenos resultados como se muestra para todos los conjuntos de consultas en general. Finalmente, comparando cada reformulación con el resto en cada conjunto de consultas, la reformulación QR1 obtiene el mejor valor para las de 2005, por lo que la idea de eliminar la parte geográfica en la consulta original puede ser una buena estrategia cuando la consulta se considera no geográfica.

6. Conclusiones y trabajo futuro

En este trabajo se proponen diferentes técnicas PLN de reformulación de consulta basadas en la modificación y/o expansión de las dos partes de una consulta geográfica: la parte temática y el ámbito geográfico. Se han evaluado cada una de las reformulaciones propuestas utilizando GeoCLEF como marco de evaluación para sistemas GIR. Esta evaluación se ha llevado a cabo desde el punto de vista de la recuperación de información, es decir, sin tener en cuenta nigún tipo de proceso de geo-reordenación posterior al proceso de recuperación de documentos relevantes. Los resultados obtenidos demuestran que, si bien el rendimiento no mejora respecto del caso base, todas las reformulaciones propuestas recuperaron documentos relevantes que no fueron recuperados mediante la consulta original. Esto nos lleva a pensar que, en determinados casos, es necesario realizar estas reformulaciones, si bien no lo es para todas las consultas.

Por tanto, como trabajo futuro, analizaremos los diferentes tipos de consultas geográficas para así estudiar con más profundidad en qué casos es aconsejable aplicar estas técnicas en un sistema GIR dependiendo del tipo de consulta. De este modo, aplicando una expansión selectiva, trataremos de mejorar el rendimiento general del sistema cuando se utiliza únicamente la consulta original.

Bibliografía

Amati, G. 2003. *Probabilistic Models for Information Retrieval based on Divergence from Randomness*. Ph.D. tesis, School of Computing Science, University of Glasgow.

Anick, Peter. 2003. Using terminological feedback for web search refinement: a log-based study. En *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, páginas 88–95, New York, NY, USA. ACM.

Baeza-Yates, Ricardo A. y Berthier Ribeiro-Neto. 1999. *Modern Information Retrie-*

val. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Buscaldi, Davide, Paolo Rosso, y Emilio Sanchis Arnal. 2005. Using the wordnet ontology in the geoclef geographical information retrieval task. En *CLEF*, volumen 4022 de *Lecture Notes in Computer Science*, páginas 939–946. Springer.

Cardoso, Nuno y Mário J. Silva. 2007. Query expansion through geographical feature types. En Ross Purves y Chris Jones, editores, *GIR*, páginas 55–60. ACM.

Fu, Gaihua, Christopher B. Jones, y Alia I. Abdelmoty. 2005. Ontology-based spatial query expansion in information retrieval. En *OTM Conferences (2)*, volumen 3761 de *Lecture Notes in Computer Science*, páginas 1466–1482. Springer.

Gan, Qingqing, Josh Attenberg, Alexander Markowetz, y Torsten Suel. 2008. Analysis of geographic queries in a search engine log. En *Proceedings of the first international workshop on Location and the web*, páginas 49–56, Beijing, China. ACM.

Gey, Fredric C., Ray R. Larson, Mark Sanderson, Hideo Joho, Paul Clough, y Vivien Petras. 2005. GeoCLEF: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. En *CLEF*, volumen 4022 de *Lecture Notes in Computer Science*, páginas 908–919. Springer.

Gravano, L., V. Hatzivassiloglou, y R. Lichtenstein. 2003. Categorizing web queries according to geographical locality. En *Proceedings of the 12th International Conference on Information and Knowledge Management*, páginas 325–333.

Jansen, Bernard J., Danielle L. Booth, y Amanda Spink. 2009. Patterns of query reformulation during web searching. *JASIST*, 60(7):1358–1371.

Jones, Christopher B. y Ross S. Purves. 2008. Geographical information retrieval. *International Journal of Geographical Information Science*, 22(3):219–228.

Jones, Rosie, Wei Vivian Zhang, Benjamin Rey, Pradhuman Jhala, y Eugene Stipp. 2008. Geographic intention and modification in web search. *International Journal of Geographical Information Science*, 22(3):229–246.

Kohler, J. 2003. Analysing search engine queries for the use of geographic terms. Master's thesis, University of Sheffield - United King.

Larson, R. 1996. Geographic information retrieval and spatial browsing. En Smith y M. Gluck, editores, *Geographic Information Systems and Libraries: Patronsand Mapsand and Spatial Information*, páginas 81–124.

Mandl, Thomas, Paula Carvalho, Giorgio Maria Di Nunzio, Fredric C. Gey, Ray R. Larson, Diana Santos, y Christa Womser-Hacker. 2008. GeoCLEF 2008: The CLEF 2008 Cross-Language Geographic Information Retrieval Track Overview. En *CLEF*, volumen 5706 de *Lecture Notes in Computer Science*, páginas 808–821. Springer.

Perea-Ortega, José M., Miguel A. García-Cumbreras, Manuel García-Vega, y L. Alfonso Ureña-López. 2008a. Comparing several textual information retrieval systems for the geographical information retrieval task. En *NLDB*, volumen 5039 de *Lecture Notes in Computer Science*, páginas 142–147. Springer.

Perea-Ortega, José M., Luis Alfonso Ureña-López, Manuel García-Vega, y Miguel Angel García-Cumbreras. 2008b. Using query reformulation and keywords in the geographic information retrieval task. En *CLEF*, volumen 5706 de *Lecture Notes in Computer Science*, páginas 855–862. Springer.

Sanderson, M. y J. Kohler. 2004. Analyzing geographic queries. En *Proceedings Workshop on Geographical Information Retrieval SIGIR*.

Spink, Amanda, Bernard J. Jansen, y Cenk H. Ozmultu. 2000. Use of query reformulation and relevance feedback by excite users. *Internet Research: Electronic Networking Applications and Policy*, 10(4):317–328.

Stokes, Nicola, Yi Li, Alistair Moffat, y Jiawen Rong. 2008. An empirical study of the effects of nlp components on geographic ir performance. *International Journal of Geographical Information Science*, 22(3):247–264.

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 49-56 recibido 04-05-12 revisado 04-06-12 aceptado 07-06-12

A clustering-based Approach for Unsupervised Word Sense Disambiguation

Una aproximación no supervisada para la desambiguación del sentido de las palabras basada en agrupamiento

Tamara Martín-Wanton
Department of Languages and Computer Systems, Universidad de Educación a Distancia (UNED)
C/ Juan del Rosal n 16. 28040 - Madrid
tmartin@lsi.uned.es

Rafael Berlanga-Llavori
Department of Languages and Computer Systems, Universidad Jaume I,
Ave. Vicent Sos Baynat, Castellón 12071
Spain
berlanga@lsi.uji.es

Resumen: Los métodos de agrupamiento han sido ampliamente usados en muchas tareas de Procesamiento de la Información con el fin de capturar categorías de objetos desconocidos. Sin embargo, el agrupamiento ha sido poco utilizado como método para etiquetar sentidos en la Desambiguación del Sentido de las Palabras (WSD), es decir, como una forma de identificar grupos formados por sentidos de palabras semánticamente relacionados que pueden ser utilizados con éxito en el proceso de desambiguación. En este artículo presentamos un método de desambiguación no supervisado basado en el agrupamiento de sentidos de palabras que además es capaz de encontrar relaciones implícitas (no presentes en WordNet) entre los sentidos de las palabras de la oración. Investigamos en profundidad el rol del agrupamiento y su contribución al WSD. En los resultados experimentales se demuestra la utilidad del agrupamiento para la desambiguación no supervisada.
Palabras clave: Desambiguación del Sentido de las Palabras, Agrupamiento

Abstract: Clustering methods have been extensively used in many Information Processing tasks in order to capture unknown object categories. However, clustering has been scarcely used as a sense labeling method for Word Sense Disambiguation (WSD), that is, as a way to identify groups of semantically related word senses that can be successfully used in a disambiguation process. In this paper, we present an unsupervised disambiguation method relying on word sense clustering that also reveals the implicit relationships (not asserted in WordNet) existing among these word senses.We also investigate in depth the role of clustering and its contribution to WSD. Experimental results demonstrate the usefulness of clustering for unsupervised WSD.
Keywords: Word Sense Disambiguation, Clustering

1. Introduction

The task of Word Sense Disambiguation (WSD) consists of assigning the appropriate meaning (sense) for a particular contextual occurrence of a polysemous word. This task is an essential research area in Natural Language Processing that contributes to almost all semantic-based text processing applications (e.g., Machine Translation, Information Extraction, Question & Answering, etc.).

Navigli (2009) broadly divides WSD approaches into supervised and unsupervised WSD. The former ones require learning a model from hand-tagged samples to disambiguate words, which give them a domain specific character. The latter ones are based on unlabeled corpora, avoiding thus the use of training samples. WSD approaches are further classified into knowledge-based and corpus-based methods. Knowledge-based methods exploit word relationships provided by external lexical resources (e.g., dictionaries, ontologies, etc.), whereas corpus-based methods do not make use of any of these resources. Currently, lexical resources like

ISSN 1135-5948

WordNet (Miller, 1995) constitute the referred source in most general purpose approaches. In this paper, we focus on unsupervised and knowledge-based methods.

The main contribution of this paper is twofold. Firstly, we extend the knowledge-based framework proposed in (Anaya-Sánchez, Pons-Porrata, & Berlanga-Llavori, 2006) for the disambiguation of nouns and present an unsupervised all-words disambiguation method derived from it. Our proposal relies on word sense clustering as a natural way to capture the reflected cohesion among the words of a textual unit. This approach is also able to reveals the implicit relationships (not asserted in WordNet) existing among these word senses.

Secondly, we explore in depth the role of clustering and its contribution to WSD. Specifically, we evaluate the capability of clustering for identifying groups of semantically related senses that can help the selection of the right ones and compare our approach with the clustering scheme of senses induced by WordNet domains (Magnini & Cavaglià, 2000). Our experimental results demonstrate the usefulness of word sense clustering for unsupervised WSD.

2. Related work

Most of the knowledge-based methods can be broadly divided into two categories, namely, similarity- and graph-based ones (Navigli & Lapata, 2010). The first category compares each sense of a target word with its surrounding context words. The sense that has the highest similarity is assumed to be the right one. In these approaches, right senses are determined for each word individually without considering the senses previously assigned.

In graph-based methods (Mihalcea, 2005; Navigli & Lapata, 2010), a graph whose nodes are word senses and edges represent meaningful relations or dependencies between them, is built from lexical resources. This graph structure is assessed to determine the importance of each node and the right sense corresponds to the most important node for each word. Experimental studies (Mihalcea, 2005; Brody, Navigli, & Lapata, 2006) show that graph-based methods outperform similarity-based ones. Like Mihalcea's method, we build a weighted graph whose nodes are word senses and edges are labeled with the similarity between them, but instead of determining the importance of a sense by using centrality algorithms, we iteratively perform a clustering method to discover the relationships existing among senses to identify the right ones.

Clustering has been explicitly used in the WSD area for clustering textual contexts of words to induce word senses by dividing the word occurrences into a number of classes or senses (Pedersen, 2006), and also for clustering of fine-grained word senses into coarse-grained ones for reducing the polysemy degree of words (Agirre & López, 2003; Navigli, 2006). However, clustering has been scarcely used as a sense labeling method in the disambiguation task. Hence, our approach shows a novel way of using clustering in this field.

To the best of our knowledge, the major effort in providing groups of semantically related word senses for disambiguation purposes consists of the definition of WordNet domains and several disambiguation algorithms use this domain categories to improve the disambiguation results (Magnini et al., 2002; Kolte & Bhirud, 2008). However, as we demonstrate in the experiments section, the granularity level of these groups is too coarse to be useful for relating word senses.

Most of the WSD methods are restricted to determine the right sense of words in a given context, but none of them give additional information about the possible relationships among the disambiguated word senses. In our proposal, we attempt to reveal the implicit relationships (not asserted in WordNet) existing among word senses.

3. A knowledge-based framework for WSD

In this paper, we extend the framework firstly introduced in (Anaya-Sánchez et al., 2006) to all-words disambiguation. The underlying idea of the framework is to use clustering as a way of identifying semantically related word senses. The goal of the framework is the disambiguation of a finite set of words W given a textual context T. Here, we do not restrict the elements of W to be in T. Our framework comprises the following elements: (1) a representation for senses, which is provided by the knowledge source; (2) a similarity measure to compare sense representations; (3) a

clustering algorithm able to group the sense representations of all words in W; (4) a filtering function for selecting sense clusters that match the best with the context T, and (5) a stopping criterion for ensuring the termination of the disambiguation process.

Assuming that these elements are given, the disambiguation process of the framework starts from a clustering distribution of all possible senses of the words in *W*. Such a clustering tries to identify cohesive groups of word senses, which are assumed to represent different meanings for the set of words. Then, clusters that match the best with the context are selected via a filtering process. If the selected clusters disambiguate all words (i.e., they contain exactly one sense for each word in *W*), the process stops and the senses belonging to the selected clusters are interpreted as the right ones. Otherwise, the clustering and filtering steps are performed again (regarding the remaining senses) until the stopping condition is satisfied. It is worth mentioning that in each iteration the clustering parameters must be refined to obtain stronger cohesive clusters. Notice that in this framework word senses are globally determined by capturing relationships among senses via the clustering process. Figure 1 shows the general steps of the framework for the disambiguation of a set of words. See (Anaya-Sánchez et al., 2006) for details.

Input: The finite set of words *W* and the textual context *T*
Output: The disambiguated word senses
1. Let *Senses* be the set of all senses of words in *W*
2. Repeat
   a. G = *clustering*(*Senses*)
   b. *Selected_G* = *filter*(*G*, *W*, *T*)
   c. $Senses = \bigcup_{g \in Selected\_G} \{s \mid s \in g\}$
   until *stopping-criterion*
3. Return *Senses*

Figure 1: Framework for the disambiguation of the set of words *W* in the textual context *T*

4. Star-based Disambiguation Algorithm

In this section, we introduce our disambiguation approach, which derives from the framework explained above. Our algorithm proceeds incrementally on a sentence-by-sentence basis. We assume that sentences are part-of-speech tagged and, therefore only content words (i.e., nouns, verbs, adjectives, and adverbs) are considered. Thus, in our case the context *T* is represented as a vector of content words in a sentence, all weighted one, and *W* is the set of all words in *T*.

For example, let us consider again the sentence "*The runner won the marathon*". In this example, the set of words *W* includes the nouns *runner* and *marathon*, and the verb *win* (lemma of the verbal form *won*), and the context is the vector *T*=< *runner*:1, *win*:1, *marathon*:1>. The rest of words are not considered for they are meaningless (i.e., stopwords). Hereafter, we will use this sentence example to explain our approach.

4.1. Sense representation

For clustering purposes, word senses are represented as feature vectors. Thus, for each word sense *s* we define a vector $<t_1{:}\sigma_1, \ldots, t_m{:}\sigma_m>$, where each feature t_i is a WordNet term highly correlated to *s* with an association weight σ_i. The set of terms for a word sense includes all its WordNet hyponyms, its directly related terms (including coordinated terms) and their filtered and lemmatized glosses.

To weight vector terms, the *tf-idf* statistics is used, considering each word as a collection and its senses as the collection documents. It is worth mentioning that the use of *tf-idf* weights allows us to distinguish a sense from the other senses of the same word. In this paper, we use the cosine as similarity measure between sense representations:

$$cos(s_i, s_j) = \frac{\vec{s}_i \cdot \vec{s}_j}{\|\vec{s}_i\| \|\vec{s}_j\|}$$

4.2. Clustering algorithm

Sense clustering is carried out by the extended star clustering algorithm (Gil-García, Badía-Contelles, & Pons-Porrata, 2003), which builds star-shaped and overlapped clusters. This clustering algorithm relies on a greedy cover of the β_0-similarity graph by star-shaped subgraphs (Aslam, Pelekhov, & Rus, 2004). A β_0-similarity graph is the undirected graph, whose vertices are word senses and there is an edge between *sense* s_i and sense s_j if the similarity between them is greater than the minimum similarity threshold β_0. Each cluster (star-shaped subgraph) consists of a single star and its satellites, where the star is the word sense with the highest connectivity within the cluster, and the satellites are those senses

connected with the star (i.e., the star neighbors in the graph).

Notice that in our approach, the disambiguation is performed over all the senses of all words in the sentence at once. The underlying hypothesis is that word sense clustering captures the reflected cohesion among the words of a sentence and each cluster reveals possible relationships existing among these word senses. Thus, the way this clustering algorithm relates word senses resembles the way in which syntactic and discourse relations link textual elements.

4.3. Filtering

For each word *w* the proposed filter selects those clusters having maximum similarity w.r.t. the context *T*. That is, we use the following function:

$$filter(G,W,T) = \bigcup_{w \in W} \arg\max_{\substack{g \in G \\ g \text{ covers } w}} cos(\overline{g},T)$$

In this definition, $cos(\overline{g},T)$ represents the cosine similarity between the centroid of cluster *g* and the context *T*, and *covers* states the relation that links a cluster with those words having senses in it.

Thus, we firstly rank all clusters according to its similarity with the context *T*, and then they are orderly processed to select clusters for covering the words in *W*. A cluster *g* is selected if it contains at least one sense of an uncovered word. Otherwise it is discarded.

4.4. Stopping Criterion

As a result of the filtering process, a set of senses for each word in *W* is obtained (i.e., the union of all the selected clusters). Words in *W* having only one sense are considered disambiguated. If some word still remains ambiguous, we must refine the clustering process to get stronger cohesive clusters of senses. In this case, all the remaining senses must be clustered again but raising the β_0 threshold. Thus, this process must be done iteratively until either all words are disambiguated or β_0 cannot be increased anymore. Initially, β_0 is defined as:

$$\beta_0(1) = pth(p, cos(Senses))$$

and at the i-th iteration (i > 1) it is updated to:

$$\beta_0(i) = \min_{q \in \{5,10,15,\ldots\}} \left\{ \begin{array}{c} \beta = pth(p+q, cos(Senses)) \\ | \ \beta > \beta_0(i-1) \end{array} \right\}$$

In these equations, *Senses* is the set of current senses, and *pth*(*p, cos*(*Senses*)) represents the *p*-th percentile value of pair-wise cosine similarities of all senses in *Senses* (i.e., $cos(Senses)=\{cos(s_i, s_j) \mid s_i,s_j \in Senses, i \neq j\} \cup \{1\}$). Here, *p* is a user-defined parameter.

Figure 5 graphically depicts the disambiguation process of the example sentence carried out by our method. The boxes in the figure represent the obtained clusters, which are sorted regarding their similarities with respect to the context (scores are under the boxes), and doubly-boxed clusters depict the selected ones by the filter.

In our example, we select *p* = 90 for obtaining the initial similarity threshold (β_0=0.048). Notice that the first cluster includes senses that cover the set of all the ambiguous words. Hence, it is selected by the filtering process and all other clusters are discarded. After this step, *Senses* is updated with the senses of the selected cluster.

At this point of the process, *Senses* does not disambiguate completely *W* because the noun *runner* has still two senses. Consequently, a new clustering must be obtained using the current set *Senses* and a new value of β_0.

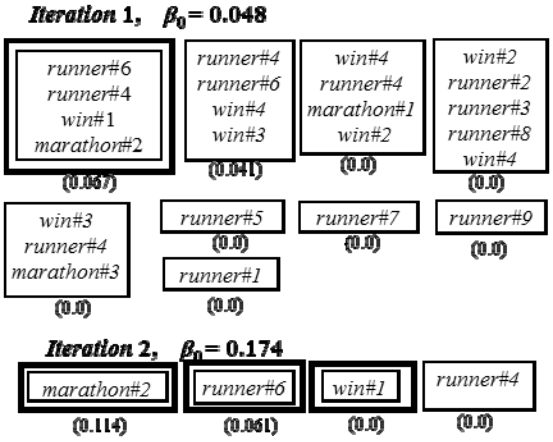


Figure 5: Disambiguation of words in "The runner won the marathon"

As *pth*(90+5, *cos*(*Senses*)) = 0.174 and 0.174 > 0.048, then $\beta_0(2)$ = 0.174. In this case, all clusters become single. Then, the final set of selected senses is *Senses*= {*marathon*#2, *runner*#6, *win*#1}, which includes only one sense for each word in *W*.

5. Experiments

In order to evaluate our proposal, we use the coarse-grained English all-words corpus of SemEval-2007 Task 07 (Navigli, Litkowski, &

Hargraves, 2007). This corpus consists of 5,377 words of running text of which 2,269 have been annotated with senses from a coarse-grained version of the WordNet 2.1 sense inventory.

We follow the evaluation methodology of SemEval-2007 and present the disambiguation results in terms of the traditional F1-measure.

5.1. Does the Extended Star Clustering Algorithm Produce Profitable Clusters for WSD?

The aim of the first experiment is to validate the performance of the extended star algorithm for clustering semantically related senses, i.e., for obtaining useful groups for WSD. With this purpose, for each corpus sentence we compare the relation between the senses generated from the clustering algorithm (namely, the set of pairs (u, v) such that senses u and v belong to a same cluster) w.r.t. the reference model consisting of all pairs of correct word senses.

We use *recall*, *precision* and F1 to evaluate these relations, which can be expressed as follows:

$$recall = \frac{c}{a} \qquad precision = \frac{c}{b} \qquad F1 = \frac{2c}{a+b}$$

where c is the number of pairs of correct senses generated from the clustering, a is the number of all pairs of correct senses in the reference model, and b is the number of sense pairs produced from the clustering that include at least one correct sense. Regarding that our reference model does not include any relation between incorrect senses, we discard all pairs of incorrect senses obtained from the clusters.

Notice that in the above definitions *recall* is a measure of the goodness on grouping together correct senses; whereas *precision* measures the accuracy of the clustering for relating correct word senses with themselves. The F1 measure is the harmonic mean between recall and precision. The higher the value of these measures, the better the clustering is for WSD.

Figure 6 shows the values of recall, precision and F1 achieved over all sentences by varying β_0 threshold. Each β_0 corresponds to a percentile value of the pair-wise similarities of the senses. As it can be appreciated, the extended star clustering algorithm produces stable results up to 70th percentile. The very high recall values obtained in this experiment (around 0.98) demonstrate the usefulness of the sense groups produced by the extended star algorithm.

The relatively low values of precision (around 0.38) were expected because no refinement of the clustering was performed in the experiment (i.e., just one iteration was done).

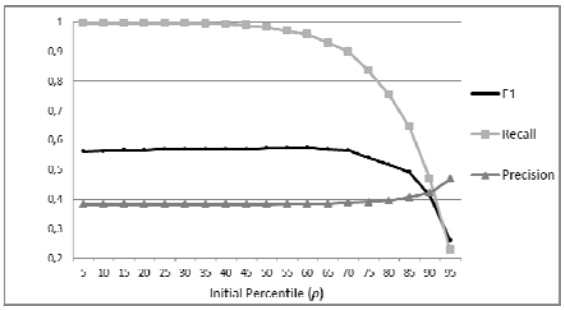


Figure 6: Performance of the extended star clustering in the identification of semantically related groups of senses for WSD

5.2. Sense Clustering and Clustering Refinement for WSD

The goal of the second experiment is to explore the role of both: clustering in the disambiguation process, and the iterative process of clustering refinement in the disambiguation.

Thus, we compare our clustering-based approach with a non-clustering based WSD algorithm obtained as an instance of the disambiguation framework by using the trivial clustering algorithm (i.e., the set of singletons consisting of a word sense) and the same filter and stopping criterion of our proposal. Note that the non-clustering approach selects the senses having maximum similarity w.r.t. the context instead of sense groups. This approach resembles those strategies based on the gloss overlap and relatedness-based measures.

For evaluating the impact of the iterative clustering refinement strategy we consider the results of our approach just after each iteration (i.e., by regarding the remaining senses as the right ones). Figure 7 summarizes the results of this experiment using different p values in the disambiguation process.

The clustering-based method outperforms the non-clustering one for all percentile values. This confirms our hypothesis that clustering provides a way to identify groups of semantically related word senses that can be useful for disambiguation tasks. Nevertheless, it is important to notice the relatively high

impact of the sense representation on the non-clustering baseline.

We can also observe that starting from an initial clustering of senses, the disambiguation results are clearly improved after performing each refinement iteration. This corroborates the idea of using the iterative process of clustering refinement.

Another interesting observation is that the accuracy of our approach is fairly consistent for all percentile values (the F1 scores remain between 0.702 and 0.722). The best F1 score was 0.722 for a 65th percentile. Thus, the greater the percentile value, a lower number of iterations is required to satisfy the stopping criterion.

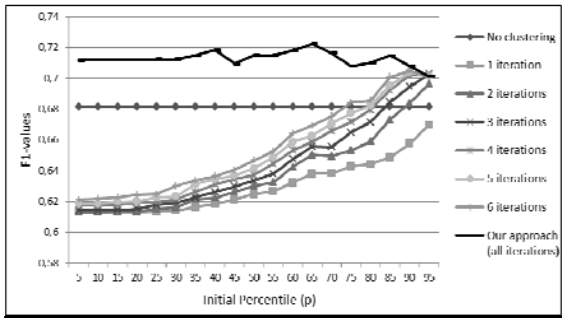


Figure 7: Performance of the star-based WSD algorithm vs. the non-clustering based method

5.3. Extended Star Clustering vs. WordNet Domains

As previously mentioned, WordNet domains have been widely used by several disambiguation algorithms. As WordNet domains induce a clustering distribution for word senses, the purpose of this experiment is to evaluate its performance in the clustering-based WSD framework. With this aim, we replace the clustering component of the star-based WSD algorithm with the clustering induced by WordNet domains.

The induced clustering considers each domain different from *Factotum* as a cluster, that is, all word senses labeled with a domain *d* (*d* ≠ *Factotum*) in a sentence belong to the same cluster. Also, all the senses of a word *w* labeled with *Factotum* domain are considered as belonging to all clusters that do not cover *w*.

In order to define an appropriate clustering refinement strategy, we consider the different levels of the domain hierarchy to generate the word sense clusters. Thus, three iterations are carried out. The first one only considers the global domains of the hierarchy. The second one relies on the basic domains, and finally the domain hierarchy leaves (i.e., the most specific domains) are regarded.

Table 2 shows the results of this experiment. It is shown that, the extended star clustering performs better than the method based on WordNet domains. It is worth mentioning that the results obtained by WordNet domains-based method also support the idea of using the clustering refinement strategy for improving the disambiguation precision.

Method	F1-value
Global domains	0.624
Basic domains	0.631
Domain hierarchy leaves	0.632
Star-based approach	**0.722**

Table 2: Performance of the star- vs. WordNet domains-based WSD algorithms

5.4. Evaluation on standard data sets

In order to contextualize our approach in the current disambiguation state-of-the-art, we evaluate our proposal on several benchmark data sets. Specifically, we use SemCor (Miller et al., 1993), Senseval-3 (Snyder & Palmer, 2004) and SemEval-2007 corpora (see details in http://www.senseval.org/). Our experiments were run in the all-words setting, where the algorithm must disambiguate all (content) words in a given document.

In this paper, we use a subset of SemCor 2.0 composed by all the documents of brown1 and brown2 corpora. It contains a total of 192,639 words (88,058 nouns, 48,328 verbs, 35,664 adjectives and 20,589 adverbs) tagged with WordNet 2.0 senses. In the case of Senseval-3, we use the all-words corpus composed by 2081 words (951 nouns, 751 verbs, 364 adjectives and 15 adverbs) annotated with WordNet 2.0.

In particular, we use the SemEval-2007 Task 7 (Navigli et al., 2007), and Task 17 (Pradhan et al., 2004) data sets. The Task 7 data set consists of 5,377 words of five articles (the first three in common with Task 17) obtained from the WSJ corpus, Wikipedia and Amy Steedman's Knights of the Art. 2,269 of these words are annotated with WordNet 2.1 senses in Task 7 and 455 (159 nouns and 296 verbs) in Task 17.

A fine-grained disambiguation was evaluated in SemCor, Senseval-3 and SemEval 2007 Task 17 corpora, whereas a coarse-

grained evaluation was done in the corpus provided by Task 7 of SemEval-2007. Our results are summarized in Table 3.

In SemEval-2007 competition we participated with the system TKB-UO (Anaya-Sánchez, Pons-Porrata, & Berlanga-Llavori, 2007), which is a previous version of the proposed method in this paper. This system was considered the best unsupervised system (Navigli, 2009). The poor performance of our proposal in the task 17 can be explained by the high polysemy degree of verbs and its relatively small number of relations in WordNet.

Recently, Navigli and Lapata (2010) evaluated several graph-based WSD methods and well- known baselines on these benchmark data sets. From these results, we find that our proposal yields competitive performance with the state-of-the-art unsupervised WSD methods.

Data set	F1-value
SemEval-2007 Task 7 (TKB-UO)	0.702
SemEval-2007 Task 17 (TKB-UO)	0.325
SemEval-2007 Task 7	**0.722**
SemEval-2007 Task 17	**0.332**
Senseval-3 all-words	0.428
SemCor	0.498

Table 3: Our results on standard data sets

6. Conclusions

In this paper, the star-based disambiguation algorithm has been introduced. This unsupervised and knowledge-based method is derived from the framework proposed in (Anaya-Sánchez et al., 2006) by using both feature vectors built from WordNet for representing word senses, and the extended star clustering algorithm. Unlike previous work in WSD, the clustering algorithm is here used to contextually group word senses according to their representations. Also, our approach profits from both sense vectors and clustering method to overcome the sparseness of WordNet relations for associating semantically related word senses.

As a result, our proposal not only is able to disambiguate all words in a sentence but also reveals the implicit relationships (not asserted in WordNet) existing among these word senses. These relations can provide evidence for the sense choices and strong clues that can be helpful for manual annotators (Navigli & Velardi, 2005).

The experiments carried out over the coarse-grained English all-words corpus of SemEval-2007 validate both the use of the extended star clustering for connecting semantically related word senses, and the iterative clustering refinement strategy for WSD. We have also shown that the proposed scheme for grouping word senses outperforms those induced by WordNet domains at any of its abstraction levels. Despite our method requires a percentile value as input parameter, we demonstrate that its accuracy is fairly consistent for almost percentile values.

Our proposal performs the best among all the unsupervised systems participating in the Task 7 of SemEval-2007 competition. It also achieves competitive results with respect to the state-of-the-art unsupervised WSD methods on existing benchmark data sets.

One of the most critical issues for clustering word senses is the representation of senses. As future work, we plan to enrich word sense vectors with other external resources (e.g., Wikipedia), in order to evaluate if they produce better disambiguation results. In particular, a proper sense representation for verbs is a key issue we must face to. Future work also regards to explore the role of the filtering component in the disambiguation and to enrich the information about the textual context. Finally, we will examine the impact of sense co-occurrences that can be obtained from lexical resources, like extended WordNet, in the clustering process.

7. References

Agirre, E., & López, O. (2003). Clustering WordNet Word Senses. In *Proceedings of the Conference on Recent Advances on Natural Language Processing (RANLP '03)*, pp. 121-130. Borovets, Bulgaria.

Anaya-Sánchez, H., Pons-Porrata, A., & Berlanga-Llavori, R. (2006). Word Sense Disambiguation based on Word Sense Clustering. In Simão Sichman, J., Coelho, H., & Oliveira Rezende, S. (Eds.), *Proceedings of 10th Ibero-American Conference on Artificial Intelligence (IBERAMIA-SBIA'2006)*, pp. 472-481. Berlin, Germany: Springer.

Anaya-Sánchez, H., Pons-Porrata, A., & Berlanga-Llavori, R. (2007). TKB-UO:

Using Sense Clustering for WSD. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval'2007)*, pp. 322-325. Morristown, NJ: ACL.

Aslam, J., Pelekhov, E., & Rus, D. (2004). The Star Clustering Algorithm for Static and Dynamic Information Organization. *Journal of Graph Algorithms and Applications, 8*, 95-129.

Brody, S., Navigli, R., & Lapata, M. (2006). Ensemble Methods for Unsupervised WSD. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'2006)*, pp. 97-104. Morristown, NJ: Association for Computational Linguistics.

Gil-García, R., Badía-Contelles, J.M., & Pons-Porrata, A. (2003). Extended Star Clustering Algorithm. In Sanfeliu, A., & Ruiz-Shulcloper, J. (Eds.), *Proceedings of the 8th Iberoamerican Congress on Pattern Recognition (CIARP'2003)*, pp. 480-487. Berlin, Germany: Springer.

Kolte, S.G., & Bhirud, S.G. (2008). Word Sense Disambiguation Using WordNet Domains. In *Proceedings of the First International Conference on Emerging Trends in Engineering and Technology (ICETET'08)*, pp. 1187-1191. Washington, DC: IEEE Computer Society.

Magnini, B., & Cavaglià, G. (2000). Integrating Subject Field Codes into WordNet. In Gavrilidou, M., Crayannis, G., Markantonatu,S.,Piperidis,S.,&Stainhaouer, G. (Eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'2000)*, pp. 1413-1418. Athens, Greece.

Magnini, B., Strapparava, C., Pezzulo, G., & Gliozzo, A. (2002). The role of domain information in Word Sense Disambiguation. *Natural Language Engineering, 8(4)*, 359-373.

Mihalcea, R. (2005). Unsupervised Large-Vocabulary Word Sense Disambiguation with Graph-Based Algorithms for Sequence Data Labeling. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP'2005)*, pp. 411-418. Morristown, NJ: Association for Computational Linguistics.

Miller, G. (1995). WordNet: A Lexical Database for English. *Communications of the ACM, 38(11)*, 39-41.

Miller, G., Leacock, C., Randee, T., & Bunker, R. (1993). A Semantic Concordance. In *Proceedings of the Workshop on Human Language Technology (HLT'93)*, pp. 303-308. Morristown, NJ: Association for Computational Linguistics.

Navigli, R. (2006). Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'2006)*, pp. 105-112. Morristown, NJ: Association for Computational Linguistics.

Navigli, R. (2009). Word Sense Disambiguation: a Survey. *ACM Computing Surveys, 41(2)*, 1-69.

Navigli, R., & Lapata, M. (2010). An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(4)*, 678-692.

Navigli, R., Litkowski, K.-C., & Hargraves, O. (2007). SemEval-2007 Task 07: Coarse-Grained English All-Words Task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval'2007)*, pp. 30-35. Morristown, NJ: Association for Computational Linguistics.

Pedersen, T. (2006). Unsupervised Corpus Based Methods for WSD. In Agirre E., & Edmonds, P. (Eds.), *Word Sense Disambiguation: Algorithms and Applications, chapter 6*. Springer.

Pradhan, S.S., Loper, E., Dligach, D., & Palmer, M. (2007). SemEval-2007 Task 17: English Lexical Sample, SRL and All Words. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval'07)*, pp. 87-92. Morristown, NJ: Association for Computational Linguistics.

Snyder, B., & Palmer, M. (2004). The English All-Words Task. In *Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pp. 41-43. New Brunswick, NJ: Association for Computational Linguistics.

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 57-64 recibido 05-05-12 revisado 31-05-12 aceptado 05-06-12

Representación Gráfica de Documentos para Extracción Automática de Relaciones *

Graph-based Document Representation for Relation Extraction

Bernardo Cabaleiro
NLP&IR Research Group
ETSI Informática, UNED, Spain
bcabaleiro@lsi.uned.es

Anselmo Peñas
NLP&IR Research Group
ETSI Informática, UNED, Spain
anselmo@lsi.uned.es

Resumen: Este artículo presenta un sistema de representación de documentos orientado a la compactación, integración y simplificación de información. El sistema genera grafos a nivel de documento a partir de árboles de dependencias sintácticas haciendo explícita la semántica de algunas aristas. El objetivo es crear una representación útil para múltiples tareas de procesamiento de lenguaje natural, entre ellas la extracción automática de relaciones, para la que realizamos una evaluación extrínseca cuantitativa.
Palabras clave: Representación de documentos, grafos semánticos, extracción de relaciones

Abstract: This paper presents a document representation system oriented to compactation, integration and simplification of information. This system generates document-level graphs from syntactic dependency trees making explicit the semantics of some edges. The goal is to create a representation useful for multiple tasks of natural language processing, as relation extraction. For this task we perform a quantitative evaluation.
Keywords: Document representation, semantic graphs, relation extraction

1. Introducción

Este artículo presenta un sistema de representación de documentos orientado a la compactación, integración y simplificación de información con el fin de avanzar el estado del arte en tareas de procesamiento de lenguaje como resolución de correferencias, extracción de información o búsqueda de respuestas.

La hipótesis de partida es que una representación de documentos completos en forma de grafo, simplificando las relaciones morfosintácticas y añadiendo información semántica, beneficia a sistemas orientados a precisión en tareas cómo la extracción automática de relaciones.

Esta representación a nivel de documento se crea a partir del árbol de dependencias, sobre el que se realizan las siguientes operaciones: Colapsado de correferencias, asignación de clases semánticas a entidades nombradas y normalización de determinadas estructuras sintácticas.

La correferencia es una relación lingüística que se establece entre dos o más expresiones que refieren a una misma entidad, sean o no isomorfas. En este artículo, al conjunto de todas estas expresiones para una entidad lo denominamos *referente de discurso* (Karttunen, 1968). Al proceso de crear un referente de discurso agrupando las menciones a medida que aparecen en un discurso lo denominaremos *colapsado*. Este procedimiento es similar a las tarjetas archivo (*file card*) empleado por Heim (1983) o las cestas (*basket*) en Recasens (2010).

Por otra parte, asignaremos clases semánticas a las entidades nombradas mediante reglas que detecten determinadas estructuras con genitivos, compuestos nominales y aposiciones.

Por último, normalizaremos distintas expresiones que no tengan diferencia semántica. Consideramos la voz gramatical, genitivos expresados de diferentes maneras, compuestos nominales, etc.

Las preguntas de investigación que aborda este trabajo son: (1) ¿Qué efecto tiene la representación gráfica a nivel de documento

* This work has been partially supported by the Spanish Ministry of Science and Innovation, through the project Holopedia (TIN2010-21128-C02).

en una tarea de extracción automática de relaciones? (2) En esta misma tarea, una vez generados los grafos, ¿supone una mejora enriquecerlos con información semántica?

El objetivo que se persigue es proporcionar una base para las técnicas de extracción de información. Si esta base es apropiada, los clasificadores dispondrán de nuevos rasgos que permitirán la captura de contextos distantes en el texto. El impacto de esta técnica se ha evaluado en las tareas Regular Slot Filling y Temporal Slot Filling correspondientes a la tarea competitiva Knowledge Base Population (Ji, Grishman, y Dang, 2011).

El artículo está estructurado de la siguiente manera: Comenzamos detallando la representación en la sección 2, en la sección 3 evaluamos el efecto de la representación propuesta en la tareas, continuamos en la sección 4 comentando brevemente otros trabajos relacionados y terminamos con las conclusiones en la sección 5 y el trabajo futuro en la sección 6.

2. Propuesta de representación

Diferenciamos dos tipos de representaciones, según la etapa de procesamiento en el que se encuentren, para poder evaluar si la agregación de información es útil para los clasificadores:

- Configuración inicial: En esta configuración están disponibles las anotaciones de los nodos y las aristas con información sintáctica, las correferencias y las relaciones temporales.
- Configuración colapsada: En esta configuración se añaden las aristas con información semántica, se normalizan las aristas sintácticas y se realiza el colapsado, por lo que desaparecen las aristas de correferencia y se emplean los referentes de discurso.

2.1. Configuración inicial

En la configuración inicial, cada documento D está representado por un grafo, G_D, con un conjunto de nodos N_D y un conjunto de aristas A_D. Cada nodo representa una unidad de información, generalmente una palabra, excepto en dos casos: una entidad nombrada de más de una palabra, o un verbo y sus auxiliares.

Los nodos están etiquetados con una serie de atributos, algunos de los cuales son comunes para todos los nodos: palabras que contiene, lemas asociados, etiquetas morfosintácticas y una cadena de caracteres representativa que denominamos *descriptor*. Esta cadena se genera de la siguiente manera: Para los nodos que no son entidades, se compone con los lemas de las palabras. En el caso de las entidades, se escoge la cadena de caracteres tal y como se encuentra en el texto.

Además, hay nodos anotados con más propiedades. Se dividen en tres categorías:

- Eventos: Se corresponden con verbos que describen una acción. Están anotados con el tiempo, el aspecto y la polaridad.
- Expresiones temporales: Identifican palabras o conjuntos de palabras que indican un instante concreto o un periodo de tiempo. Contienen un valor temporal normalizado según el estándar TimeX3.
- Entidades nombradas: Entidades reconocidas en el texto. Se anotan con el tipo de entidad, por ejemplo, organización, persona, lugar, etc. En caso de ser persona también se etiqueta el género y la edad.

Las aristas representan cuatro tipo de relaciones entre los nodos:

- Sintáctica: Indica que existe una relación sintáctica entre dos nodos. Se corresponden con las etiquetas sintácticas del Penn Treebank.
- Coreferencia: Indica que dos nodos son menciones de un mismo referente de discurso.
- Semántica: Indica que existe una relación semántica entre dos nodos. Distinguimos cuatro etiquetas semánticas subespecificadas: is, has, hasClass y hasProperty.
- Temporal: Indica que existe una relación temporal entre un evento y una expresión temporal. Las relaciones pueden ser de los tipos: *before*, *after*, *within*, *throughout*, *beginning* y *ending*.

2.2. Configuración colapsada

El grafo con configuración inicial G_D se transforma para crear el grafo con la configuración colapsada G_C. Cada grupo de nodos relacionados por correferencias $n_0, \ldots, n_i, \ldots, n_k \in N_D$ se agrupan en

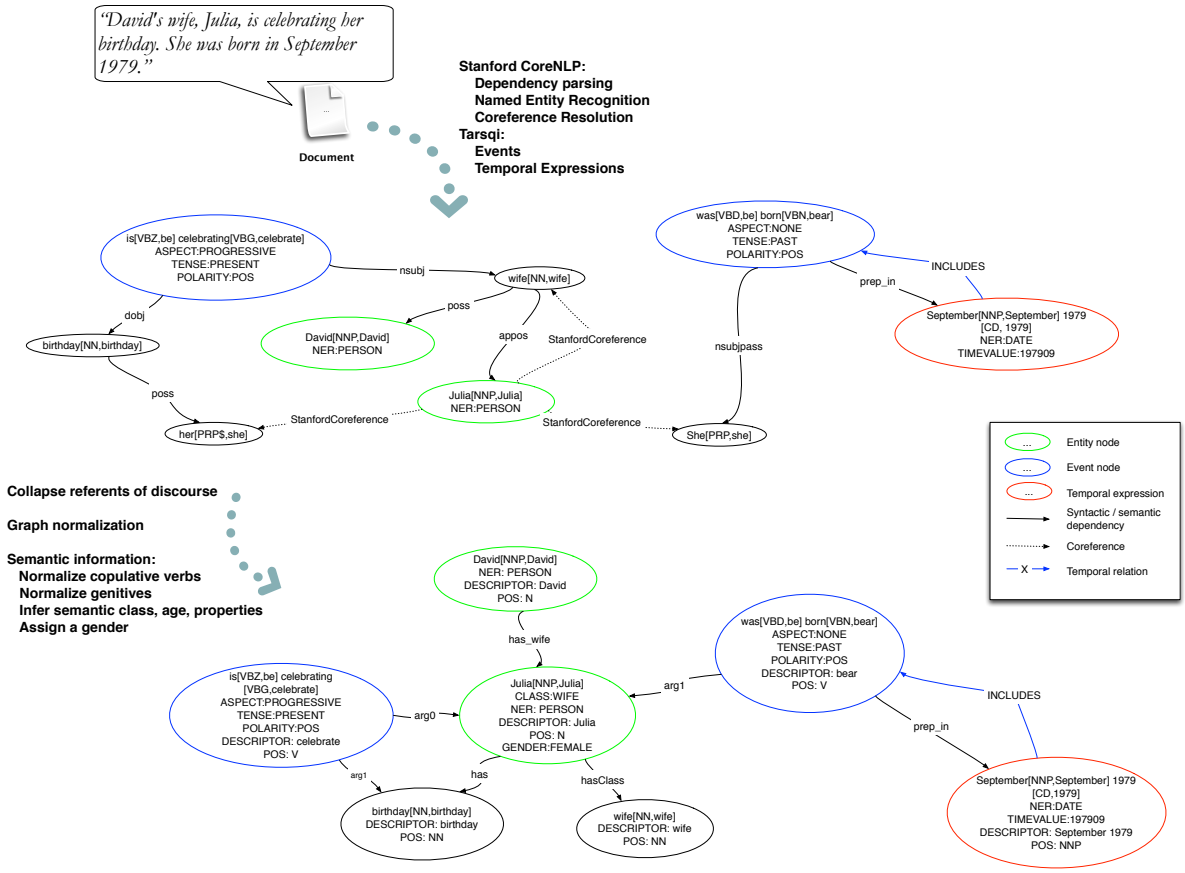


Figura 1: Transformación del grafo inicial, G_D, al colapsado, G_C para el documento de ejemplo: "David's wife, Julia, is celebrating her birthday. She was born in September 1979".

un referente de discurso $r = \cup\, n_0, \ldots, n_k$, creando así un nuevo conjunto de referentes de discurso R_C. Dado otro referente de discurso $r' = \cup\, n'_0, \ldots, n'_j, \ldots, n'_k$, unimos ambos referentes con una arista si algunos de sus nodos estaban unidos, es decir: $\exists\, a(n_i, n'_j) \implies \exists\, a'(r, r')$

Los atributos que consideramos para los referentes de discurso son ligeramente distintos que los de los nodos. Conservan un descriptor y una etiqueta morfosintáctica, pero no los lemas ni las palabras de los nodos origen, así como las categorías de los nodos que los componen (eventos, expresiones temporales y entidades nombradas).

El colapsado provoca que nodos anotados con diferentes atributos se agrupen. Estos atributos pueden reforzar una evidencia o aportar datos complementarios, pero en ocasiones también muestran datos contradictorios. En estos casos, se han tomado decisiones distintas dependiendo del atributo que se esté tratando.

El descriptor debe ser la cadena de caracteres más representativa del referente de discurso. Por ello, como primera aproximación se ha empleado el descriptor más largo de los nodos de origen. En un futuro, podría crearse una base de conocimiento con las distintas concurrencias de descriptores, y a partir de ella escoger el correcto. Esto funcionaría de manera similar a un sistema de desambiguación.

Para la etiqueta morfosintáctica, hemos escogido asignar a las entidades nombradas el valor *N* y a los eventos *V*, mientras que para el resto de palabras se mantiene la correspondiente al nodo de origen. Esto es una aproximación sencilla que busca la normalización de los nodos que suelen contener la información más importante.

El colapsado puede suponer también que algunos referentes de discurso tengan varias etiquetas de una misma categoría. Esto no supone un problema en el caso de los eventos o las expresiones temporales, ya que siempre agregan información complementaria, pero sí en el caso de las entidades nombradas, que pueden tener tipos distintos. En este caso, se escogen aquellos tipos que sean más co-

munes, descartando los demás. Al igual que los descriptores, esta información se puede almacenar en una base de conocimiento y recuperarla para mejorar el proceso de asignación de tipos.

Tanto el proceso de colapsado como el resto de tareas, incluyendo la explicitación de la información semántica y la simplificación y normalización del grafo se han realizado mediante las reglas incluidas en el cuadro 1.

2.3. Procesamiento

Para crear esta representación, se ha seguido un proceso que consta de las siguientes fases: (1)Análisis morfosintáctico, reconocimiento de entidades nombradas y resolución de correferencia. Efectuado con el programa Stanford CoreNLP (Klein y Manning, 2003). (2)Etiquetado de eventos y expresiones temporales, e inclusión de aristas con relaciones temporales mediante el programa Tarsqi Toolkit (Verhagen et al., 2005). (3)Colapsado de nodos en referentes de discurso. (4)Simplificación y normalización del grafo mediante reglas.

En la Figura 1 se pueden observar los grafos con las dos configuraciones. En la configuración inicial (grafo superior) se han realizado los pasos 1 y 2, mientras que en la configuración colapsada (grafo inferior) se han añadido los pasos 3 y 4.

3. Evaluación

Dado que el objetivo de la transformación de documentos en grafos es mejorar el rendimiento de aplicaciones que empleen esta representación, realizaremos la evaluación de su funcionamiento de manera extrínseca. Para evaluar el sistema hemos escogido la tarea Slot Filling (Ji, Grishman, y Dang, 2011) en el marco del Knowledge Base Population en la conferencia Text Analysis Conference.

Para esta tarea se distribuye una base de conocimiento creada a partir de las infoboxes de Wikipedia que contiene las referencias para entrenar los sistemas, más una colección 1.7 millones de documentos de diversas fuentes, como noticias, conversaciones telefónicas y texto web, que será el espacio de búsqueda de soluciones.

Slot Filling se divide en dos subtareas, Regular Slot Filling (RSF) y Temporal Slot Filling (TSF). La evaluación de la representación la hemos realizado sobre esta última.

El objetivo es encontrar el valor de una lista cerrada de atributos para diversas entidades y acotarlo temporalmente. Las entidades pueden ser personas u organizaciones, mientras que los atributos dependen de la entidad, para personas son:*estados de residencia*, *ciudades de residencia*, *títulos*, *miembro de*, *empleado de*, *esposos*; mientras que para organizaciones es unicamente *directivos*.

Para esto se emplean una serie de consultas compuestas por la dupla $\langle entidad, atributo \rangle$, a las que los sistemas deben responder con una tripleta $\langle entidad, atributo, valor \rangle$. Por ejemplo, para la consulta $\langle Barack\ Obama, spouse \rangle$ la respuesta debería ser $\langle Barack\ Obama, spouse, Michelle\ Obama \rangle$.

La respuesta correcta a un atributo puede consistir en una lista de valores. Solo se valoran los valores correctos, y las respuestas redundantes son ignoradas.

En la subtarea TSF se pide además acotar temporalmente las tripletas. Para ello, se define un intervalo temporal impreciso como una tupla de cuatro valores (t_1, t_2, t_3, t_4), que indican que la relación comienza en un punto entre t_1 y t_2 y termina en otro entre t_3 y t_4. Si el valor de t_1 o t_3 está sin rellenar significa que el valor es $-\infty$, mientras que para t_2 o t_4 es $+\infty$.

Para responder a las preguntas de investigación planteamos los siguientes experimentos:

(1) Medir nuestro sistema frente a otros participantes para comprobar si es apropiada la representación en forma de grafo para la tarea.

(2) Entrenar dos clasificadores con representaciones distintas para evaluar la utilidad del colapsado, la normalización y el enriquecimiento semántico. En el primero empleamos la configuración inicial, mientras que en el segundo utilizamos la configuración colapsada.

3.1. Resultados

El cuadro 2 muestra los resultados generales de la tarea Temporal Slot Filling. En ella se puede observar que el funcionamiento del sistema ha sido similar a otros sistemas en el estado del arte, siendo la precisión la más alta de todos los participantes. Además, a pesar de la baja cobertura, el sistema obtiene la tercera mejor medida F_1.

Esto implica que la representación gráfica supone una posibilidad prometedora en la

Colapsado de Referentes de Discurso	
X correference Y	$r = X \cup Y$
X dep1 Y, X dep2 Y, $dep1 = dep\ 2$	X dep1 Y
X nn Y, X amod Y	X nn Y
X dep X	
X class Y, X correference Y	X class Y
Clases Semánticas	
Antecedente	**Consecuente**
NE nn NN	NE hasClass NN
NE appos NN	NE hasClass NN
NE abbrev NN	NE hasClass NN
NN appos NE	NE hasClass NN
NN abbrev NE	NE hasClass NN
NE nsubj NN	NE hasClass NN
NE is NN	NE hasClass NN
Propiedades	
Antecedente	**Consecuente**
JJ nsubj X, JJ cop Y	X has_property JJ
JJ arg0 X, JJ cop Y	X has_property JJ
Genitivos	
Antecedente	**Consecuente**
NN nn NE	NE has_NN NN
NN poss NE	NE has_NN NN
X poss Y	Y has X
NN prep_of NE	NE has_NN NN
X has NE , NE has Class	X has_Class NE, NE has Class
NN nsubj NE	NE has_NN NN
Normalización de Argumentos Verbales	
Antecedente	**Consecuente**
V nsubj X	V arg0 X
V xsubj X	V arg0 X
V csubj X	V arg0 X
V agent X	V arg0 X
V nsubjpass X, V arg1 Y	V arg1 Y, V arg2 X
V nsubjpass X	V arg1 X
V dobj X	V arg1 X
V iobj X	V arg2 X
X partmod V	V arg1 X
V xcomp X	V arg1 X
V ccomp X	V arg1 X
V xcomp X, V arg1 Y	V V arg1 Y, arg2 X
V ccomp X, V arg1 Y	V arg1 Y, V arg2 X
Copulativos	
Antecedente	**Consecuente**
NN nsubj X, NN cop Y	X is NN
Edad	
Antecedente	**Consecuente**
PERSON appos NUMBER	NE hasAge NUMBER
PERSON abbrev NUMBER	NE hasAge NUMBER
Género	
Antecedente	**Consecuente**
NE coreference he	NE hasGender MALE
NE coreference she	NE hasGender FEMALE
he coreference NE	NE hasGender MALE
she coreference NE	NE hasGender FEMALE

Cuadro 1: Reglas para la transformación de los grafos. Cada columna se corresponde con una tripleta $\langle governor,\ dependency,\ dependant \rangle$, donde X e Y son dos nodos diferentes del grafo. Nótese que los antecedentes que no aparecen en el consecuente son borrados.

tarea de extracción automática de relaciones, ya que a pesar de encontrarse en las primeras fases de desarrollo permite competir con los sistemas en el estado del arte.

El sistema está formado por varios componentes que funcionan en cadena: Recuperación de información, representación de documentos, aprendizaje semisupervisado, ex-

System	Precision	Recall	F1
BLENDER2	0.1749	0.3261	0.2277
BLENDER1	0.1749	0.3172	0.2255
BLENDER3	0.1642	0.3099	0.2147
IIRG1	0.2404	0.1299	0.1711
Colapsado	**0.2571**	**0.0656**	**0.1045**
Inicial	**0.2299**	**0.0620**	**0.0977**
Stanford 12	0.0206	0.1724	0.0369
Stanford 11	0.0211	0.1491	0.0370
USFD20112	0.0099	0.0053	0.0069
USFD20113	0.0019	0.0004	0.0006

Cuadro 2: Resultados finales de la tarea Temporal Slot Filling

tracción de relaciones y acotación temporal. Este tipo de sistema se ve afectado por la propagación de errores, por lo que es interesante aislar el impacto de la representación de la influencia de otros factores. Por ello estudiaremos la precisión, cobertura y medida F_1 tras la etapa de extracción de relaciones, ya que la acotación temporal es una etapa que es relativamente independiente de la representación.

En el cuadro 3 se muestra los resultados en la tarea Temporal Slot Filling tras la fase de extracción de relaciones. En ella se muestra la precisión y la cobertura calculadas teniendo en cuenta el número de tripletas $\langle$*entidad*, *atributo*, *valor*$\rangle$ obtenidas correctamente.

Los datos muestran que tanto la precisión como la cobertura obtienen valores muy similares en ambos tipos de representaciones, siendo ligeramente mejores en el caso de los grafos colapsados.

Una inspección manual de los datos muestra que los grafos generados contienen numerosos errores, principalmente en la fase de identificación de descriptores en los grafos colapsados, aunque también otros menores a lo largo del procesamiento. Sin embargo, estos errores parecen compensarse con las mejoras, ya que los resultados no sólo no bajan sino que mejoran.

Con estos datos podemos afirmar que la fase de enriquecimiento es útil para la tarea de extracción automática de relaciones, ya que el balance entre las ganancias resultantes de colapsar los grafos y los errores introducidos por este procesamiento es positivo, y esperamos que un refinamiento posterior haga que mejore todavía más.

Configuración	Inicial	Colapsada
Cobertura	0.08	0.08
Precisión	0.42	0.45
F1	0.14	0.14

Cuadro 3: Resultados de la tarea Temporal Slot Filling en la fase de extracción

4. Trabajo Relacionado

El estudio de la representación de los textos es muy amplio. Los primeros acercamientos están basados en las teorías de dependencia conceptual de Schank (1972), y la de significado-texto (*meaning-text theory*) de Mel'cuk y Polguère (1987). Posteriormente se utilizaron técnicas de representación en grandes colecciones de textos, como por ejemplo en el trabajo de Pradhan et al. (1994) aplicado al dominio médico. Sin embargo cada tarea de procesamiento de lenguaje tiene unas necesidades de representación distintas.

La extracción de información tal y como la entendemos en este trabajo comenzó a estudiarse en la Message Understanding Conference (MUC) (Grishman y Sundheim, 1996; Beth, 1995; Chinchor, 1998) y continuó con el programa Automatic Content Extraction (ACE) (Maynard, Bontcheva, y Cunningham, 2003) hasta 2008. Estas fueron las primeras evaluaciones cuantitativas de sistemas de este tipo y en ellas se profundizó en el procesamiento automático de texto.

En 2009, la Text Analisis Conference (TAC) (McNamee y Dang, 2009) tomó el relevo de ACE en las evaluaciones de sistemas de recuperación de información. En esta conferencia se propuso la tarea Knowledge Base Representation (KBP), que se puede ver como una combinación de extracción de información y búsqueda de respuestas en la que se complicaba la tarea al forzar la adquisición de información en múltiples documentos. En esta tarea se enmarcan las subtareas Regular Slot Filling (SF) y Temporal Slot Filling (TSF).

Como se puede ver en (Ji y Grishman, 2011), SF y TSF permanecen como problemas de investigación abiertos ya que los competidores todavía no son capaces de acercarse a los resultados de la anotación manual. Los participantes en esta edición utilizan sistemas similares de representación, por ejemplo, los sistemas de CUNY (Artiles et al.,

2011) y Stanford (Surdeanu et al., 2011) emplean también tokenización, segmentación, detección de entidades nombradas, resolución de correferencia y análisis de dependencias sintácticas. En el caso del sistema de CUNY también emplean n-gramas de varias longitudes como modelos de bolsa de palabras. Sin embargo ninguno de ellos emplea una representación gráfica a nivel de documento como aquí se propone.

5. Conclusiones

Este trabajo muestra una representación de documentos como grafos morfosintácticos enriquecidos semánticamente. Con ella, disponemos de la base para realizar múltiples tareas de procesamiento de lenguaje natural como resolución de correferencias, extracción de información o búsqueda de respuestas.

Los resultados obtenidos nos permiten comprobar que este tipo de representación resulta prometedora para el funcionamiento de los clasificadores en la tarea de extracción automática de relaciones. El estado de evolución de la representación nos invita a pensar que quedan muchas alternativas por explorar y que una versión más avanzada de la misma nos permitirá mejorar los resultados.

A pesar de los errores del procesamiento, comprobamos cómo los grafos colapsados se comportan mejor que los grafos iniciales. Estas nuevas estructuras nos permitirán realizar procesos como agregación de información, asignación automática de clases semánticas o desambiguación de entidades, que podrían utilizarse como base de conocimiento para enriquecer nuevos grafos.

Además, el proceso de creación de esta representación está basado en herramientas externas intercambiables, lo que permite tener una aplicación modular y flexible en la que encajan fácilmente futuros cambios.

Sin embargo, todavía existe mucho margen de mejora en la representación. Dado que nuestro objetivo nos obligaba a agrupar información muy heterogénea, hemos empleado técnicas sencillas en cada uno de los pasos.

6. Trabajo futuro

La información de un texto que es fácilmente predecible por el lector tiende a omitirse, como se apunta en (Peñas y Ovchinnikova, 2012). Sustituir las aristas genéricas de estructuras que suelan codificar este tipo de información, como genitivos o compuestos nominales, por otras más específicas que indiquen la naturaleza de la relación entre los componentes podría ayudar a mejorar los sistemas de extracción de información.

Otra posibilidad de mejora es incluir un sistema de correferencia de eventos (Humphreys, Gaizauskas, y Azzam, 1997; Hasler, Orasan, y Naumann, 2008), y realizar un proceso de colapsado similar al que se emplea con la correferencia de entidades nombradas. De esta manera se conseguiría una mayor cohesión de la información.

Por último, sería muy interesante generar bases de conocimiento agregando automáticamente la información de distintos documentos de manera similar a (Peñas y Hovy, 2010; Banko et al., 2007; Clark y Harrison, 2009). Por ejemplo, podríamos seleccionar subgrafos para seleccionar estructuras morfosintácticas o semánticas interesantes, o crear ontologías a partir de las clases semánticas inferidas. Esta información podría utilizarse para enriquecer los grafos y mejorar así la cobertura del sistema de extracción de relaciones.

Bibliografía

Artiles, Javier, Qi Li, Taylor Cassidy, Suzanne Tamang, y Heng Ji. 2011. Cuny blender tac-kbp2011 temporal slot filling system description. En *Proceedings of the Text Analysis Conference*.

Banko, Michele, Michael J. Cafarella, Stephen Soderl, Matt Broadhead, y Oren Etzioni. 2007. Open information extraction from the web. En *In IJCAI*, páginas 2670–2676.

Beth, Sundheim. 1995. Proceedings of the sixth message understanding conference. MUC-6, Columbia, MD.

Chinchor, Nancy A. 1998. Overview of proceedings of the seventh message understanding conference. En *Proceedings of the Seventh Message Understanding Conference*, MUC-7, Fairfax, VA.

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.

Grishman, Ralph y Beth Sundheim. 1996. Message understanding conference-6: a

brief history. En *Proceedings of the 16th International Conference on Computational Linguistics*, ICCL, páginas 466–471.

Hasler, Laura, Constantin Orasan, y Karin Naumann. 2008. Nps for events: Experiments in coreference annotation.

Heim, Irene, 1983. *File Change Semantics and the Familiarity Theory of Definiteness*, páginas 164–189. Walter de Gruyter.

Humphreys, Kevin, Robert Gaizauskas, y Saliha Azzam. 1997. Event coreference for information extraction.

Ji, Heng y Ralph Grishman. 2011. Knowledge Base Population: Successful Approaches and Challenges. En *ACL HLT 2011*, páginas 1148–1158.

Ji, Heng, Ralph Grishman, y Hoa Trang Dang. 2011. Overview of the tac2011 knowledge base population track. En *Text Analysis Conference, TAC 2011 Workshop, Notebook Papers*.

Karttunen, Lauri. 1968. *What do referential indices refer to?* Rand Corporation: [Paper]. Rand Corp.

Klein, Dan y Christopher D. Manning. 2003. Accurate unlexicalized parsing. En *ACL 2003*, páginas 423–430.

Maynard, Diana, Kalina Bontcheva, y Hamish Cunningham. 2003. Towards a semantic extraction of named entities. En *In Recent Advances in Natural Language Processing*.

McNamee, Paul y Hoa T. Dang. 2009. Overview of the TAC 2009 knowledge base population track. En *TAC 2009*.

Mel'cuk, Igor y Alain Polguère. 1987. A formal lexicon in the meaning-text theory (or how to do lexica with words). *Computational Linguistics*, 13(3-4):261–275.

Peñas, Anselmo y Eduard Hovy. 2010. Filling knowledge gaps in text for machine reading. En *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, páginas 979–987, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peñas, Anselmo y Ekaterina Ovchinnikova. 2012. Unsupervised acquisition of axioms to paraphrase noun compounds and genitives. En *(Ed.): CICLing 2012, Part I, LNCS 7181*, páginas 388–401, Springer-Verlag.

Pradhan, Malcolm, Gregory Provan, Blackford Middleton, y Max Henrion. 1994. Knowledge engineering for large belief networks. En *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*, UAI'94, páginas 484–490, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Recasens, Marta. 2010. Coreference: Theory, annotation, resolution and evaluation.

Schank, Roger C. 1972. Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology*, 3(4):552–631, October.

Surdeanu, Mihai, Sonal Gupta, John Bauer, David McClosky, Angel X. Chang, Valentin I. Spitkovsky, y Christopher D. Manning. 2011. Stanford's distantly-supervised slot-filling system. En *Proceedings of the Fourth Text Analysis Conference (TAC 2011)*, Gaithersburg, Maryland, USA, November.

Verhagen, Marc, Inderjeet Mani, Roser Sauri, Robert Knippen, Seok Bae Jang, Jessica Littman, Anna Rumshisky, John Phillips, y James Pustejovsky. 2005. Automating temporal annotation with TARSQI. En *ACLdemo'05*.

Lexicografía y Terminología Computacionales

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 67-74 recibido 20-04-12 revisado 29 -05-12 aceptado 05-06-12

Co-occurrence Graphs Applied to Taxonomy Extraction in Scientific and Technical Corpora*

Grafos de coocurrencia aplicados a la extracción de taxonomías en corpus científico-técnicos

Rogelio Nazar(1) **Jorge Vivaldi**(1) **Leo Wanner**(2)
(1) University Institute for Applied Linguistics
(2) Department of Information and Communication Technologies
and Catalan Institution for Research and Advanced Studies (ICREA)
Universitat Pompeu Fabra, C/ Roc Boronat, 138, 08018 Barcelona
{rogelio.nazar; jorge.vivaldi; leo.wanner}@upf.edu

Resumen: Los grafos de coocurrencia léxica han sido utilizados en lingüística computacional en experimentos de desambiguación de sentidos pero hasta ahora no para la extracción de relaciones de hiperonimia, donde la metodología más usual ha sido la aplicación de patrones léxico-sintácticos. En este artículo mostramos que es posible extraer relaciones de hiperonimia entre términos utilizando estadísticas de coocurrencia. La clave del método reside en que las relaciones de coocurrencia no suelen ser simétricas en el caso de las relaciones de hiperonimia y, en consecuencia, es posible generar grafos dirigidos de coocurrencia que guardan una apariencia similar a la de una taxonomía. En el presente artículo presentamos experimentos con textos de la Wikipedia en castellano ordenados aleatoriamente, pero los resultados sugieren que la coocurrencia asimétrica entre términos es una propiedad intrínseca y macroscópica del discurso argumentativo en general.
Palabras clave: Construcción de ontologías; estadísticas de coocurrencia; extracción de taxonomías; lingüística cuantitativa; semántica distribucional.

Abstract: Word co-occurrence graphs have been used in computational linguistics mainly for word sense disambiguation and induction, but until very recently, not for the extraction of hypernymy relations, where the methodology most often applied is the use of lexico-syntactic patterns. In this paper, we show that it is possible to use word co-occurrence statistics to extract IS-A relations between entities in scientific and technical corpora. We exploit the fact that word co-occurrence often has a direction, that is, a term might co-occur with another, but this is very often not true the other way round. This means that one can represent co-occurrence as a directed graph and this graph resembles a taxonomy. In this paper we present an experiment with texts randomly extracted from the Spanish Wikipedia, but our findings suggest that this co-occurrence behavior is a macroscopic and intrinsic property of argumentative discourse in general.
Keywords: Co-occurrence statistics; distributional semantics; ontology learning; quantitative linguistics; taxonomy extraction.

1 Introduction

Having at our disposal software that could automatically induce structured data such as a taxonomy of concepts from unstructured text would be, without any doubt, a substantial improvement to our ability to conduct scientific research[1]. Presently, the corpus of scientific literature has reached such a volume that it is becoming increasingly difficult for a single individual or a group of researchers to follow related work and spot all relevant advances in their fields. The desire to obtain structured data from texts has motivated decades of efforts in the area of automatic information extraction (Harris, 1958;

* This research was funded by project APLE (Spanish Ministry of Science and Education: Ref. FFI2009-12188-C05-01/FILO) lead by Dr. M. Teresa Cabré. The authors would like to thank the anonymous reviewers and Chris Norrdin for proofreading.

[1] The paper is based on a chapter of the first author's PhD thesis (Nazar, 2010).

Grishman, 1997). In this paper, we focus on the single case of the extraction of hypernymy relations between scientific concepts, and present a novel approach to the topic from the perspective of term co-occurrence statistics.

The vast majority of the proposals produced so far (see next section) have opted for the so called "pattern-based" approach to taxonomy extraction, in which the strategy is language dependent and consists of generating lists of lexico-syntactic patterns that presumably convey a hypernymy relation as in the case, for instance, of the following pattern: **X** *is a kind of* **Y**. The limitation of this approach is that lexical patterns not always convey the expected relation and, moreover, hypernymy relations often appear in the corpus expressed in ways that the researcher was not able to anticipate.

The present proposal is different from previous work from the starting point. It disregards the use of explicit lexico-syntactic patterns in favor of an "immanent" approach, that is, a corpus-based approach with no previous conceptions about the language or the domain. Another characteristic of the present proposal is that, because our starting point is not based on patterns, our approach is entity-by-entity based, which means that given a set of input terms the output will be the assignment of the most probable hypernym for each term.

A basic sketch of our reasoning can be easily grasped by considering the following statements, assuming that A, B, C, D, E and F are terms that denote real or conceptual entities in a given language and domain, and when we say that one term has a tendency to co-occur with another we mean a significant frequency of co-occurrence of these terms in a given context window, such as a search engine's snippet. Thus, if:

- A tends to occur with B and C
- B tends to occur with C and D
- C tends to occur with D
- D tends to occur with E and F

then, we assume that B and A are hyponyms of C, while C, in turn is a hyponym of D. Naturally, conclusions of this kind are not based on a small number of cases as in this sketch, but on hundreds of contexts of occurrence of the input terms found in a corpus. A few heuristics, described in detail in Section 3, are added to the procedure. The first is the operational definition of what counts as an entity. For simplicity, terms denoting entities are not selected observing authoritative sources that certify their normative status. Instead, they are selected according to a statistic criterion as word n-grams which show a significant frequency of occurrence. This procedure is overly simplistic and causes a certain amount of noise in the results. It is to be expected that strategies based on syntactic chunking could offer room for improvement in this aspect, these strategies being linked to the fields of Terminology Extraction (TE) and Named Entity Recognition (NER). In any case, it should be clear that the purpose of the present paper is not to offer the best possible results in the extraction of taxonomy links, a task that would demand a combination of different methods and resources, but to test how much can be done using only limited and essentially simple information such as co-occurrence statistics and, eventually, elementary inferences from these data.

2 Related Work

The fields of TE and NER have been evolving independently of efforts in automatic conceptual relation extraction, but are relevant to all methods of taxonomy extraction because before any attempt to extract relations between terms can be undertaken, these terms must be defined in some way. Lack of space prevents us from offering a detailed introduction which would refer to other works on TE (Kageura & Umino, 1996; Vivaldi & Rodríguez, 2011; Nazar, 2011) and NER (Grishman & Sundheim, 1996; Nadeau & Sekine, 2007).

With respect to efforts in automatic taxonomy extraction, reports began to appear shortly after the availability of the first copies of digitalized lexicographic material, sharing the point of view and methodology: of crafting a specific script for each dictionary and processing the definitions identifying the head of the defining phrase as the hypernym candidate (Amsler, 1981; Chodorow et al, 1985; Fox et al, 1988; Alshawi, 1989). These rules are written in the form of lexico-syntactic patterns and can capture not only hypernymy relations but others such as Part-of, Object-of, Location, Purpose, Manner,

Size, Time, Agent, Act-of, Set-of, Inhabitant-of, Follower-of, etc.

When corpus linguistics gained momentum, at the beginning of the nineteen nineties, researchers in the area started to try to derive taxonomies directly from corpora instead of dictionaries. However, the underlying methodology and assumptions were essentially the same as in previous attempts with machine readable dictionaries (Hearst, 1992; Pearson, 1998; Morin, 1999; Meyer, 2001; Rydin, 2002; Auger & Barrière, 2008). Under the influence of these authors, researchers conducting studies on the subject of conceptual relation extraction will typically define first a corpus of a domain to work with and then apply a routine that searches through the space of this corpus for any occurrence of the members of a set of handcrafted patterns with the aid of a concordance extraction tool. Then, for each context of occurrence, they will inspect what kind of entities are at each side of the pattern and, if these entities really hold the desired conceptual relation, then the outcome will be considered a success.

Few works depart from this perspective. There are reports on the use of machine learning techniques to automatically extract the lexico-syntactic patterns, saving thus the effort of creating them manually (Snow et al, 2006). Patterns are learned with the help of seed-patterns or seed term-pairs which instantiate the relation in question, which is gradually expanded with similar instances found in a corpus. Other authors have proposed the use of statistical techniques to find semantic similarities between entities, inducing vector-based thesauri (Grefenstette, 1994; Lin, 1998). The reasoning is that entities which can be classified as, say, beverages, have a distinctive distributional similarity (e.g., *a bottle of X*, *drinking too much X*, etc.).

3 Methods

As already mentioned in the introduction, our approach to taxonomy induction from corpora is based on statistics of term co-occurrence. The context window –or the space for that co-occurrence– is a paragraph of text (in practice, the text between two newline characters). Using a large corpus, we have been able to observe how hypernymy relations are correlated with term co-occurrence and, given the asymmetric property of term association, taxonomic links can be automatically derived without explicit linguistic or ontological knowledge. A description of each of the steps of the experiment follows. Among them, the most important are the study of first order co- occurrence or syntagmatic association (Section 3.3), the study of second order co-occurrence or paradigmatic relation (Section 3.4) and, finally, the representation of these co-occurrence relations in a directed graph (Section 3.5).

3.1 Selection of a sample of terms to be used as input

The idea of analyzing terms in batches of hundreds instead of one-by-one will become apparent in the following subsections. The basic motivation is that the algorithm needs large numbers to obtain reliable estimations. Given a set of input terms in the same language and domain, each will result in the assignment of the most probable hypernym. In practice, these terms can be obtained from a glossary or database or be the result of an extraction of terms from a corpus.

3.2 Compilation of a reference corpus of the language

A reference corpus is needed in order to develop a language model that will allow us to score and highlight the most significant terms. This model consists of the frequencies of occurrence of words and word n-grams in a corpus of general language. A corpus of press articles of an extension of two million words is sufficient to be used as a language model. Of course, more data would produce better results.

3.3 Analysis of first order co-occurrence

The analysis of first order co-occurrence consists in extracting terms that are syntagmatically related to an input term, which is done by sorting the co-occurring vocabulary in decreasing order of frequency. For illustration, consider an example in the field of medicine (Table 1). These are the most frequent n-grams in the first 100 snippets returned by a web search engine using the term *chronic obstructive pulmonary disease* (COPD).

This co-occurring vocabulary is defined as a set of n-grams ($n \leq 3$) with term frequency and document frequency ≤ 3. Units of a

Rank	Term	Freq.
1	copd	45
2	disease	23
3	lung	21
4	lung disease	20
5	chronic bronchitis	18
6	chronic	18
7	chronic obstructive	14
8	bronchitis and emphysema	12
9	emphysema	10
10	known as copd	9
11	copdgroup of lung	9
12	obstructive	8

Table 1: Terms that most frequently co-occur with *chronic obstructive pulmonary disease*.

length of less than 4 characters and multiword units with a first or last element of a length less than 4 characters are eliminated. The rest of the vocabulary is weighted in order to keep only those units that show a significant frequency. The weight is calculated as shown in (1), where i is the n-gram from the frequency lists, $f_o(i)$ the observed relative frequency of i in the analyzed corpus and $f_e(i)$ the expected relative frequency of i, which is its frequency in the reference corpus. We eliminate all units with a score above an empirically determined threshold (0.01), as well as all n-grams with a first or last element in the same condition.

$$w(i) = \log(\frac{f_o(i)}{(f_e(i)+1)}) \quad (1)$$

If a lexical resource for the analyzed language is available, it can be used at this point to filter out units that are not nouns and to change plural forms into their lemmata. In case this linguistic resource is not available, a workaround can be to proceed with a pseudo-lemmatization based on orthographic similarity using a similarity coefficient such as Dice (2) with letter bigrams as features.

$$Dice(I,J) = \frac{2|I \cap J|}{|I| + |J|} \quad (2)$$

In order to avoid the possibility of two components of the same n-gram competing for positions in the rank, we eliminate overlapping units. That is, if a unit forms part of another and both have the same frequency, such as in the case of *chronic* and *chronic bronchitis*, it means that every time *chronic* occurs, it is followed by *bronchitis*. In such cases, only the n-gram with higher n is kept. The absolute frequency of the remaining n-grams is multiplied by their corresponding n.

3.4 Analysis of second order co-occurrence

The analysis of second order co-occurrence is very similar to the first order, the only difference being that it is the re-iteration of the analysis for each of the terms that were found co-occurring with the input term. Thus, if on the first analysis for the term *chronic obstructive pulmonary disease* we found that it is related to *disease* and *lung disease*, etc., now we will submit these new terms to the same process. The result is that, for the initial term *chronic obstructive pulmonary disease*, we will find terms that are related to it, and also terms that are related to these latter terms. That enables us to calculate a dispersion coefficient $D(i,j)$ which measures how recurrent a term j is among the lists of related terms generated from a term i. The rationale is that the correct hypernym term j of a term i not only appears syntagmatically related to i, but it is also related to other co-hyponyms that co-occur with i. This dispersion is calculated by multiplying the observed frequency of a hypernym candidate with the number of times the candidate appeared in the frequency lists, as shown in (3), where $f_o(i,j)$ is the observed frequency of j in the contexts of i and all of its related terms, and $D(i,j)$ is the dispersion of term j in the analysis of i. Table 2 shows the terms that have a significant second-order co-occurrence with the initial input term.

$$wD(i,j) = \log(1 + f_o(i,j) * D(i,j)) \quad (3)$$

Rank	Term	1st ord	2nd ord
1	disease	23	142
2	symptoms	7	134
3	chronic	18	77
4	pain	124	3
5	causes	86	3
6	lung	21	63
7	chronic obstructive	7	36
8	lung disease	10	30
9	chronic bronchitis	9	28

Table 2: Terms that show high frequency of second order co-occurrence with *chronic obstructive pulmonary disease*.

3.5 Generation of a directed graph of term co-occurrence

Once the first and second order co-occurrence has been calculated for a sample of terms, the next step is to establish the hypernymy relations between the terms. The central criteria for the selection of a term j as a hypernym of the term i is that j is in first or second order co-occurrence with i and that j is in the same situation with respect to other input terms. This is represented as $wT(j)$ in (4): the number of times j appears in the hypernym candidate lists of other terms in the input term list, with H as the hypernym list of term i, H_i as the hypernym list of $i \in H$, and $|j \in H_i|$ as the number of times j occurs in H_i).

$$wT(j) = \sum_{i=1}^{|H|} |j \in H_i| \qquad (4)$$

With the hypernym j of i determined, it can be assumed that all members of the candidate list of i that also have j as the most generic term are hypernyms of i. Due to the transitivity property, it is also possible that i ends up as hyponym of a term that is not syntagmatically related. In any case, the result of the process is that for each input term the system will attempt to return one term as the best hypernym candidate available.

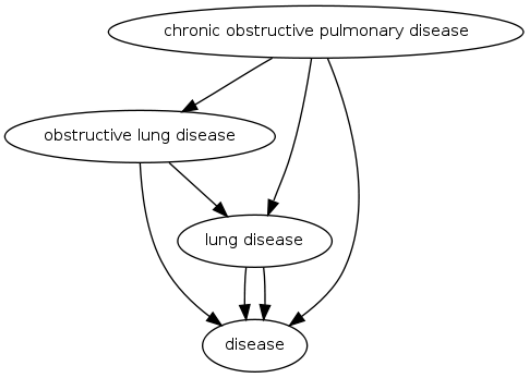


Figure 1: Example of a co-occurrence graph resembling a taxonomy chain for the term *chronic obstructive pulmonary disease*

Figure 1 shows that COPD has a tendency to co-occur with the terms *obstructive lung disease*, *lung disease* and *disease*, but none of these three show a special tendency to co-occur with COPD (the relation is not reciprocal). The term *obstructive lung disease*, in turn, shows a similar tendency to co-occur with *lung disease* and *disease*, but again, none of these two selects *obstructive lung disease* as a frequent co-occurrence. Finally, *lung disease* only co-occurs with *disease* and this last term does not show a tendency to co-occur with any of the above. Coding these relationships as arrows in the graph, we can interpret the figure as a taxonomy, having the number of incoming arrows as a natural expression of hypernymy. The interpretation is that COPD is a kind of disease.

3.6 Analogical inference

Working with corpus based methods has the shortcoming that very often the terms we are analyzing are not found in the corpus with sufficient frequency. In order to overcome this limitation, we have extended the co-occurrence model with the addition of a layer of analogical inference.

In essence, the idea is that if a given term is not found in the corpus, we will attempt to find some similarity with other terms which have been found in the corpus and effectively assigned a hypernym. For instance, when the algorithm has found that in repeated occasions terms such as *lung disease* or *celiac disease* have the term *disease* as hypernym, then it can safely assume that another term that was not found in the corpus but has clear similarities such as, for instance, *Knights disease*, is also another kind of disease. Notice that this cannot be computed by a simple overlapping measure, i.e., we cannot assume that *Knights disease* is a *disease* just because the word *disease* is included in the term *Knights disease*. Doing so would also lead us to other wrong assumptions, for instance that *lichen planus* is a kind of *planus* or that a *transforming growth factor* is a kind of *factor* when actually they are a kind of *disease* and a kind of *protein*, respectively. The procedure is, thus, not just to find overlapping sequences but to learn to associate features in the terms with the hypernyms they have been assigned by the co-occurrence method. This reasoning allows us to operate in the same way in cases where there is actually no overlapping. For instance, when the algorithm finds that terms such as *Carpenter syndrome* and *Asperger syndrome* consistently receive the term *disease* as hypernym, it will assume that other terms that were not found in the corpus but share the same element *syndrome*, such as *Meretoja syndrome* or *Maffuci syndrome*,

can also be considered diseases by means of simple analogy.

The same reasoning is applied at the morphological level, which in this case is defined as the first and last four letters of each word of a term. The motivation behind this procedure is that very often in specialized terminology the units that pertain to a particular semantic class share some morphological features. Thus, if the algorithm finds a persistent morphological pattern within the terms of a determined semantic class, it will learn to associate such pattern with that class. For example, if it finds that terms ending with a sequence of letters such as *-itis* or *-osis* which are frequently assigned the hypernym *disease*, as in the case of *arthritis* or *endometriosis*, then it will assume that other terms that were not found in the corpus such as *acrodermatitis* or *pneumocystosis* can also be classified as diseases.

The main benefit of this analogical inference is that it grants the algorithm a great amount of flexibility and generalization power, because there is no need for explicit information about the entities nor any kind of previous training phase. The learning is conducted on the fly using the result of the co-occurrence method.

4 Results and Evaluation

In order to evaluate the strategy, we took a sample of 375 terms in Spanish from the Mosby (2003) dictionary, pertaining to the classes of bones, disorders, ganglia, glands, hormones, drugs, organs, proteins and viruses in unequal proportions. This facilitates the task of evaluation because we know that the correct hypernym of each input term must pertain to some of these classes (which is information that the algorithm does not have). As mentioned earlier, in a real life scenario these input terms would be obtained, for instance, by term extraction from LSP corpora. In this experiment we used texts from the Spanish Wikipedia of the year 2010 as corpus, in random order and excluding, of course, all metadata and structural information that could be used to construct a taxonomy, leaving a single text file of approximately 455 million tokens. The choice of this corpus does not mean that we are particularly interested in Wikipedia. In fact, we believe the experiment could be replicated with any other corpus if it is large and diverse enough to contain the input terms with sufficient frequency. It is probable that better results could have been obtained using the web as corpus, but then there would be other factors that we would not be able to quantify, such as the particular ranking of results provided by each search engine.

Only 164 of the 375 input terms appeared with sufficient frequency to select co-occurring terms. Those which did appear in the corpus were assigned a hypernym-candidate, which in the majority of the cases was indeed a correct hypernym and when it was not, it was a semantically related concept. Table 3 shows the evaluation figures for all the experiments. Results are reported both with the co-occurrence method as explained in Sections 3.1. to 3.5. and including the analogical inference layer described in Section 3.6. As we can see, the inference layer dramatically increases figures of recall. This is of course important for a practical application, but the key point that we wanted to demonstrate with this experiment is that asymmetric co-occurrence by itself is sufficient to show a significant correlation with hypernymy relations.

It should also be noticed that the task of assigning a hypernym to a given term is performed with variable levels of precision depending on the domains. In the case of the terms pertaining to the class of ganglia, the system found virtually no occurrence of them in the corpus, and this explains why there are so many zeros in their case (they are false negatives). However, because of their extremely regular form – most of them happen to be terms such as *ganglio intercostal* (intercostal node), *ganglio inguinal superficial* (superficial inguinal node), *ganglio gástrico* (gastric ganglion), and so on – this makes it possible for the inference engine to assign a correct hypernym *ganglio* (node/ganglion) to all of them. There are also many zeros in the case of organs, but this time for a different reason. They did appear but the performance was extraordinarily poor because organs are not associated with the hypernym *órgano* (organ) but, instead, to a meronym like *cuerpo* (body) and are, therefore, false positives. Something similar occurs with organs that are related to diseases, e.g. *próstata* (prostate) as a type of cancer.

		Co-occurrence						Co-occurrence + Inference					
Domain	**Trials**	**tp**	**fp**	**fn**	**P**	**R**	**F1**	**tp**	**fp**	**fn**	**P**	**R**	**F1**
Bones	36	27	1	8	96.43	75	84.38	32	1	3	96.97	88.89	92.75
Disorders	68	27	3	38	90	39.71	55.10	56	2	10	96.55	82.35	88.89
Ganglia	29	0	0	29	0	0	0	29	0	0	100	100	100
Glands	15	3	3	9	50	20	28.57	15	0	0	100	100	100
Hormones	43	25	7	11	78.13	58.14	66.67	31	7	5	81.58	72.09	76.54
Drugs	69	4	4	61	50	5.80	10.39	27	7	35	79.41	39.13	52.43
Organs	29	0	22	7	0	0	0	0	22	7	0	0	0
Proteins	65	15	15	35	50	23.08	31.58	31	24	10	56.36	47.69	51.67
Virus	21	2	6	13	25	9.52	13.79	15	6	0	71.43	71.43	71.43
TOTAL	375	103	61	211	62.80	27.47	38.22	236	69	70	77.38	62.93	69.41

Table 3: Results after 375 experiments.

In this case, there is nothing the inference engine can do, because it does not rectify the output of the co-occurrence method. Other causes of mistakes have been the wrong segmentation of multi-word terms and also errors due to problems of polysemy.

We cannot offer a comparison to other authors' results because it is technically impossible given the fact that our method is entity based, thus there is no way of replicating a pattern based method using our dataset. In any case, compared with state-of-the-art techniques, an F1 of %69 is not a very impressive result. In fact, we can expect to obtain higher precision figures with a very simple baseline, such as taking the lexical unit that is the head of the phrase in multiword terminology (normally the first noun from the left in the case of Spanish noun phrases). That said, our results are more meaningful than those that can be obtained with a trivial baseline: undoubtedly, *bocio* (goitre) is the correct hypernym of *bocio coloide* (colloid goiter), and *citocromo* (cytochrome) is of *citocromo P-450* (cytochrome P450), but for most NLP applications, hypernyms such as *enfermedad* (disease) and *enzima* (enzyme) are more meaningful. This is one of those cases in computational linguistics where the interesting point is not to have obtained figures representing better precision than other methods but to have found a methodology to extract information that could not be obtained otherwise.

5 Conclusions and Future Work

This paper has presented an experiment in taxonomy extraction from corpus using purely statistical methods. Our approach to the topic is fundamentally theoretical, though based on empirical evidence. We believe we have found a measurable pattern of term co-occurrence which is characteristic of hypernymy relations, and this property is expected to be inherent to argumentative discourse independent of the language and the domain. Still, much experimentation has to be carried out before reaching conclusive results, especially to sustain the claim of language independence. However, and despite our theoretical motivation, nothing seems to prevent an application of this algorithm as a method for automatic development of taxonomies from corpora at least in the languages where it has already been tested (English and Spanish, so far).

As an interpretation of why this method yields results, we recall two discursive strategies that can be identified in scientific or argumentative discourse. One of the strategies is to introduce and define concepts in discourse according to the knowledge established in the community targeted by the text. To present something new is the purpose of the other strategy used in the text. In this case, the statements convey external or empirical information which cannot be directly inferred from the established knowledge (or at least not trivially). One of the consequences of these two forces acting upon discourse is that we can expect a certain degree of coincidence in the passages where authors introduce concepts in their texts, and this coincidence can be measured in the selection of relevant conceptual features, often hypernym terms.

Future work will include replicating the same model in different languages and domains and introducing different degrees of explicit linguistic knowledge such as POS-tagging, chunking, lexico-syntactic patterns and also ontologies and other semantic re-

sources. We expect to automatically produce high quality taxonomies in the near future with a combination of different techniques.

References

Alshawi, H. 1989. Computational lexicography for natural language processing. *Analysing the dictionary definitions.* Longman Publishing Group, (White Plains, NY, USA): 153–169.

Amsler, R. 1981. A taxonomy for English nouns and verbs. Proceedings of 19th annual meeting on Association for Computational Linguistics. (Morristown, NJ, USA): 133–138.

Auger, A. & Barrière, C. 2008. Pattern-based Approaches to Semantic Relation Extraction Special issue of Terminology. *Terminology* 14(1).

Chodorow, M. & Byrd, R. & Heidorn, G. 1985. Extracting semantic hierarchies from a large on-line dictionary. Proceedings of the 23rd annual meeting on Association for Computational Linguistics, July 08-12, 1985 (Chicago, Illinois, USA): 299–304.

Fox, E. & Nutter, J. & Ahlswede, T. & Evens, M. & Markowitz, J. 1988. Building a large thesaurus for information retrieval. Proceedings of the 2nd conference on Applied natural language processing, Morristown, NJ, USA. Association for Computational Linguistics: 101–108.

Grefenstette, G. 1994. *Explorations in Automatic Thesaurus Construction.* Kluwer, Dordrecht, The Netherlands.

Grishman, R. 1997. Information Extraction: Techniques and Challenges. *Information Extraction*, ed. Maria Teresa Pazienza, Springer-Verlag.

Grishman, R. & Sundheim, B. 1996. Message Understanding Conference 6: a brief history. Proceedings of the 16th International Conference on Computational Linguistics (Copenhagen, Denmark): 466-471.

Harris, Z. 1958. Linguistic transformations for information retrieval. Proceedings of the 16th International Conference on Scientific Information. National Academy of Sciences-National Research Council (Whashington DC, USA).

Hearst, M. 1992. Automatic acquisition of hyponyms from large text corpora. Proceedings of the 14th International Conference on Computational Linguistics (Nantes, France): 539–545.

Kageura, K. & Umino, B. 1996. Methods of Automatic Term Recognition. *Terminology*, 3(2): 259–290.

Lin, D. 1998. Automatic retrieval and clustering of similar words. Proceedings of COLING-ACL: 768–774.

Meyer, I. 2001. Extracting knowledge-rich contexts for terminography. Recent Advances in Computational Terminology.

Morin, E. 1999. Extraction de liens sémantiques entre termes à partir de corpus de textes techniques. PhD Thesis in Computer Sciences, Université de Nantes, 1999.

Mosby 2003. Diccionario Mosby de medicina, enfermería y ciencias de la salud. VI Edición. Madrid: Elsevier.

Nadeau, D., Sekine, S. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes* 30(1):3–26.

Nazar, R. 2010. A Quantitative Approach to Concept Analysis. PhD thesis, Universitat Pompeu Fabra.

Nazar, R. 2011. A Statistical Approach to Term Extraction. *International Journal of English Studies* 11(2):153-176.

Pearson, J. 1998. *Terms in context.* John Benjamins.

Rydin, S. 2002. Building a hyponymy lexicon with hierarchical structure. Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition. Association for Computational Linguistics (Morristown, NJ, USA): 26–33.

Snow, R., Jurafsky, D. & Ng, A. 2006. Semantic taxonomy induction from heterogeneous evidence. Proceedings of the 21st International Conference on Computational Linguistics (Sydney, Australia): 801–808.

Vivaldi, J.; Rodríguez, H. 2011. *Extracting Terminology from Wikipedia. Procesamiento del lenguaje natural* 47: 65–73.

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 75-82 recibido 05-05-12 revisado 01-06-12 aceptado 05-06-12

Análisis del uso de métodos de similitud léxica con conocimiento semántico superficial para mapear la información de enfermería en español*

Analyzing the Use of Shallow Semantic Knowledge with Similarity Methods for Mapping Nursing Information in Spanish

Jorge Cruanes
Dep. Leng. y Sist. Inf.
Universidad de Alicante
03690, Alicante(Spain)
jcruanes@dlsi.ua.es

M. Teresa Romá-Ferri
Dep. Enfermería
Universidad de Alicante
03690, Alicante(Spain)
mtr.ferri@ua.es

Elena Lloret Pastor
Dep. Leng. y Sist. Inf.
Universidad de Alicante
03690, Alicante(Spain)
elloret@dlsi.ua.es

Resumen: Uno de los problemas actuales en el dominio de la salud es reutilizar y compartir la información clínica entre profesionales, ya que ésta se encuentra escrita usando terminologías específicas. Una posible solución es usar un recurso de conocimiento común sobre el que mapear la información existente. Nuestro objetivo es comprobar si la adición de conocimiento semántico superficial puede mejorar los mapeados establecidos. Para ello experimentamos con un conjunto de etiquetas de NANDA-I y con un conjunto de descripciones de SNOMED-CT en castellano. Los resultados obtenidos en los experimentos muestran que la inclusión de conocimiento semántico superficial mejora significativamente el mapeado léxico entre los dos recursos estudiados.
Palabras clave: PLN, similitud léxica, salud, enfermería, NANDA-I, SNOMED-CT

Abstract: In health domain, one of the current problems is the reusing and the sharing the clinical information between professionals, due to this information is written using specific terminologies. One possible solution is to use a common knowledge resource for mapping the existing information. In this paper, our aim is to analyze if the use of lexical similarity algorithms enriched with shallow semantic knowledge can improve these mappings. In order to achieve this, we experiment with a set of NANDA-I labels and a set of SNOMED-CT descriptions in Spanish. The results obtained show that the addition of shallow semantic knowledge significantly improves the lexical mapping between both studied resources.
Keywords: NLP, lexical similarity, health, nursing, NANDA-I, SNOMED-CT

1. Introducción

El reto actual en el dominio de la salud es contar con sistemas que permitan una rápida y eficiente manera de compartir información. Esta información en el caso específico de enfermería es fundamental, ya que suelen reflejar tanto los problemas del paciente o los diagnósticos de enfermería como las intervenciones realizadas y los resultados obtenidos.

Con el objetivo de conseguir esta eficiencia en la comunicación entre los profesionales de la salud, muchos países han comenzado a utilizar la Historia Clínica Electrónica (HCE). Estos sistemas permiten a los profesionales incluir texto estructurado y texto libre en los apartados para narración, empleando sus propios términos. Esto conlleva el empleo de diferentes palabras o etiquetas para referirse a un mismo concepto, generando una heterogeneidad en la información que dificulta compartirla y reutilizarla (Romá-Ferri y Palomar, 2008; Zwaanswijk et al., 2011).

Una posible solución pasa por la existencia de un recurso de conocimiento que permita la interconexión de las terminologías existentes. Para conseguir mapear la información escrita en lenguaje natural sobre estos recursos de

* Este artículo ha sido cofinanciado por el Ministerio de Ciencia e Innovación (proyecto TIN2009-13391-C04-01), y la Conselleria d'Educació de la Generalitat Valenciana (proyectos PROMETEO/2009/119, ACOMP/2010/286 y ACOMP/2011/001). Damos las gracias a los revisores por los comentarios constructivos, que nos han permitido mejorar este trabajo.

conocimiento es necesario el uso de métodos automáticos, debido a la gran cantidad de información existente así como a las constantes actualizaciones de las terminologías. En la actualidad, una de las soluciones adoptadas es la aplicación de métodos de similitud léxica (Meizoso, Allones y Taboada, 2011), ya que éstos son independientes del idioma y no requieren de corpus previamente etiquetados. Estas técnicas determinan cuán relacionadas están entre sí dos expresiones en base a su similitud de escritura. Sin embargo, están limitadas, pues no son capaces de discernir la semántica.

Tratando de ayudar a superar este problema, en este trabajo proponemos una aproximación de mapeado de información en lenguaje natural en el dominio de enfermería en castellano, basada en el uso de métodos de similitud léxica junto con conocimiento semántico superficial. El uso de semántica superficial pretende mejorar la evaluación de la similitud así como aportar una mayor eficacia en el mapeado.

Para verificar la eficacia de nuestra propuesta usaremos como origen la terminología NANDA-I (*North American Nursing Diagnosis Association International*) (NANDA-I, 2010) y como destino la terminología SNOMED-CT (*Systematized Nomenclature of Medicine-Clinical Terms*) (International Health Terminology Standards Development Organisation -IHTSDO-, 2010). NANDA-I consiste en una lista de diagnósticos de enfermería expresados en lenguaje natural y SNOMED-CT es una terminología clínica integral, formalizada y que permite la interoperabilidad semántica entre distintos sistemas, dando soporte a la diversidad lingüística. La terminología SNOMED-CT es la terminología de referencia que ha sido seleccionada para la HCE del Sistema Nacional de Salud español[1].

Este artículo está organizado en cinco secciones diferentes, comenzando con esta introducción. En la sección 2 se describe el estado de la cuestión. Las secciones 3, 4 y 5 describen, respectivamente, los materiales usados en las pruebas, nuestra propuesta y los experimentos que fueron realizados. La sección 6 muestra las conclusiones y se proponen mejoras para futuras investigaciones.

2. Estado de la cuestión

Para solucionar el problema de mapear diferentes terminologías y texto en lenguaje natural entre sistemas, una de las estrategias es hacer uso de técnicas de similitud léxica. Aunque no es la única, y se han empleado otras estrategias basadas en jerarquías de ontologías (Meizoso, Allones y Taboada, 2011) o métodos estadísticos (Nyström et al., 2010). Sin embargo, en el dominio de enfermería en castellano estas alternativas no son una opción inicialmente. En este ámbito no se cuenta con terminologías formalizadas que especifiquen explícitamente el significado de sus términos, ni con corpus etiquetados.

En el mapeado de información a SNOMED-CT mediante uso de técnicas de similitud léxica en lengua inglesa podemos destacar los trabajos de Patrick y Budd (2006), Wang et al. (2006), Patrick, Wang y Budd (2007) y Stenzhorn et al. (2009). Para el castellano podemos destacar el trabajo de Farfán Sedano et al. (2009). En estos trabajos la terminología origen para el mapeo es diversa. En unos casos se optó por las expresiones incluidas en las notas clínicas (Patrick y Budd, 2006; Patrick, Wang y Budd, 2007) o bien por la información sobre medicamentos incluida en una base de datos de farmacia hospitalaria (Farfán Sedano et al., 2009).

Una de las características comunes en los estudios sobre similitud léxica es la de realizar un preprocesamiento (normalización) a los textos a comparar. Esta preparación está ligada a la necesidad de, por ejemplo, reducir errores que se producen al comparar expresiones en mayúsculas con expresiones en minúsculas. Uno de los preprocesamientos más comunes es el de eliminar las palabras conocidas como 'stopwords' y signos de puntuación (Wang et al., 2006).

En cuanto a las técnicas de similitud más usadas son la similitud léxica exacta (Patrick y Budd, 2006; Wang et al., 2006; Patrick, Wang y Budd, 2007; Farfán Sedano et al., 2009) y el algoritmo del Coseno (Stenzhorn et al., 2009). En algunos casos las cadenas comparadas son expandidas para cubrir ciertas variaciones léxicas, como por ejemplo sustituir las abreviaturas por sus expresiones completas (Wang et al., 2006). Otros métodos hacen uso de un lexicón para mejorar el

[1]Sistema Nacional de Salud español. Referencia Web: http://www.msps.es/profesionales/hcdsns/areaRecursosSem/snomed-ct/snomedHCD.htm. Último acceso 5/5/2012.

rendimiento en la búsqueda. En los trabajos de Patrick y Budd (2006) y de Patrick, Wang y Budd (2007) este lexicón es un índice de todas las palabras existentes en SNOMED-CT, asociadas con los identificadores de concepto en los que aparecen. La búsqueda consiste en encontrar el identificador de SNOMED-CT que contenga la mayor subcadena coincidente.

En algunos trabajos existe una etapa de postprocesamiento donde vuelven a usarse técnicas de similitud léxica. Esta etapa persigue mejorar los mapeados, bien combinando elementos entre sí para lograr un término más general en SNOMED-CT que los englobe (Patrick y Budd, 2006; Patrick, Wang y Budd, 2007), o bien realizando comparaciones de subcadenas para aquellas expresiones para las que no se ha encontrado un equivalente satisfactorio (Wang et al., 2006).

Tras esta revisión podemos observar que las técnicas de similitud léxica han sido ampliamente usadas en los mapeados de terminologías en el dominio de la salud. Sin embargo, los trabajos estudiados presentan algunas carencias, como la falta de consideración semántica de las palabras. Por ejemplo, las expresiones "alteración nutricional por deficiencia" y "alteración nutricional por exceso", aunque son léxicamente parecidas, son opuestas en significado. Otra carencia existente es por la omisión de stopwords, algunas de las cuales pueden llegar a cambiar drásticamente la semántica y, por tanto, el resultado de la similitud. Por ejemplo, en una comparación léxica de las expresiones "alimento con gluten" y "alimento sin gluten", si no se tienen en cuenta las stopwords se estaría comparando la misma expresión: "alimento gluten". Es por ello, que la evaluación a nivel léxico no es suficiente, y es necesario aportar, al menos, una capa de semántica superficial.

3. Materiales

Para poder analizar el uso de conocimiento semántico superficial sobre métodos de similitud léxica, hemos creado una serie de materiales de trabajo. Con estos materiales se ha diseñado una batería de pruebas para comprobar la eficiencia de nuestra aproximación.

El primer material de trabajo está basado en la terminología NANDA-I en castellano. Esta terminología identifica una serie de diagnósticos de enfermería estandarizados para el cuidado de pacientes. Cada diagnóstico es una etiqueta que describe una situación y cuenta con un código propio que la identifica. El material de trabajo se creó a partir de la recopilación de las etiquetas de NANDA-I en castellano de las versiones 2001-2002, 2005-2006, 2007-2008 y 2009-2011. Con ello se recopilaron todas las descripciones y variantes léxicas existentes (variaciones de términos y signos de puntuación). La recopilación aportó un total de 728 etiquetas, de las cuales hemos utilizado seis para evaluar manualmente los resultados obtenidos en la experimentación.

El segundo material de trabajo consiste en un subconjunto de SNOMED-CT en castellano. SNOMED-CT es una terminología clínica, estructurada jerárquicamente, que proporciona contenido para informes y documentación clínica. En SNOMED-CT, cada concepto tiene un código único, una descripción completa (llamada *full*), una descripción preferente (*preferred*) y sus posibles sinónimos (*synonyms*). Cada concepto tiene una única representación terminológica (*full*), aunque incluye las diferentes etiquetas usadas en el dominio de salud para mencionar a dicho concepto (*synonyms*), incluso manteniendo aquellas que ya no están en uso. Por ejemplo, el concepto 85623003 tiene como descripción completa 'Termorregulación ineficaz (hallazgo)' y como descripción preferente tiene la etiqueta 'Termorregulación ineficaz'. Para nuestra experimentación se ha seleccionado un subconjunto de SNOMED-CT de su versión española del 30/04/2011, extrayéndolo a partir de los términos clave de los seis diagnósticos NANDA-I seleccionados. Este subconjunto está compuesto por 36 descripciones, que han sido evaluadas manualmente para establecer su equivalencia con las etiquetas NANDA-I. Tras la evaluación manual se estableció que solo nueve eran descripciones equivalentes a las etiquetas de NANDA-I.

Para las funciones de conocimiento semántico se crearon 3 recursos complementarios y con conocimiento ajustado a este dominio. Por una parte, se creó una bolsa con 68 antónimos y otra bolsa con 34 sinónimos; en ambos casos se incluyeron tanto sustantivos como adjetivos. Por otra parte, se añadió una 'expresión existencial'. Denominamos 'expresiones existenciales' a aquellas expresiones condicionales que modifican la semántica de toda la frase. No usamos recursos semánticos

de dominio abierto, como por ejemplo EuroWordNet, ya que, tras un estudio inicial, se detectó que no se ajustaban a las necesidades del dominio en estudio. Por ejemplo, las palabras 'exceso' y 'defecto', que en el dominio de enfermería son consideradas como antónimos, no lo son en ninguno de los recursos estudiados.

Finalmente, la experimentación se realizó usando 15 algoritmos de similitud léxica, mediante la implementación proporcionada por la librería de código abierto Java SimMetrics (versión 1.6.2) (Chapman, 2006). Los algoritmos usados han sido: (i) Coseno, (ii) Levenshtein, (iii) Similitud de Dice, (iv) Distancia Euclídea, (v) Similitud de Jaccard, (vi) Distancia Jaro-Winkler, (vii) Coeficiente de Matching, (viii) Needleman Wunch, (ix) Smith Waterman, (x) Coeficiente de Superposición, (xi) Monge Elkan, (xii) Distancia de Bloque, (xiii) Desviación de Distancia de Chapman, (xiv) Q Grams Distance y (xv) Soundex.

4. Método

En esta sección mostramos nuestra propuesta para el mapeado entre etiquetas de NANDA-I y SNOMED-CT para el idioma castellano. Para ello proponemos el uso de tres aproximaciones, complementarias al análisis de similitud léxica, para determinar la similitud entre etiquetas: (i) detección de antonimia, (ii) detección de sinonimia y (iii) detección de *'expresiones existenciales'*. Estas aproximaciones semánticas ayudan a establecer un grado de similitud entre dos etiquetas comparadas sin necesidad de entrenamiento ni corpus etiquetados. Las tres aproximaciones propuestas se han diseñado independientemente, de forma que pueden ser aplicadas de forma individual o combinándose entre ellas. El proceso se detalla en el Algoritmo 1.

Según se describe en el Algoritmo 1, el conocimiento semántico superficial se aplica en las funciones 'areRelated', 'areAntonyms' y 'getSynonyms', mientras que la función 'getSimilarity' devuelve el resultado de la comparación léxica de las etiquetas. El orden aplicado se corresponde a la mayor eficiencia de su ejecución en caso de que se haga uso de todas las funciones de conocimiento semántico superficial definidas. A continuación se detalla el funcionamiento de cada una de las funciones empleadas.

La función 'areRelated' se encarga de comprobar si hay expresiones en las etiquetas que establezcan su 'no equivalencia' semántica. Por ejemplo, si una de las etiquetas es "Riesgo de desequilibrio nutricional" y otra es "Desequilibrio nutricional", se establecen como 'no equivalentes', puesto que la primera expresa una probabilidad, mientras que la segunda expresa un hecho. En el caso de dos etiquetas con una 'expresión existencial' se consideran como posibles equivalentes cuando en ambas está presente dicha 'expresión existencial' o un sinónimo.

Algoritmo 1 Obtención de la similitud entre etiquetas, usando comparación léxica con conocimiento semántico superficial.

```
if areRelated(label1, label2) then
  if areAntonyms(label1, label2) then
    return 0;
  else
    {Expansión de sinónimos}
    labelSynons1=getSynonyms(label1);
    labelSynons2=getSynonyms(label2);
    {Para todas las expansiones}
    for all syn1 in labelSyns1 do
      for all syn2 in labelSyns2 do
        {Comparación léxica}
        tmpScore = getSimilarity(syn1,syn2);
        {Guardamos el mejor resultado}
        if tmpScore > maxScore then
          maxScore = tmpScore;
        end if
      end for
    end for
    return maxScore;
  end if
  return 0;
end if
```

La función 'areAntonyms' se encarga de comprobar la antonimia de las etiquetas. Por ejemplo, si una etiqueta es "Aumento de la fiebre" mientras que la otra es "Disminución de la fiebre", con significado antónimos, son establecidas como 'no equivalentes'.

La función 'getSynonyms', a partir de una etiqueta origen, devuelve un conjunto de etiquetas sinónimas generadas por la expansión de sus sinónimos. Por ejemplo, expandiendo la etiqueta "Aumento de la fiebre" obtendríamos como sinónima la nueva etiqueta "Incremento de la fiebre". Todas las expansiones de las etiquetas son usadas en la comparación léxica, buscando maximizar la

Prueba	Etiqueta NANDA-I	Etiqueta SNOMED-CT	Resultado ideal
I	1 etiqueta NANDA-I en uso	1 descripción completa y 1 preferente	Ambas son equivalentes a la etiqueta NANDA-I
II	1 etiqueta NANDA-I en desuso	1 descripción completa y 1 preferente	Ambas son equivalentes a la etiqueta NANDA-I
III	1 etiqueta NANDA-I en uso	2 descripciones completas, 2 preferentes y 4 sinónimas	Ninguna es equivalente
IV	1 etiqueta NANDA-I en uso	2 descripciones completas, 2 preferentes y 4 sinónimas	Ninguna es equivalente
V	1 etiqueta NANDA-I en uso	2 descripciones completas, 2 preferentes y 4 sinónimas	Sólo una descripción completa, una preferente y tres sinónimas son equivalentes a la etiqueta NANDA-I
VI	1 etiqueta NANDA-I en uso	2 descripciones completas, 2 preferentes y 4 sinónimas	Ninguna es equivalente

Tabla 1: Descripciones de las pruebas diseñadas y resultado ideal esperado.

Prueba	Etiqueta NANDA-I	Etiqueta SNOMED-CT	Singularidades
I	Termorregulación ineficaz	Termorregulación ineficaz	Iguales léxicamente
III	Desequilibrio nutricional: ingesta superior a las necesidades	Riesgo de desequilibrio nutricional: ingesta superior a las necesidades	Aunque tienen alta similitud léxica, la etiqueta SNOMED-CT está escrita como probable y la de NANDA-I como hecho
V	Riesgo de desequilibrio nutricional: ingesta superior a las necesidades	Riesgo de desequilibrio nutricional: ingesta inferior a las necesidades	Aunque tienen alta similitud léxica, son opuestas en significado
VI	Riesgo de infección	Trastorno nutricional: potencial de exceso para los requerimientos corporales	Léxicamente muy diferentes. Fue diseñado como prueba de control.

Tabla 2: Ejemplos de etiquetas usadas en los experimentos junto con sus singularidades.

puntuación obtenida.

Finalmente, la función 'getSimilarity' se encarga de evaluar la similitud léxica entre dos etiquetas. Esta función devuelve un valor entre 0 y 1. Estos valores representan el grado de similitud entre dos expresiones, lo cual consideraremos como el resultado de la comparación. Los resultados con un valor de 1 indican que las dos etiquetas son completamente equivalentes.

4.1. Diseño de los experimentos

Con el fin de comprobar la validez del método propuesto se diseñó un conjunto de seis pruebas. Puesto que nuestro objetivo es analizar la inclusión de conocimiento semántico superficial, como valores de referencia en la evaluación usamos los resultados de la comparación de similitud léxica simple (valores base). Para cada prueba se seleccionaron manualmente una etiqueta NANDA-I y varias de SNOMED-CT (descripciones completas, preferentes y sinónimas).

El esquema de cada prueba se encuentra indicado en la Tabla 1, donde se detallan las etiquetas incorporadas así como el resultado ideal esperado. Para completar la descripción, en la Tabla 2, se muestran algunos ejemplos de las etiquetas empleadas en las pruebas y las singularidades a superar por nuestra propuesta. Por ejemplo, en la prueba V (Tabla 2), las palabras subrayadas mues-

tran que las etiquetas comparadas, aunque léxicamente sean muy parecidas, no pueden considerarse equivalentes.

Por una parte, tras la ejecución del algoritmo propuesto, se realizó un análisis manual de los resultados para determinar si las etiquetas comparadas debían catalogarse como 'equivalentes' o 'no equivalentes'. Por otra parte, se analizaron los resultados respecto a los valores de los umbrales. La hipótesis de partida fue que si, dado un cierto umbral, el resultado obtenido es igual o mayor a él, entonces el algoritmo establecía como 'equivalentes' las etiquetas comparadas.

Posteriormente, se relacionaron ambos análisis y las etiquetas comparadas que eran equivalentes reales se consideraron 'verdaderos positivos' cuando el resultado alcanzaba o superaba el umbral analizado, o bien 'falso negativos' cuando no llegaba al umbral. De igual forma, cuando dos etiquetas comparadas no son en realidad equivalentes, se consideraron como 'falso positivo' cuando el resultado alcanzaba o superaba el umbral y, en caso contrario, se consideraba 'verdadero negativo'. Por ejemplo, al aplicar el algoritmo para comparar las etiquetas "Riesgo de infección" e "Infección Potencial", manualmente catalogadas como equivalentes, se obtiene como resultado 0.68, si se analizaba el resultado utilizando el umbral de 0.65, entonces se consideraba un resultado 'verdadero positivo', al ser equivalentes las etiquetas comparadas. Sin embargo, si se empleaba como umbral 0.7 el resultado sería considerado como un 'falso negativo', pues no sería identificada la comparación como 'equivalente'.

Los valores de los umbrales se fijaron entre 0.3 y 0.85 con un incremento de 0.05 para el análisis de los resultados. Esto se estableció acorde a un estudio piloto, donde se comprobó que los valores de los umbrales inferiores a 0.3 producían gran cantidad de falsos positivos (ruidos para los potenciales usuarios) y los umbrales superiores a 0.85 eran excesivamente restrictivos, obteniendo verdaderos positivos para aquellos casos en los que la etiqueta comparada tenía una escritura casi idéntica (silencio o perdida de información significativa para los usuarios).

5. Resultados

Para determinar la eficacia del conocimiento semántico superficial aplicado sobre los 15 algoritmos estudiados, se analizaron los resultados en base a la cobertura y la precisión, en tres ejecuciones diferentes: (i) sin utilizar conocimiento semántico, (ii) usando conocimiento parcial de antonimia y 'expresiones existenciales' y (iii) usando conocimiento semántico superficial completo. Los algoritmos de similitud léxica que mejor respondieron a la detección de etiquetas 'equivalentes' o 'no equivalentes' se muestran en las Tablas 3 y 4. En cada caso, al nombre del algoritmo se le ha añadido la letra 'C' cuando se ha complementado con conocimiento semántico superficial completo, la letra 'P' cuando se ha empleado conocimiento parcial y sin especificar letra para la aplicación del algoritmo básico. En ambas tablas, cada columna representa los valores del umbral utilizado, y los resultados obtenidos son mostrados como porcentajes, resaltando los mejores resultados para cada umbral.

De entre los 15 algoritmos estudiados, los mejores valores de cobertura fueron aportados por Monge Elkan (ME) y Soundex (Tabla 3). Con el algoritmo de ME los mejores resultados se obtuvieron con una aportación parcial de conocimiento semántico, y se mejoró la cobertura entre un 6.2 % (umbrales entre 0.3 a 0.45) y un 12.5 % (umbrales superiores a 0.5). Los resultados base del algoritmo ME solo mejoraron, con conocimiento completo, los resultados en los umbrales 0.55 y 0.8 (un 12.5 % superior). Por contra, el algoritmo de Soundex obtuvo mejores resultados usando el conocimiento semántico completo; su mejora osciló entre 12.5 % para los umbrales 0.6 y 0.65 hasta el 7.1 % para los umbrales 0.8 y 0.85. Con conocimiento parcial, Soundex solo mejoró la cobertura un 6.3 % a partir del umbral 0.7.

Respecto a la precisión, los algoritmos que aportaron los mejores resultados base fueron Jaro y Levenshtein, mientras que los mejores complementados con conocimiento semántico fueron los algoritmos Levenshtein y Q Grams Distance (QGD) (Tabla 4)[2]. Los algoritmos Levenshtein y QGD se comportaron de forma idéntica con conocimiento parcial o completo respecto a los resultados base. El algoritmo Levenshtein aporta una mejora significativa situada por encima del 50 % para

[2]Se incluyen los resultados de QGD sin conocimiento semántico para facilitar la comparación de la mejora aportada por el conocimiento semántico P y C.

Algoritmos	**Umbrales**											
	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85
Monge Elkan	93.8	93.8	93.8	93.8	87.5	75.0	75.0	75.0	75.0	75.0	62.5	62.5
Monge Elkan P	**100**	**100**	**100**	**100**	**100**	87.5	87.5	87.5	**87.5**	**87.5**	**75.0**	**75.0**
Monge Elkan C	93.8	93.8	93.8	93.8	87.5	87.5	75.0	75.0	75.0	75.0	**75.0**	62.5
Soundex	**100**	**100**	**100**	**100**	**100**	**100**	87.5	87.5	62.5	62.5	50.0	50.0
Soundex P	**100**	**100**	**100**	**100**	**100**	**100**	87.5	87.5	68.8	68.8	56.3	56.3
Soundex C	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	71.4	71.4	57.1	57.1

Tabla 3: Resultados de cobertura aplicando algoritmos de similitud básicos, con conocimiento semántico superficial parcial (P) y completo (C).

Algoritmos	**Umbrales**											
	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85
Jaro	34.9	34.9	34.9	35.7	32.5	34.2	40.0	42.9	42.1	28.6	**100**	**100**
Levenshtein	27.6	27.6	30.8	36.4	44.4	33.3	22.2	50.0	33.3	**100**	**100**	**100**
Levenshtein P	88.9	88.9	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Levenshtein C	88.9	88.9	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Q Grams Distance	35.7	35.7	41.7	29.4	20.0	16.7	22.2	25.0	25.0	33.3	**100**	**100**
Q Grams Distance P	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Q Grams Distance C	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

Tabla 4: Resultados de precisión aplicando algoritmos de similitud básicos, con conocimiento semántico superficial parcial (P) y completo (C).

los umbrales hasta 0.7 (mínima mejora del 50 % en el umbral 0.65 y máxima de 77.8 % en el umbral 0.6). El algoritmo QGD, aunque no obtuvo los mejores resultados base en precisión, la inclusión tanto de conocimiento parcial como completo mejoró los resultados entre un 58.3 % (umbral 0.4) y un 83.3 % (umbral 0.55), además de lograr una precisión del 100 % para los umbrales entre 0.3 y 0.75.

5.1. Discusión

La inclusión de conocimiento semántico superficial mejora significativamente la comparación léxica simple, llegando a mejorar hasta en un 12.5 % la cobertura y un 83.3 % la precisión, alcanzando porcentajes de cobertura del 87.5 % y precisión del 100 % para umbrales hasta el 0.75. Estos resultados suponen además una mejora significativa respecto al actual estado del arte, donde se reportan coberturas del 80 % y del 83 % para los trabajos de Wang et al. (2006) y Meizoso, Allones y Taboada (2011) respectivamente. En cuanto a precisión, estos dos mismos trabajos reportan precisiones del 50 % y del 95 % respectivamente. El resto de trabajos no aportan datos claros y completos respecto a su cobertura y precisión. Las principales mejoras aportadas por el método presentado frente a la comparación léxica exacta o el algoritmo del Coseno reside en que nuestra aproximación premia las subcadenas coincidentes y las similitudes parciales de cadenas, además de aportar comparación de sinonimia y restricciones semánticas.

Las penalizaciones impuestas por las restricciones de antonimia y 'expresiones existenciales' han prevenido emparejamientos incorrectos y, por lo tanto, mejorado significativamente los niveles de precisión respecto a los valores base (usar solo similitud léxica).

Sin embargo, estudiando detenidamente los resultados observamos que el uso de conocimiento semántico de sinónimos produce resultados dispares. Por un lado hace caer los resultados de cobertura con Monge Elkan, mientras en el caso del algoritmo de Soundex sirve para mejorarlos. En cuanto a los algoritmos de Levenshtein y Q Grams Distance no produce ningún cambio. Por otro lado, se detectó un efecto ruido en el experimento control diseñado. La sustitución de 'riesgo' por su sinónimo 'potencial' provocó un resultado superior en la compara-

ción y generó un falso positivo al superar los umbrales de análisis.

6. Conclusiones y trabajos futuros

Este artículo propone un método para el mapeado automático entre terminologías en el dominio de enfermería en castellano, combinándolo los métodos de similitud léxica clásicos con conocimiento semántico superficial en tres vías diferentes: sinónimos, antónimos y 'expresiones existenciales'.

Tras la experimentación hemos podido constatar que nuestra aproximación mejora los resultados de cobertura y de precisión respecto a métodos de similitud léxica, considerados como resultados base. Sin embargo, el uso de sinónimos en la expansión de etiquetas léxicamente muy diferentes puede producir ruido cuando comparten un término. Es por ello que esta aproximación debe ser mejorada para evitar mapeados erróneos.

Como trabajo futuro, para mejorar nuestra aproximación mediante el uso de sinónimos y prevenir falsos positivos, proponemos establecer un mínimo de términos en común entre dos etiquetas para poder ser comparadas. De esta forma, pretendemos evitar la comparación entre dos etiquetas no relacionadas. Este mínimo número de términos en común deberá ser establecido mediante experimentación, para establecer el margen más adecuado.

Bibliografía

Chapman, S. 2006. SimMetrics. Recuperado el 20 de Noviembre de 2011, desde http://sourceforge.net/projects/simmetrics.

Farfán Sedano, F. J., M. Terron Cuadrado, E. M. García Rebolledo, Y. Castellanos Clemente, P. Serrano Balazote y A. Gomez Delgado. 2009. Implementation of SNOMED CT to the medicines database of a general hospital. *Studies in Health Technology and Informatics*, 148:123–130.

International Health Terminology Standards Development Organisation -IHTSDO-. 2010. SNOMED Clinical Terms User Guide. Informe técnico, IHTSDO.

Meizoso, M., J. Allones y M. Taboada. 2011. Automated mapping of observation archetypes to SNOMED CT concepts. En *4th international conference on Interplay between natural and artificial computation, IWINAC 2011*, volumen 6686, páginas 550–561. Springer-Verlag, Berlin.

NANDA-I. 2010. *Diagnósticos Enfermeros: Definiciones y Clasificación, 2009-2011*. Elsevier, Barcelona.

Nyström, M., A. Vikström, G. H. Nilsson, H. Åhlfeldt y H. Örman. 2010. Enriching a primary health care version of ICD-10 using SNOMED CT mapping. *Journal of Biomedical Semantics*, 1:7. Doi:10.1186/2041-1480-1-7.

Patrick, J. y P. Budd. 2006. Automatic conversion of clinical notes into snomed ct at point of care. En J. Westbrook J. Callen G. Margelis y J. Warren, editores, *Proceedings of HIC2006 and HINZ2006*, páginas 209–213. Health Informatics Society of Australia (Aotea Centre, New Zealand).

Patrick, J., Y. Wang y P. Budd. 2007. An automated system for conversion of clinical notes into SNOMED clinical terminology. En *Proceedings of the fifth Australasian symposium on ACSW frontiers*, volumen 68, páginas 219–226. Australian Computer Society (Ballarat).

Romá-Ferri, M. T. y M. Palomar. 2008. Análisis de terminologías de salud para su utilización como ontologías computacionales en los sistemas de información clínicos. *Gaceta Sanitaria*, 22(5):421–433.

Stenzhorn, H., E. J. Pacheco, P. Nohama y S. Schulz. 2009. Automatic Mapping of Clinical Documentation to SNOMED CT. *Studies in health technology and informatics*, 150:228–232.

Wang Y., J. Patrick, G. Miller y J. O'Halloran. 2006. Linguistic mapping of Terminologies to SNOMED CT. En *Proceedings of Semantic Mining Conference on SNOMED*. Network of Excellence Semantic Mining (Copenhagen).

Zwaanswijk, M., R. A. Verheij, F. J. Wiesman y R. D. Friele. 2011. Benefits and problems of electronic information exchange as perceived by health care professionals: an interview study. *BMC Health Services Research*, 11:256. Doi:10.1186/1472-6963-11-256.

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 83-90 recibido 05-05-12 revisado 31-05-12 aceptado 05-06-12

Hypernymy relations from definiens-definiendum co-occurrence in multiple dictionary definitions*

Relaciones de hiperonimia a partir de la coocurrencia definiens-definiendum en múltiples definiciones de diccionario

Irene Renau **Rogelio Nazar**
University Institute for Applied Linguistics
Universitat Pompeu Fabra
Roc Boronat 138
08018 Barcelona, Spain
{irene.renau;rogelio.nazar}@upf.edu

Resumen: Presentamos una metodología basada en estadísticas de coocurrencia entre *definiens* y *definiendum* con el fin de extraer relaciones hiperonímicas de un corpus lexicográfico, como parte de un proyecto más extenso dedicado a la creación de una ontología general de nombres aplicada al estudio de las relaciones predicado-argumento. La idea de la presente propuesta es hacer emerger las relaciones de hiperonimia mediante la combinación de distintas fuentes lexicográficas. Encontramos que los hiperónimos de una palabra son los que aparecen con más frecuencia en las definiciones de esa palabra en diccionarios y que, del mismo modo, sus hipónimos suelen ser los que contienen frecuentes menciones a esta palabra en sus definiciones. Esto crea una asociación estadística entre palabras y permite estructurar un vocabulario en forma de taxonomía. Resultados preliminares muestran una precisión de 71,57% en hiperónimos y de 67,97% en hipónimos.
Palabras clave: estadística de coocurrencia, extracción de taxonomías, lexicografía computacional, relaciones hiperonímicas

Abstract: We present a methodology based on co-occurrence statistics between headwords and words in their definitions in order to derive hypernymy relations from a lexicographic corpus, as part of a more extensive project devoted to the creation of a general purpose Spanish ontology of nouns and its application to the study of predicate-argument structures. The idea of the present proposal is to extract these semantic relations using a statistical technique that allows to combine diverse lexicographic resources. We find that hypernyms of a word are frequently used in its definitions and, similarly, its hyponyms usually are those which have frequent mentions to this word in their definitions. This creates a statistical association between words that allows for a taxonomic structuring of a vocabulary. Preliminary results show precision figures of of 71,57% in hypernyms and of 67,97% in hyponyms.
Keywords: co-occurrence statistics, computational lexicography, hypernymy relations, ontologías, taxonomy extraction

1 Introduction

This paper explores the possibility of using definiens-definiendum co-occurrence statistics to derive hypernymy relations from a lexicographic corpus. The idea of the present proposal is to try to extract hypernymy relations from a combination of diverse lexicographic resources using co-occurrence information and to propose a method to deal with problems related to polysemy, one of the most important challenges in ontology extraction methods.

It is a already an established idea that the meaning of a word can be deduced from the words related to it on the syntagmatic and paradigmatic axes (Harris, 1954; Sinclair, 2004, among others).

* This research was funded by project: "Agrupación semántica y relaciones lexicológicas en el diccionario", lead by J. DeCesaris (HUM2009-07588/FILO); APLE: "Procesos de actualización del léxico del español a partir de la prensa", 2010-2012, lead by M. T. Cabré (FFI2009-12188-C05-01/FILO). We would like to thank the anonymous reviewers for their comments and Chris Norrdin for proofreading.

ISSN 1135-5948

However, corpus based studies of predicate-argument relations must face the problem of retrieving the semantic information we cannot deduce directly from the context. Similarly as corpus linguists do when they lemmatize a corpus to obtain a more accurate count of the vocabulary frequency, in this type of distributional analysis of meaning we must classify the instantiated words in semantic classes. Consider, for example, the Spanish verb *calmar* 'to calm'. In the corpus, we find a large number of arguments in direct object position, such as *ansiedad, dolor, hambre, jóvenes, niños, sed, señora, temor*, 'anxiety, pain, hunger, young people, children, thirst, lady, fear', etc. It is easy to manually separate these nouns into two groups: a) human and b) feelings and appetites, and this is a necessary step for conducting the semantic analysis which will led us to conclude that the transitive structures of *calmar* are linked to two meanings: a) 'to make a person calm down' and b) 'to alleviate pain or necessity'. An ontology of nouns would be thus helpful for a faster and more accurate semantic analysis of corpus by replacing the arguments as they appear in corpus with their corresponding hypernyms (e.g., *a lady is a kind of human*, *anxiety is a kind of feeling*). It would be possible then to encode every particular instance of the verb occurring with each different kind of human or feeling to produce general patterns such as *to calm + human* or *to calm + feeling/appetite*. This operation would increase the power of generalization of corpus analysis by translating the myriads of possible arguments in more general and useful categories. This is why we need a wide coverage Spanish ontology or, more precisely, a device capable of producing hypernyms for any given input word instead of the static and limited ontological resources available such as the Spanish WordNet.

Using dictionaries to extract semantic information is a well-known methodology (Section 2), but the strategy of combining multiple lexicographic resources and treating them as a unified corpus, to our knowledge, is a novel approach. Using word frequencies in multiple dictionary definitions has been attempted before with the goal of extracting hypernyms (Nazar & Janssen, 2010). In this paper we elaborate on this simple idea and take it further to extract both hypernyms and hyponyms of a word by iteration of the algorithm that counts the frequency of co-occurrence between definiens and definiendum. In this approach, we reduce a group of dictionaries to a two-column matrix. One of the columns corresponds to headword entries and the other contains the vocabulary of all the definitions of such headword (including definitions for the different senses of the word as well as examples of use), in decreasing order of frequency. Thus, the interest lies in the study of the statistical association between words in both columns, without using linguistic or ontological knowledge and ignoring all other aspects of the definitions text such as syntactic parsing, which results in a very simple and yet robust and flexible mechanism to derive the taxonomy of a language on the fly.

Our experiment is part of a broader project devoted to the field of semantic analysis, one of the most important and challenging problems in NLP. The strategy presented here, as it is dictionary-driven and not corpus-driven, cannot be the only one used to conduct semantic analysis, but it can be part of a group of combined strategies. We are conducting extensive experiments on taxonomy induction using different methods of distributional semantics, of which the method described in the present paper is just a particular case. In Nazar & Renau (in press), we carry out an experiment with data from a corpus of general language text, and in Nazar & Renau (in preparation) we create clusters of words that have similar profiles of co-occurence and are, thus, semantically similar (in most cases sharing the same hypernym). As we will show in Section 5, the next step to follow is mixing these different experiments in order to improve the results. Finally, it is also important to emphasize that although our experiments are carried out in Spanish, there is nothing language specific in the approach and therefore it should be possible to replicate the experiment in other languages and with other lexicographic resources.

2 Related Work

Efforts in automatic taxonomy extraction have been reported for decades. First attempts involved the extraction of taxonomies from dictionaries, and were mainly based on the parsing of the definitions (Calzolari et al, 1973; Calzolari, 1977; Amsler, 1981;

Chodorow et al, 1985; Nakamura & Nagao, 1988; Wilks et al, 1989). These authors assume that, for instance, the first noun in the definition of a noun is in general the genus term, as in the example (1) for the definition of the English word *sedan* in the *Longman Dictionary of Contemporary English*, where the noun (*car*) will be treated as the hypernym of *sedan*:

(1) **Sedan**: a car that has four doors, seats for at least four people, and a trunk.

Of course, this reasoning will not be useful in slightly more complicated cases such as a definition of *truck* (2) in the same dictionary, where the first noun, "piece", cannot be considered the genus term:

(2) **Truck**: a simple piece of equipment on wheels used to move heavy objects.

According to Chodorow et al. (1985), these difficulties can be circumvented using what the authors call "empty heads", consisting of a closed list of non-content words that should be ignored (e.g., *piece of, variety of, type of*, and so on), which would result in the tuple *truck* IS-A *equipment*.

Exploiting dictionaries in this way has continued until the present (Gonçalo Oliveira et al, 2011). There is, however, another trend that started with the arrival of corpus linguistics, when authors started to explore the possibility of hypernymy extraction not from dictionaries but directly from corpora, applying lexico-syntactic patterns that explicitly convey hypernymy relations (Hearst, 1992). Thus, if a word W_1 is a hyponym of a word W_2, such patterns could be W_1 *is a kind of* W_2, W_1 *is a type of* W_2, W_1 *and other* W_2, among many others.

Strickingly, the vast majority of the work in taxonomy extraction has long disregarded the application of co-occurrence statistics or quantitative methods in general. Machine learning methods, which are quantitative by nature, have been applied for taxonomy extraction (Snow et al, 2006; Pantel & Pennacchiotti, 2006) but, again, only for the extraction of lexico-syntatic patterns *à la* Hearst. To our knowledge, no study takes the frequency of co-occurrence to increase the certainty of a given hyponym-hypernym pair. Naturally, if *sedan* appears multiple times in different instances of Hearst's patterns with the word *car*, then that should be taken as a clear indication the hypernymy relation is correct in that case.

Another aspect that authors in the field of taxonomy induction often ignore is the problem of polysemy, as pointed out by some authors (Guthrie et al, 1990; Klapaftis & Manandhar, 2010). If not addressed properly, polysemy can potentially harm the results of any taxonomy extraction attempt, because such taxonomy would fail to offer the inheritance and transitivity properties. Problems of polysemy are not limited to homographs. They can be of different natures, for instance, regular polysemy or systematic polysemy, among others (Pustejovsky, 1995; Agirre & Edmonds, 2006; Jezek & Hanks, 2010).

In parallel to advances in automatic taxonomy extraction, there are also relevant pieces of work related to manual development of taxonomies, as in the case of WordNet (Miller, 1995) and EuroWordNet (Vossen, 1998), probably the most widely known current hand-crafted project of this kind. In this tool, the general idea was to follow the criteria "is a" for connecting the hypernyms with its hyponyms. The reason why WordNet is not an adequate tool for our purposes is that it is based on the concept of "synsets" or "synonyms sets", which are groups of words connected by a common meaning. Every synset is connected to the others and all of them create the taxonomic tree. Thus, there is no lexical approach in this idea, but a kind of 'conceptual approach' not appropriate for the analysis of lexical units. If Spanish WordNet gives us a synset made from *delito, falta, fechoría, malhecho* 'crime, offense, misdeed, bad action' as equivalents of the English synset containing *misbehavior* and *misdeed*, it is based on a criterion than is too loose for lexical analysis, as lexical units as *delito* and *falta* cannot be considered exact synonyms. In other cases, the taxonomic hierarchy fails when selecting one specific lexical unit from the synset and trying to connect to other specific ones in the upper levels. In the case of *aguja* meaning 'stylus', we find that the English hypernym *device* has been translated into a synset with three hypernyms: *aparato, dispositivo, mecanismo*, but in Spanish this three words are very different from each other, and in the case of *aguja* 'stylus', it is probably closer to *dispositivo* and *mecanismo* than to *aparato*. All these problems derive from the synset-centric approach, which is not compatible

with our lexico-centric approach. Authors such as Palmer (1998) and Hanks & Pustejovsky (2005) have raised similar objections.

Necessities of lexical analysis can be probably better achieved through methods such as Corpus Pattern Analysis (Hanks, 2004), which offers a systematic way of manually analyzing lexico-syntactic patterns. Patterns are constituted by syntactic information related to the arguments of the analyzed verb, and these arguments are also semantically analyzed by linking them to semantic types taken from an ontology. Thus, for example, in the case of the verb *calmar*, mentioned in Section 1, it would be divided into two different patterns as the following: [[Human 1]] calmar a [[Human 2]] and [[Human — Eventuality]] calmar [[Emotion — Appetite]]. This methodology is an important guide to our approach based on the Theory of Norms and Exploitations (Renau & Nazar, 2011), which offers a systematic way of manually analyzing lexico-syntactic patterns.

3 Methods

As stated in the introduction, our approach to the problem of taxonomy extraction is based on statistics of co- occurrence between defined words and the words in their definitions. The first step of our methodology is to compile a large lexicographic corpus from the web (Section 3.1). This corpus is then converted into a two-column matrix that encodes the frequency of co-occurrence of the defined words and those used in their definitions. With this matrix, we can compute two operations for any given input word: first, we look-up the word in the definiendum column of the matrix and then retrieve the most frequent words used in its definitions (Section 3.2). Second, we proceed exactly the other way round: the input word is looked up this time in the definitions of other words (Section 3.3). With the first operation we retrieve hypernym candidates, and with the second we retrieve hyponym candidates and also improve the certainty in the selection of both kinds of candidates by eliminating words that appear at the same time as hyponyms and hypernyms candidates. With respect to the problem of polysemy in the assignment of hypernymy (Section 3.4), our approach is based on an iteration of the same process: for every particular hypernym candidate proposed for a given input word, the system retrieves all the other probable hyponyms. Then, by representing these relations in the form of a directed graph, we observe a natural clustering of the words according to their different senses.

3.1 Compilation of a lexicographic corpus

Experiments were conducted on a lexicographic corpus crawled from the web. With this corpus, we create an index that registers words that tend to appear in the definitions of other words. With this index, we derive conclusions such as the hypernymy structure of a vocabulary not from a single lexicographic authority but from the aggregation of a multiplicity of sources. This makes up for the fact that the downloaded corpus is rather noisy and that many of the definitions are not entirely satisfactory if taken in isolation. The aggregated material as a whole, in contrast, offers better certainty as a consequence of the cumulative effect of the definiens-definiendum co-occurrence.

3.2 The input word in the definienda

In this phase of the procedure, the analysis of co-occurrence of words means to obtain, for any given word, a list of the most frequent content words that occur in its definitions. Thus, for instance, in the case of the word *sedán* 'sedan' the most frequent word in all the definitions is *automóvil* 'car'. Similarly, the most frequent word in the definitions of *triquinosis* 'trichinosis' is *enfermedad* 'disease'. This is a very stable pattern, however it is not sufficient for the development of a full scale ontology, and that is why we still need to carry out further operations.

3.3 The input word in the definiens

In this phase, we take the result of the previous one and iterate it. For each hypernym candidate obtained, we analyze in which other definitions they appear (definitions of words other than the initial input word). This gives us the possibility to find not only hypernym candidates (which we represent by outgoing arrows) but also hyponym candidates (represented by incoming arrows). To use one of the previous examples, Figure 1 shows a graph in which the initial input word *sedán* is linked to the hypernym candidate

automóvil 'car'. This hypernym, in turn, has links (incoming arrows) from different words apart from *sedán*. These are other words that, similarly as *sedán*, also have *automóvil* 'car' as the most frequent word in the definitions, e.g. *taxi, camión, berlina, grúa* 'taxi, truck, berlin, wrecker'. Of course, not all of these links are correct. There are also elements which are parts of an automobile and are confused with hyponyms: *luz, luneta, intermitente* 'light, rear-view window, indicator light', etc. In the case of *sedán* itself, unsurprisingly it does not appear to have hyponyms (it is not the most frequent word used in the definitions of any other word). Finally, the graph also reflects the relationship between *automóvil* and a higher order hypernym *vehículo* 'vehicle', thus, a hierarchy of two levels. This new link is derived from *automóvil* by an iteration of the same process that started from *sedán*, i.e., *vehículo* 'vehicle' is the most frequent word in the definitions of *automóvil*.

We could continue to iterate the process, for instance with *vehículo*, but that would depend on the particular task at hand. Thus, the number of iterations of the process is an execution parameter.

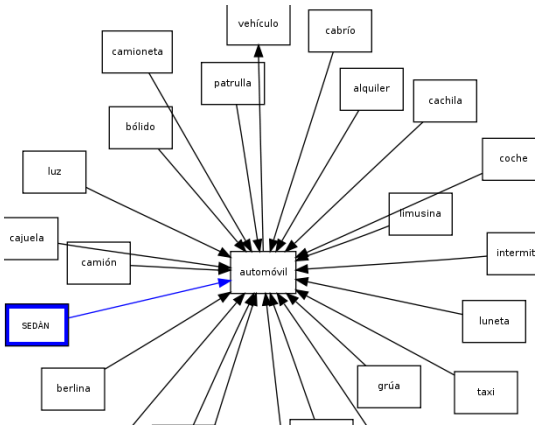


Figure 1: The noun *sedan* linked to *automóvil* as its hyponym and to other kinds of vehicles as co-hyponyms. At the same time, *automóvil* is linked to *vehículo* as its hyponym.

3.4 Word sense induction

Our attempt to solve the problems of polysemy by the use of word co-occurrence graphs is motivated by a distinctive geometric property of these graphs, which is to have attractors or hubs, defined as regions of the graph depicting nodes with large numbers of incoming arrows. In the case of word co-occurrence graphs, these hubs naturally represent the different senses of a word and can be of help in the process of disambiguation.

As we can see in Figure 2, for the case of a polysemous word such as *langosta* ('locust'/'lobster'), there are certain nouns in the network that have an important number of incoming links. Notice that different nouns co-occurring with them are clustered. These clusters represent the two different meanings of the word *langosta*, one for the land animal (*locust*) and the other one for the marine animal (*lobster*). Both groups are created around the hypernyms *insecto* ('insect') and *crustáceo* ('crustacean').

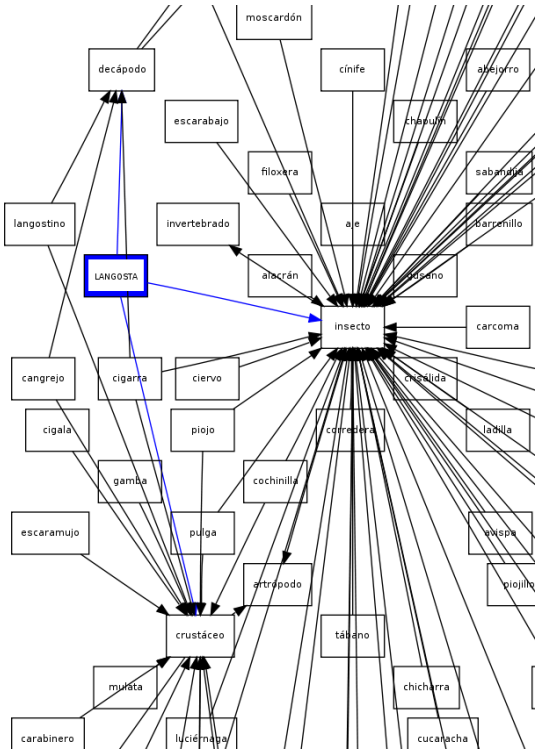


Figure 2: Fragment of the co-occurrence graph of the noun *langosta*, divided into two general meanings (clusters): *insecto* in the 'locust' sense and *decápodo* or *crustáceo* in the 'lobster' sense.

4 Results and Evaluation

For the evaluation of the system, we manually analyzed 173 nouns from the taxonomy, divided in the following groups:

a) 8 nouns typically used as hypernyms of many other nouns, in order to know how hyponyms were detected by the system: *calzado* 'footwear', *mueble* 'furniture', *embarcación* 'vessel', *queso* 'cheese', *herramienta* 'tool',

sombrero 'hat', *mineral* 'mineral', *utensilio* 'utensil'.

b) 15 polysemous nouns with at least two clear different meanings, in order to get information about how the hypernyms were detected (the English translation is given only for the most typical meaning): *águila* 'eagle', *manta* 'blanket', *aguja* 'needle', *mono* 'monkey', *araña* 'spider', *ratón* 'mouse', *dragón* 'dragon', *rémora* 'remora', *emperador* 'emperor', *sirena* 'siren', *fraile* 'monk', *tritón* 'newt', *gato* 'cat', *zapatero* 'cobbler', *langosta* 'lobster'.

For example, in the case of the word *águila*, we would expect to find the meanings 'eagle' and 'sharp person', being thus the hypernyms *ave* 'bird' and *persona* 'person', respectively. Similarly, in the case of *sirena* 'siren', we should see at least the meanings of 'piece of equipment' and 'aquatic nymph', etc.

c) 150 nouns randomly sampled from the dictionaries.

Despite the fact that groups *a* and *b* were created to focus on the problem of hypernymy and homonymy separately, both semantic relationships were evaluated in the three groups. The basic criterion used to evaluate the results is that nouns linked to a hypernymy relationship must strictly accept the test "is a" or "is a kind of". Thus, for instance in the case of *calzado* 'footwear', it could be possible to create a sentence like "A *bota* 'boot' is a kind of *calzado*", in which *bota* is a hyponym candidate.

The results of the overall evaluation are summarized in Table 1. The algorithm detected correctly 71.57% of the hypernyms and 67.97% of the hyponyms. For 42 nouns (24%) there were no results.

	Correct		**Incorrect**		**Total**
	n	%	n	%	
Hypernyms	214	71.57	85	28.43	299
Hyponyms	399	67.97	188	32.03	587
					886
No results:	42 (24 %)				

Table 1: Results of hypernyms and/or hyponyms candidates.

For group *a*, the algorithm detects many types of objects related to the hypernyms evaluated, that is, it detects *brie, cabrales, camembert, feta*, etc., as kinds of cheeses; *abanico, cuchara, grapadora, matamoscas* 'fan, spoon, stapler, flyswatter', etc., as kinds of utensils, and so on. In the case of group *b*, of polysemous words, co-occurrence graphs detect the most prototypical meaning in the majority of the cases. Table 2 shows the meanings detected for every polysemous word taken from the evaluation.

Nouns	**Hypernyms detected**
águila	2: *ave, moneda*
aguja	4: *barra, instrumento, pez, varilla*
araña	4: *arácnido, lámpara, planta, red animal, embarcación, pez, planta,*
dragón	6: *reptil, soldado*
emperador	4: *dignidad, pez, soberano, título*
fraile	2: *monje, montón (de uva)*
gato	2: *mamífero, persona*
langosta	3: *crustáceo, decápodo, insecto*
manta	2: *pieza, tela*
mono	1: *persona*
ratón	2: *dispositivo, mamífero*
rémora	1: *pez*
sirena	3: *aparato, instrumento, ninfa*
tritón	2: *anfibio, dios*
zapatero	4: *insecto, mueble, persona, pez*

Table 2: Number of hypernyms detected in 15 of the candidates (group *b*).

In order to estimate recall, we measured the coincidence of hypernyms and hyponyms of groups *a* and *b* with those provided by WordNet. To obtain this estimation we first had to manually check if the units provided by WordNet were indeed correct, and only then we compared those with the result of our algorithm. In total, WordNet provides 211 units (only in the first level and leaving aside multiword terms), but 49 of them are incorrect (e.g., we find nouns such as *diapasón* 'diapason', *remo* 'paddle' or *cepillo* 'brush' as hyponyms of *herramienta*, which is not exact because *herramienta* is translated from 'implement', but 'implement' can have other meanings as well, such as *utensilio*, probably closer to the mentioned hyponym candidates). For the remaining 162 units from WordNet, 55 are also in our taxonomy (34% recall). This figure, however, is only approximate because there are also units in our taxonomy that are correct and are not included in WordNet.

With respect to error analysis, the most important problems derive from the confusion between semantic relations: a) confusion hypernym-hyponym, e.g. *ratón* 'mouse' is taken as the hypernym of *múrido* 'murine', when in reality a mouse is a kind of murine and not vice-versa; b) confusion

hypernym/hyponym-synonym, especially in the case of slang variants, e.g. *minino* 'kitty' is put in the place of the hyponym of *gato* 'cat', when it is actually a colloquial synonym; confusion hypernym/hyponym-meronym/ holonym, e.g. *oro* 'gold' is taken by the hypernym of *águila* because the latter can be a kind of coin made of gold, thus the material of the coin is taken as if it was a hypernym. In other cases, the reason for the incorrect results are basically due to the lexicographic nature of the data, in which abbreviations, etymology or other informations are offered. Sometimes, the absence of data represents the lack of consensus in the different dictionaries, e.g. in the case of *rémora* taken as a 'hindrance', definitions were varied, and the concept was defined with hypernyms such as *cosa* 'thing', *obstáculo* 'obstacle' or *impedimento* 'impediment'.

5 Conclusions and Future Work

This paper has shown a language independent statistical method which uses a large number of online dictionaries as input to create a corpus-driven ontology with no human supervision. The results shown in the previous section give an approximate idea of the usefulness and limitations of the system. On the one hand, a 67-71 of precision seems to be sufficient as a basis of a ready-to-use tool to complement and facilitate the analysis by human expert. At the same time, the recall obtained seems to be appropriate for detecting the main meanings of a noun, that is, the meanings in which different lexicographical authorities agree upon, and in this sense the algorithm can be used as a reference for basic semantic analysis. On the other hand, many important meanings of some words were not detected, and often it is in these cases where the lexical analysis is most needed, and not in the main or more frequent meanings.

A general limitation of the strategy has already been pointed out in Section 1: as the sources used to run the experiment are dictionaries, the strategy lacks a real corpus-driven approach with direct contact with real data. In this sense, we are confident that the strategy presented in this paper can be combined with other strategies in order to create a ready-to-use taxonomy, having the advantage, in comparison with other dictionary-based methodologies, of representing the consensus and variety of information founded in diverse lexicographical resources.

Among the lines of future work, extensive evaluation of our word disambiguation strategy has to be carried out. We will also address the study of multiword expressions, which were not included in this paper. Another remaining factor is that our solution proposes for each word different hypernym candidates at each link of the chain, as in the example of *langosta* described in Section 3.4, where we had different hypernyms according to the 'locust' or 'lobster' senses. Strictly speaking, this is not a solution for the structural problem that polysemy represents for an ontology. In this taxonomy, we could start from a given word such as *tarántula* 'tarantula' and go up one level to a correct hypernym, such as *araña* 'spider', and then to an upper level hypernym such as *arácnido* 'arachnid'. However, being *araña* a polysemous word, we should not also end up in an incorrect hypernym such as *lámpara* (see Table 2). In a semasiological approach like ours, the solution for the problem of polysemy has to be tackled differently in every particular task at hand, that is, it has to be treated as a problem of disambiguation. This can include the computation of distributional similarity coefficients between the context of a particular instance of a word and those contexts of the different senses that the word can have. Again with the *langosta* example, our ontology should be able to assign a correct hypernym for a target word depending on the words that are found in the context, depending on whether they are related to insects or to lobsters.

References

Agirre, E. & Edmonds, P. (eds.) 2006. Word Sense Disambiguation. Algorithms and Applications. Dordrecht, Springer.

Amsler, R. 1981. A taxonomy for English nouns and verbs. Proc. of 19th annual meeting on ACL (Morristown, NJ, USA): 133–138.

Calzolari, N. 1977. An empirical approach to circularity in dictionary definitions. *Cahiers de Lexicologie* 31(2): 118–128.

Calzolari, N., Pecchia, L. & Zampolli, A. 1973. Working on the Italian machine dictionary: a semantic approach. Proc. of 5th Conference on Computational Linguistics (Morristown, NJ, USA): 49–52.

Chodorow, M., Byrd, R. & Heidorn, G. 1985. Extracting semantic hierarchies from a large on-line dictionary. Proc. of the 23rd annual meeting on ACL (Chicago, Illinois, USA): 299–304.

Guthrie, L., Slator, B., Wilks, Y. & Bruce, R. 1990. Is there content in empty heads? Proc. of the 13th International Conference on Computational Linguistics, COLING'90 (Helsinki, Finland): 138–143.

Gonçalo Oliveira, H., Antón Pérez, L., Costa, H. & Gomes, P. 2011. Uma rede léxico-semântica de grandes dimensões para o português, extraída a partir de dicionários electrónicos. *Linguamática* 3(2): 23–38.

Hanks, P. 2004. The Syntagmatics of Metaphor and Idiom. *International Journal of Lexicography* 17(3): 245–274.

Hanks, P. & Pustejovsky, J. 2005. A Pattern Dictionary for Natural Language Processing. *Revue française de linguistique appliquée* 10(2): 63–82.

Harris, Z. 1954. Distributional structure. *Word* 10(23): 146–162.

Hearst, M. 1992. Automatic acquisition of hyponyms from large text corpora. Proc. of the 14th International Conference on Computational Linguistics (Nantes, France): 539–545.

Jezek, E. & Hanks, P. 2010. What lexical sets tell us about conceptual categories. *Lexis* 4: 7–22.

Klapaftis, I. & Manandhar, S. 2010. Taxonomy Learning Using Word Sense Induction. Proc. of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL (Los Angeles, USA).

Miller, G. A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* 38(11): 39–41.

Nakamura, J. & Nagao, M. 1988. Extraction of Semantic Information from an Ordinary English Dictionary and its Evaluation. Proc. of the 12th International Conference on Computational Linguistics COLING-88 (Budapest, Hungary): 459-464.

Nazar, R. & Janssen, M. 2010. Combining Resources: Taxonomy Extraction from Multiple Dictionaries Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10): 1055–1061.

Nazar, R. & Renau, I. In press. A Co-occurrence Taxonomy from a General Language Corpus Proc. of EURALEX 2012. (Oslo, 7-11 August, 2012).

Nazar, R. & Renau, I. In preparation. Agrupación semántica de sustantivos basada en similitud distribucional. Implicaciones lexicográficas. V Congreso Internacional de Lexicografía Hispánica (Madrid, 25-27 June, 2012).

Palmer, M. 1998. Are WordNet sense distinctions appropriate for computational lexicons? SIGLEX-98, SENSEVAL (Herstmonceux, Sussex, UK, Sep 2-4, 1998).

Pantel, P. & Pennacchiotti, M. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. Proc. of 21st International Conference on Computational Linguistics and 44th annual meeting of the ACL (Sydney, Australia): 113–120.

Pustejovsky, J. 1995. The Generative Lexicon. Cambridge: MIT Press.

Renau, I.; Nazar, R. 2011. Propuesta metodológica para la creación automática de patrones léxicos usando el Corpus Pattern Analysis. Proc. of the 27th Conference of SEPLN (Huelva, September 5-7, 2011).

Sinclair, J. 2004. Trust the Text: Language, Corpus and Discourse Routledge.

Snow, R., Jurafsky, D. & Ng, A. 2006. Semantic taxonomy induction from heterogeneous evidence. Proc. of the 21st International Conference on Computational Linguistics (Sydney, Australia): 801–808.

Vossen, P. 1998. EuroWordNet: A Multilingual Database with Lexical Semantic Networks. *Computers and the Humanities* 32(2-3).

Wilks, Y., Fass, D., Guo, C., McDonald, J., Plate, T., Slator, B. 1989. A Tractable Machine Dictionary as a Resource for Computational Semantics. Computational Lexiography for Natural Language Processing. B. Boguraev and T. Briscoe (eds): 193-228. Essex, UK: Longman.

Reconocimiento y Síntesis del Habla

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 93-100 recibido 05-05-12 revisado 05-06-12 aceptado 08-06-12

A Simple Approach to Use Bilingual Information Sources for Word Alignment

Una manera sencilla para usar fuentes de información bilingüe para el alineamiento de palabras

Miquel Esplà-Gomis, Felipe Sánchez-Martínez, Mikel L. Forcada
Dep. de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant, Spain
{mespla,fsanchez,mlf}@dlsi.ua.es

Resumen: En este artículo se describe un método nuevo y sencillo para utilizar fuentes de información bilingüe para el alineamiento de palabras en segmentos de texto paralelos. Este método puede ser utilizado *al vuelo*, ya que no requiere de entrenamiento. Además, puede ser utilizado con corpus comparables. Hemos comparado los resultados de nuestro método con los obtenidos por la herramienta GIZA++, ampliamente utilizada para el alineamiento de palabras, obteniendo unos resultados bastante similares.
Palabras clave: Alineamiento de palabras, fuentes de información bilingüe

Abstract: In this paper we present a new and simple method for using sources of bilingual information for word alignment between parallel segments of text. This method can be used *on the fly*, since it does not need to be trained. In addition, it can also be applied on comparable corpora. We compare our method to the state-of-the-art tool GIZA++, widely used for word alignment, and we obtain very similar results.
Keywords: Word alignment, sources of bilingual information

1 Introduction

In this paper we describe a method which uses sources of bilingual information (SBI) such as lexicons, translation memories, or machine translation, to align the words of a segment with those in its translation (parallel segments) without any training process. Our approach aligns the sub-segments in a pair of segments S and T by using any SBI available, and then aligns the words in S and T by using a heuristic method which does not require the availability of a parallel corpus. It is worth noting that many SBIs which could be used to align words with our method are currently freely available in the Internet: MT systems, such as Apertium[1] or Google Translate;[2] bilingual dictionaries, such as Dics.info;[3] or Word Refference[4] or translation memories, such as Linguee[5] or MyMemory.[6] This method is inspired on a previous approach (Esplà-Gomis, Sánchez-Martínez, and Forcada, 2011) that was proposed to detect sub-segment alignments (SSAs) and help translators to edit the translation proposals produced by translation-memory-based computer-aided translation tools by suggesting the target words to change. A similar technique was also successfully applied to cross-lingual textual entailment detection (Esplà-Gomis, Sánchez-Martínez, and Forcada, 2012). Here, we propose to use these SSAs to obtain word alignment *on the fly*.

Related works. Many previous works tackle the problem of word alignment. The existing approaches may be divided in statistical approaches and heuristic approaches. One of the most remarkable works in the first group is the one by Brown et al. (1993), which describes a set of methods for word alignment based on the expectation-maximisation algorithm (Dempster, Laird, and Rubin, 1977), usually called *IBM models*. In this work, au-

[1] http://www.apertium.org
[2] http://translate.google.com
[3] http://www.dics.info
[4] http://www.wordreference.com
[5] http://www.linguee
[6] http://mymemory.translated.net

ISSN 1135-5948

thors propose five models, from a very simple one considering just one-to-one alignments between words, to more complex models which allow a word to be aligned with many words. Other authors (Vogel, Ney, and Tillmann, 1996; Dagan, Church, and Gale, 1993) propose using a hidden Markov model for word alignment. Both methods were combined and extended by Och and Ney (2003), who also developed the tool GIZA++, implementing all these methods.

Some heuristic approaches have also been proposed. Rapp (1999) proposes an approach based in the idea that groups of words which usually appear together in a language should also appear together in other languages. To obtain word alignments from this idea, the author uses a window of a given number of words to look for the most usual groups of words in each monolingual corpora. Then, coocurrence vectors are computed for the words appearing frequently together inside the window and word alignments are computed by comparing these coocurrence vectors. Fung and McKeown (1997) propose a similar method which introduces some SBIs. In this case, authors use bilingual dictionaries to obtain an initial alignment between *seed words* in a parallel text. To choose reliable seed words, they use only those words having a univocal translation in both directions and appearing with enough frequency to become useful references in both texts of the parallel corpus. Then, these initial alignments are used to align other words appearing around them in the parallel texts using a similar method to that used by Rapp (1999). Another family of heuristic methods for word alignment are based on cognates. Schulz et al. (2004) use word similarity between Spanish and Portuguese for word alignment. The most important limitation of this work is that it is only useful for closely-related languages. Other works (Al-Onaizan and Knight, 2002) try to overcome this problem by using transliteration to obtain the way in which a word in a language may be written in another language. In this case, Al-Onaizan and Knight (2002) use transliteration to find out the most likely way in which English proper nouns could be written in languages such as Arabic or Japanese in order to find their translations. Although statistical approaches have proved to obtain better results than heuristic ones, one of the advantages of heuristic approaches is that they can be used not only on parallel corpora, but in comparable corpora.

Novelty. In this work we propose a method for word alignment using previously existing bilingual resources. Although some works in the bibliography also use SBIs to perform alignment (Fung and McKeown, 1997), the main difference between this work and the previous approaches is that our method does not need any training process or bilingual corpus, i.e. it can be run *on the fly* on a pair of parallel segments. This kind of alignment method may be useful in some scenarios, as is the case of some computer-aided translation systems, to help users to detect which words should be post-edited in the translation proposals (Kranias and Samiotou, 2004; Esplà, Sánchez-Martínez, and Forcada, 2011). In addition, this method can be applied on comparable corpora to find partial alignments.

The paper is organized as follows: Section 2 describes the method used to collect the bilingual information and obtain the word alignment; Section 3 explains the experimental framework; Section 4 shows the results obtained for the different features combination proposed; finally, the paper ends with some concluding remarks.

2 Methodology

The method presented here uses the available sources of bilingual information (SBIs) to detect parallel sub-segments in a given pair of parallel text segments S and T written in different languages. Once sub-segments have been aligned, a simple heuristic method is used to extract the most likely word alignments from S to T and from T to S. Finally, both alignments are symmetrised to obtain the word alignments.

Sub-segment alignment. To obtain the sub-segment alignments, both segments S and T are segmented in all possible ways to obtain sub-segments of length $l \in [1, L]$, where L is a given maximum sub-segment length measured in words. Let σ be a sub-segment from S and τ a sub-segment from T. We consider that σ and τ are aligned if any of the available SBIs confirm that σ is a translation of τ, or vice versa.

Suppose the pair of parallel segments *S=Costarà temps solucionar el problema*, in Catalan, and *T=It will take time to solve the problem*, in English. We first obtain all the

possible sub-segments σ in S and τ in T and then use machine translation (MT) as a SBI by translating the sub-segments in both directions. We obtain the following set of SSAs:

temps	$\leftrightarrow$	*time*
problema	$\leftrightarrow$	*problem*
solucionar el	$\rightarrow$	*solve the*
solucionar el	$\leftarrow$	*to solve the*
el problema	$\leftrightarrow$	*the problem*

It is worth noting that multiple alignments for a sub-segment are possible, as in the case of the sub-segment *solucionar el* which is both aligned with *solve the* and *to solve the*. In those cases, all the sub-segment alignments available are used. Figure 1 shows a graphical representation of these alignments.

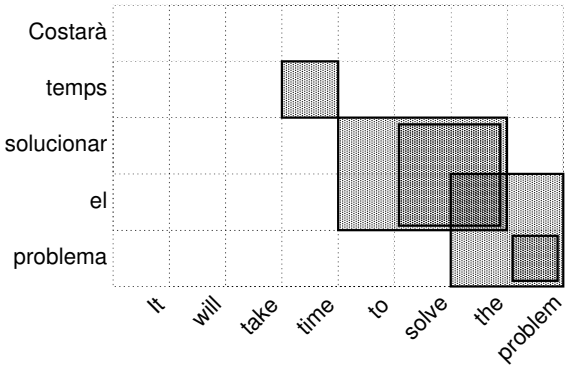


Figure 1: Sub-segment alignments.

Word alignment from sub-segment alignments. The information provided by the SSAs can then be used for word alignment. We define the *alignment strength* A_{jk} between the j-th word in S and the k-th word in T as

$$A_{jk}(S,T,M) = \sum_{(\sigma,\tau)\in M} \frac{\text{cover}(j,k,\sigma,\tau)}{|\sigma|\cdot|\tau|}$$

where M is the set of SSAs detected for the pair of parallel segments S and T, $|x|$ is the length of segment x measured in words, and $\text{cover}(j,k,\sigma,\tau)$ equals 1 if σ covers the j-th word in S and τ the k-th word in T, and 0 otherwise. This way of computing the alignment strengths is based on the idea that SSAs apply *alignment pressures* on the words; so the larger the surface covered by the SSA, the weaker the word-alignment strength obtained. Following our example, the alignment strengths for the words covered by the SSAs are presented in Figure 2. The words *temps* and *time* are only covered by a SSA (*temps*,*time*), so the surface is 1 and the alignment strength is $A_{1,4} = 1$. However, words *the* and *el* are covered by three SSAs: (*solucionar el*, *solve the*), (*solucionar el*, *to solve the*), and (*el problema*, *the problem*). So the alignment strength is $A_{3,6} = 1/4 + 1/6 + 1/4 = 2/3$.

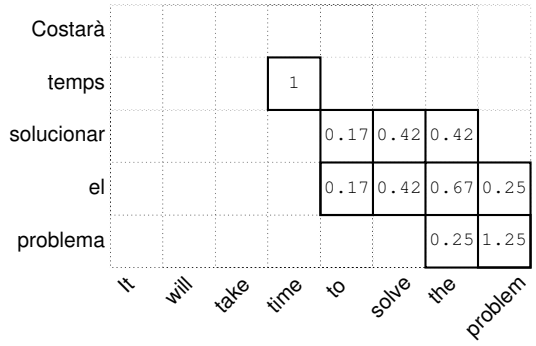


Figure 2: Alignment strengths.

The alignment strengths are then used to obtain word alignments. We simply align the j-th word in S with the k-th word in T if $A_{jk} > 0 \wedge A_{jk} \geq A_{jl}, \forall l \in [1, |T|]$, and vice versa. Note that one word in one of the segments can be aligned with multiple words in the other segment. Figures 3 and 4 show, respectively, the Catalan-to-English and the English-to-Catalan word alignments for the running example.

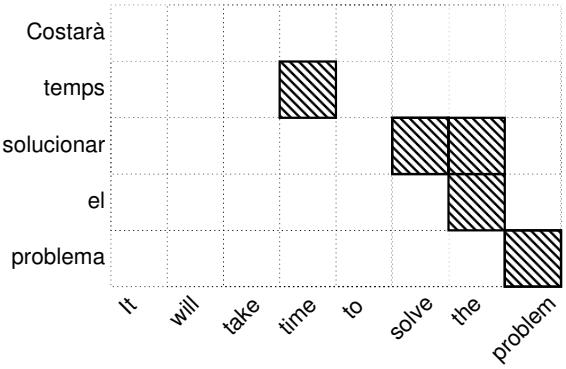


Figure 3: Resulting Catalan to English word alignment.

Figure 5 shows two possible symmetrised word alignments obtained by computing, in the first case, the intersection of the alignments shown in Figures 3 and 4, and, in the second case, the the widely-used *grow-diag-final-and* heuristic (Koehn, Och, and Marcu, 2003). It is worth noting that some words remain unaligned in Figure 5. This is a situation which can also be found in other state-of-the-art word alignment methods and, in this case, can be caused both by the symetrisation method, such as the word *to* in the alignment symetrysed through the intersection, or

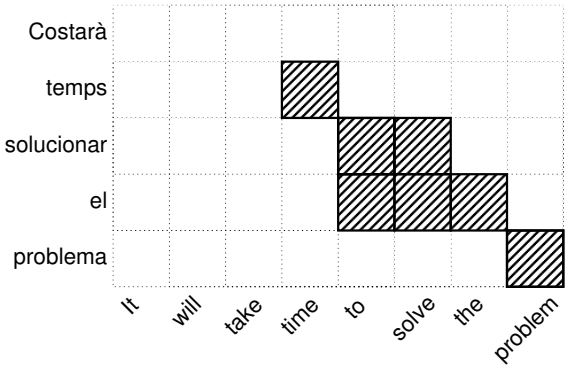


Figure 4: Resulting English to Catalan word alignment.

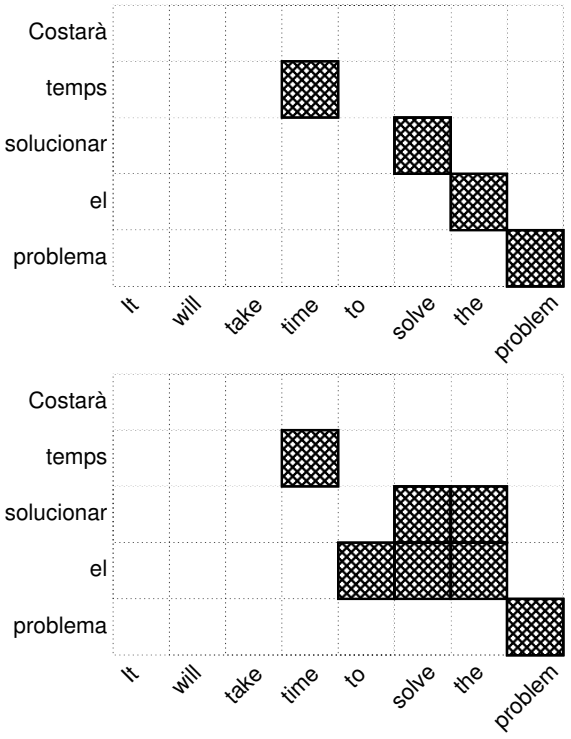


Figure 5: Two possible symmetrised word alignments, the first one using the intersection heuristic and the second one using the grow-diag-final-and heuristic.

by the lack of bilingual evidence relating the words, such as the words *Costarà*, *It*, *will*, and *take*. Depending on the needs of the task, more bilingual sources can be used in order to reduce the number of unaligned words. However, it is worth noting that unaligned words can also be caused by incorrect or excessively free translations, so keeping them unaligned may improve the overall alignment quality.

In addition, alignment strengths can be seen as a measure of the confidence on the relationships between the words. In future works, we plan to use the average alignment strength as a measure of the confidence on the SSAs. In this way, it could be possible to set a threshold to discard less-trusted SSAs. In the running example, the average alignment strength for the SSA (*solucionar el*, *to solve the*) is 0.37, whereas for the SSA (*el problema*, *the problem*) the average alignment strength is 0.60. Therefore, we see that (*el problema*, *the problem*) is a more reliable SSA than (*solucionar el*, *to solve the*).

3 Experimental setting

We evaluated the success of our system for word alignment using a *gold-standard* English–Spanish parallel corpus in which word alignments are annotated. We ran our method in both directions (Spanish to English and English to Spanish) and symmetrised the alignment obtained through the *grow-diag-final-and* heuristic (Koehn, Och, and Marcu, 2003) implemented in Moses (Koehn et al., 2007). We compared the performance of our system with that obtained by GIZA++ (Och and Ney, 2003) in different scenarios.

Test corpus. We used the test parallel corpus from the *tagged EPPS corpus* (Lambert et al., 2005) as a gold-standard parallel corpus.[7] It consists of 400 pairs of sentences from the English–Spanish Europarl (Koehn, 2005) parallel corpus and is provided with the corresponding gold-standard for word alignment. Two levels of confidence are defined for word alignments in this corpus, based on the judgement of the authors of the gold-standard: *sure* alignments and *possible* (less trusted) alignments.

Sources of bilingual information. We used three different MT systems as SBIs to translate the sub-segments from English into Spanish and vice versa:

- *Apertium*:[8] a free/open-source platform for the development of rule-based MT systems (Forcada et al., 2011). We used the English–Spanish MT system from the project's repository[9] (revision 34706).
- *Google Translate*:[10] an online MT system

[7] `http://gps-tsc.upc.es/veu/LR/epps_ensp_alignref.php3` [last visit: 2nd May 2012]

[8] `http://www.apertium.org` [last visit: 2nd May 2012]

[9] `https://apertium.svn.sourceforge.net/svnroot/apertium/trunk/apertium-en-es/` [last visit: 2nd May 2012]

[10] `http://translate.google.com` [last visit: 2nd May 2012]

by Google Inc (translations performed on 28th April 2012).

- *Microsoft Translator*:[11] an online MT system by Microsoft (translations performed on 27th April 2012).

Metrics. We computed the precision (P) and recall (R) for the alignments obtained both by our approach and by the baseline:

$$P = 100\% \cdot \frac{|\text{WA} \cap \text{GS}|}{|\text{WA}|}$$

$$R = 100\% \cdot \frac{|\text{WA} \cap \text{GS}|}{|\text{GS}|}$$

where WA is the set of alignments obtained and GS is the set of alignments in the gold standard. Then, we combined both measures to obtain the F-measure:

$$F = \frac{2 \cdot P \cdot R}{P + R}$$

These three metrics were computed, only for the sure alignments and also for both sure and possible alignments.

Baseline. We compared the performance of our word-alignment method to that of GIZA++ (Och and Ney, 2003), a toolkit for word alignment which implements different statistical alignment strategies. We run GIZA++ in both directions (source to target and target to source) and then we combine both sets of alignments through the *grow-diag-final-and* heuristic (Koehn, Och, and Marcu, 2003).

GIZA++ is widely used for word-alignment in statistical MT. In this scenario, it is usually trained on the parallel corpus to be aligned. However, it is also possible to use pre-trained models to align new pairs of segments, in order to avoid training a new alignment model for each new alignment task. As our system is aimed at performing word alignment on the fly, we consider that the most adequate scenario to compare our approach with GIZA++ is using pre-trained alignment models to align the test corpus. Therefore, for a better comparison of our method to state-of-the-art techniques, we define two baselines. In the first one, henceforth *basic-GIZA++ baseline*, we train and run GIZA++ on the test corpus. In the second one, henceforth *pre-trained-GIZA++ baseline*, we train GIZA++ on a larger parallel corpus and use the resulting models to align the test corpus. To train the alignment models for the pre-trained-GIZA++ baseline, we used the parallel corpus from the News Commentary corpus distributed for the machine translation task in the Workshop on Machine Translation 2011.[12] This corpus was lowercased, tokenized and cleaned to keep only those parallel segments containing up to 40 words. After this process, we obtained a corpus of 126,419 pairs of segments.

[11] `http://www.microsofttranslator.com` [last visit: 2nd May 2012]

segs.	sure			sure ∪ possible		
	P	**R**	**F**	**P**	**R**	**F**
100	57.1	59.9	58.5	64.7	47.4	54.7
200	57.5	61.2	59.3	64.9	47.5	54.9
300	59.7	63.6	61.6	67.8	50.1	57.7
400	59.9	64.2	62.0	68.2	50.5	58.0

Table 2: Precision (P), recall (R), and F-measure (F) obtained by the basic-GIZA++ baseline for sure alignments, and for all sure and possible alignments when aligning the gold-standard corpus in portions of 100, 200, 300, and 400 pairs of segments (segs).

4 Results and discussion

Table 1 shows the results obtained by our system and both baselines based on GIZA++: the basic-GIZA++ baseline and the pre-trained-GIZA++ baseline.

As can be seen, the method proposed in this paper obtains F-measures very similar to those obtained by both GIZA++-based baseline approaches. Another important detail is that our method obtains better precision in alignment than the two baselines proposed, although the results on recall obtained by the basic-GIZA++ baseline are better than ours.

Table 2 presents the results obtained by the basic-GIZA++ baseline when using portions of the test corpus with a different number of pairs of segments. The results presented in this table are useful to understand that, although the basic-GIZA++ yields slightly better results than the other approaches in Table 1, it clearly depends on the size of the parallel corpus to align. Of course, using this approach is not possible when trying to align a pair of segments on the fly, and obtains lower results when trying to align a very small set of parallel segments.

[12] `http://www.statmt.org/wmt11/translation-task.html`

Alignment kind	SBI-based approach			basic-GIZA++			pre-trained-GIZA++		
	P	R	F	P	R	F	P	R	F
sure	68.5%	57.6%	62.6%	59.9%	64.2%	62.0%	61.5%	55.8%	58.5%
sure ∪ possible	75.7%	43.9%	55.6%	68.2%	50.5%	58.0%	67.3%	42.2%	51.8%

Table 1: Precision (P), recall (R), and F-measure (F) obtained for the sure alignments, and also for all sure and possible alignments when aligning the gold-standard corpus. The results included correspond to our SBI-based approach and to both the basic-GIZA++ baseline and the pre-trained-GIZA++ baseline.

These results confirm that the approach proposed here can obtain alignments of a quality comparable to that obtained by the state-of-the-art GIZA++ tool, at least when trying to align small corpora, without needing any training process. These results set a bridge between the work of Esplà, Sánchez-Martínez, and Forcada (2011) and Esplà-Gomis, Sánchez-Martínez, and Forcada (2011), allowing to use SBI-based word alignment to help users to modify the translation proposals of a computer-aided translation system. It is worth noting that the weakness of our method is the recall, which may be improved by combining other SBIs.

5 Concluding remarks

In this work we have presented a new and simple approach for word alignment based on SBIs. This method can use any bilingual source of sub-sentential bilingual knowledge to align words in a pair of parallel segments on the fly. In this way, this process can be run without any training, which is useful in some scenarios, as is the case of computer-aided translation tools, in which word alignment can be used to guide translators when modifying the translation proposals (Kranias and Samiotou, 2004; Esplà, Sánchez-Martínez, and Forcada, 2011). In the experiments performed, our approach obtained results similar to those obtained by the state-of-the-art word-alignment GIZA++ tool. It is worth noting that the method proposed in this paper is a naïve approach which could be extended to obtain better results. Currently, we are evaluating new possibilities to improve the results obtained, such as using stemming or adding other SBIs available on-line.

In addition, we are developing a machine-learning-based approach which uses the ideas presented in this paper to perform word alignment in a more elaborate way, in order to improve the results obtained by the current approach. In this work we simply rely on the idea of *alignment pressures* to obtain the alignment strengths. However, it is possible to fit a maximum-entropy function, using a set of features obtained from the sub-segment alignments in order to obtain better alignment strengths. Although fitting the function would require a training process, once it is performed it could be applied to any new pair of segments on the fly. Another possible improvement may be to set weights for the different SBIs used for alignment, in order to promote those sources which are more reliable.

Acknowledgements: Work supported by the Spanish government through project TIN2009-14009-C02-01 and by Universitat d'Alacant through project GRE11-20. Google Translate service was provided by the *University Research Program for Google Translate*.

References

Al-Onaizan, Y. and K. Knight. 2002. Translating named entities using monolingual and bilingual resources. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 400–408, Philadelphia, Pennsylvania.

Brown, P.F., S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Dagan, I., K.W. Church, and W.A. Gale. 1993. Robust bilingual word alignment for machine aided translation. In *Proceedings of the Workshop on Very Large Corpora*, pages 1–8, Columbus, USA.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. volume 39 of *Series B*. Blackwell Publishing, pages 1–38.

Esplà, M., F. Sánchez-Martínez, and M.L. Forcada. 2011. Using word alignments to assist computer-aided translation users by marking which target-side words to change

or keep unedited. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, pages 81–88, Leuven, Belgium.

Esplà-Gomis, M., F. Sánchez-Martínez, and M.L. Forcada. 2011. Using machine translation in computer-aided translation to suggest the target-side words to change. In *Proceedings of the 13th Machine Translation Summit*, pages 172–179, Xiamen, China.

Esplà-Gomis, M., F. Sánchez-Martínez, and M.L. Forcada. 2012. UAlacant: using online machine translation for cross-lingual textual entailment. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 472–476, Montreal, Quebeq, Canada.

Forcada, M.L., M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Fung, P. and K. McKeown. 1997. Finding terminology translations from non-parallel corpora. pages 192–202.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*, pages 79–86, Phuket, Thailand.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 177–180, Prague, Czech Republic.

Koehn, P., F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Edmonton, Canada.

Kranias, L. and A. Samiotou. 2004. Automatic translation memory fuzzy match post-editing: A step beyond traditional TM/MT integration. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 331–334, Lisbon, Portugal.

Lambert, P., A. De Gispert, R. Banchs, and J. Mariño. 2005. Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation*, 39(4):267–285.

Och, F.J. and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Rapp, R. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526, College Park, USA.

Schulz, S., K. Markó, E. Sbrissia, P. Nohama, and U. Hahn. 2004. Cognate mapping: a heuristic strategy for the semi-supervised acquisition of a Spanish lexicon from a Portuguese seed lexicon. In *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland.

Vogel, S., H. Ney, and C. Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 836–841, Copenhagen, Denmark.

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 101-108 recibido 05-05-12 revisado 30-05-12 aceptado 05-06-12

Diseño y desarrollo de un sistema de evaluación automática de la pronunciación para el euskara

Design and development of an automatic pronunciation evaluation system for Basque

Igor Odriozola
UPV/EHU
Urkixo zum., z/g
igor@aholab.ehu.es

Oliver Jokisch
TU Dresden
Chair for Sys. Theory and Speech Tech.
Oliver.Jokisch@tu-dresden.de

Inma Hernáez
UPV/EHU
Urkixo zum., z/g
inma@aholab.ehu.es

Rüdiger Hoffmann
TU Dresden
Chair for Sys. Theory and Speech Tech.
Ruediger.Hoffmann@tu-dresden.de

Resumen: En este artículo, se presentan los primeros pasos en el desarrollo de un sistema de enseñanza de la pronunciación asistida por ordenador (CAPT, *Computer-Assisted Pronunciation Teaching*) para el euskara. El punto de partida es un sistema estándar de reconocimiento automático del habla (ASR) basado en modelos ocultos de Markov (HMM) que maneja parámetros de confianza GOP (*Goodness of Pronunciation*) para la verificación de fonemas. Dicho ASR se integrará en AzAR, el software de entrenamiento de la pronunciación desarrollado para el alemán y varias lenguas eslavas. En este artículo se presentan los primeros pasos del diseño del currículum para el euskara, los problemas generados en la verificación por el uso de HMMs creados a partir de una base de datos de ASR, y algunos resultados iniciales.
Palabras clave: evaluación de la pronunciación, verificación de fonemas, parámetros de confianza GOP, sistema AzAR.

Abstract: In this paper, the first steps of the development of a computer-assisted pronunciation teaching (CAPT) system for Basque are introduced. The baseline is a standard automatic speech recognition (ASR) system based on hidden Markov models (HMMs) that manages GOP (goodness of pronunciation) scores for phoneme verification. This ASR will be integrated into AzAR, the pronunciation training software developed for German and other Slavonic languages. This paper presents the initial steps in the design of the curriculum for Basque, some verification problems caused by the use of HMMs created from an ASR database, and some preliminary results.
Keywords: pronunciation evaluation, phoneme verification, GOP confidence scores, AzAR system.

1 Introducción

Para desarrollar un sistema de enseñanza de la pronunciación asistida por ordenador o CAPT (*Computer-Assisted Pronunciation Teaching*), es una tarea esencial adaptar los algoritmos utilizados para el reconocimiento automático del habla (ASR, *Automatic Speech Recognition*). En este artículo presentamos los primeros pasos del proceso de integración del euskara en el sistema AzAR (*Automat zur Akzentreduktion* – Autómata para la reducción del acento), el cual es el resultado de un proyecto desarrollado en el IAS (*Institut für Akustik und Sprachkommunikation*) de la Universidad Técnica de Dresden (Jokisch et al., 2005).

El sistema AzAR se diseñó inicialmente para entrenar y mejorar la pronunciación de estudiantes de la lengua alemana (L2) cuya lengua nativa (L1) fuera alguna de las pertenecientes al grupo de lenguas eslavas. Dentro del proyecto de cooperación *Euronounce* (Demenko et al., 2009), el sistema se amplió para el polaco, eslovaco, checo y ruso

ISSN 1135-5948

como L2. La base de datos *Euronounce* incluye lecciones especiales para entrenar la pronunciación de unos fonemas concretos, y se incluyen ciertas frases para utilizarlas como referencia en la práctica de la entonación o prosodia. De todas maneras, este último aspecto no tiene un *feedback* automático por parte del sistema. En siguientes desarrollos, el concepto *Euronounce* fue testado también con estudiantes de alemán de origen chino (Ding, Mixdorff y Jokisch, 2010).

El euskara es una lengua aislada que no pertenece al grupo de lenguas indoeuropeas, al contrario que cabría esperar si nos fijamos en su ubicación geográfica (Hualde, 1991). Convive con dos lenguas vecinas de gran potencial demográfico: el español y el francés. La influencia que ejercen esas dos lenguas sobre el euskara es importante, ya que el euskara se encuentra en situación de diglosia tanto en su parte norte (bajo la administración francesa) como en la sur (bajo la administración española), y tiene un estatus jurídico y grado de oficialidad heterogéneo, bastante complejo, dependiendo de la región.

Tal y como se ha mencionado anteriormente, las lenguas vecinas tienen gran influencia en el euskara, sobre todo para las personas que la estudian como L2. Además, *Euskaltzaindia*, la Real Academia de la Lengua Vasca, no ha tomado aún una decisión sobre cuál debe ser la entonación o prosodia para el euskara estándar, debido a la variedad de acentuaciones y entonaciones que muestra el euskara de un dialecto a otro. Eso hace que los estudiantes de euskara como L2 no tengan una referencia clara de la entonación que deben utilizar, lo que conlleva que muchas veces se decanten por la de su propio L1.

Por otro lado, a nivel fonético, el euskara posee ciertos fonemas que no existen en el inventario de las lenguas vecinas, y deben ser, por tanto, objeto de atención en el desarrollo del sistema de evaluación de la pronunciación de fonemas.

Por tanto, la primera conclusión clara que se ha obtenido es que en el diseño de un sistema CAPT para el euskara hay que tener en cuenta dos aspectos: el segmental o relativo a un sólo fonema (en este caso, la realización fonética o modo de pronunciación) y el suprasegmental o relativo a un grupo de fonemas (en este caso, el acento y la entonación).

2 Diseño del currículum para el euskara

Tal y como se ha explicado en el apartado anterior, se va a implementar una parte para la evaluación segmental y otra para la suprasegmental. Por tanto, el currículum de referencia para el euskara o colección de las características principales de un idioma que un hablante debe adquirir y que son objeto de entrenamiento debe comprender ejercicios para evaluar ambos aspectos.

Los primeros pasos para la evaluación de la parte suprasegmental, más concretamente la prosodia, fue implementada en AzAR bajo el proyecto *Euronounce*, pero el sistema se basaba solamente en la percepción auditiva del alumno, ya que el alumno debía comparar la entonación de una frase repetida con una de referencia. Por tanto, no había ninguna evaluación automática, es decir, ningún *feedback* por parte del sistema. Para el euskara, se ha considerado que la evaluación de la parte suprasegmental es esencial y, por tanto, se va a añadir un módulo de análisis de la prosodia que constará de una representación gráfica de las curvas de frecuencia fundamental (f0) y una puntuación obtenida de forma automática comparando la curva de f0 grabada por el alumno con la de referencia.

En esta parte se evaluarán dos características:

- La acentuación a nivel de palabra.
- La entonación a nivel de frase.

La acentuación y entonación que se han escogido como de referencia son las correspondientes al dialecto central del euskara, por ser aquélla la más estable en un área mayor. El acento característico a nivel de palabra de esta área, el cual no es exclusivo del dialecto central, sigue mayoritariamente un patrón acentual [+2, -1] para palabras aisladas de más de tres sílabas (Álvarez, 1986); es importante tener en cuenta que el euskara es una lengua aglutinante posposicional y que las marcas de los casos gramaticales se añaden al sintagma nominal en forma de sufijos; las palabras que llevan dichas marcas también siguen el mismo patrón.

Esta parte del currículum consta, en total, de 20 palabras aisladas y 50 frases.

En lo que se refiere a la parte segmental, las características más destacables del euskara, tanto fonéticas como fonológicas, son las siguientes:

- Características fonéticas:
  - El sistema vocálico y sistema de diptongos e hiatos.
  - Fonemas que no existen en L1 (español y francés), como, por ejemplo, el /*ts`*/ (ver ref. *Sampa Basque*).
  - Diferenciación entre las 6 sibilantes sordas del euskara: /*s`*/, /*s*/ y /*S*/ (fricativas) y /*ts`*/, /*ts*/ y /*tS*/ (africadas).

- Características fonológicas:
  - El proceso de palatalización de las consonantes /*l*/ y /*n*/ en el contexto /*i*CV/ (V es cualquier vocal).
  - El proceso de pérdida de sonoridad del primer fonema de la palabra siguiente a la partícula negativa *ez* (*ez dator* > /es`tatorr/) o desaparición de la sibilante de la partícula negativa (*ez nator* > /enatorr/).

Esta parte del currículum consta de 60 pares de palabras (contrastes) y 125 frases.

La voz de referencia fue grabada con un locutor nativo de euskara. Las señales se grabaron en formato digital de 16 bits a 16 kHz, en el estudio de grabación del IAS.

3 Pruebas iniciales y adaptación

3.1 El ASR base

Para la parte segmental, se ha utilizado el sistema de verificación de fonemas para el euskara desarrollado por el grupo Aholab (Odriozola et al., 2012). La verificación se realiza mediante un ASR basado en modelos ocultos de Markov (HMM, *Hidden Markov Models*), mediante el procedimiento de alineamiento forzado. De este modo, el sistema produce un parámetro GOP (*Goodness of Pronunciation*) para cada fonema, que se utiliza como medida de confianza.

Por lo general, los sistemas CAPT utilizan grabaciones de hablantes nativos vs. no-nativos para evaluar las señales grabadas por los estudiantes. Por tanto, una base de datos desarrollada específicamente para la verificación de fonemas debe tener grabaciones de los dos tipos de hablantes. Además, utilizando conocimiento previo, se puede prever en cierta forma cuáles son algunos de los fonemas más conflictivos para el estudiante de un L1 concreto, y, por tanto, las bases de datos se completan habitualmente con grabaciones donde aparecen dichos fonemas (Demenko, Wagner, y Cylwik, 2010).

El euskara es una lengua con recursos limitados que no tiene bases de datos acústicas suficientes para desarrollar tecnologías del habla que puedan competir con sus lenguas vecinas. Actualmente, hay tres bases de datos acústicas diseñadas para crear sistemas de reconocimiento de voz: la base de datos *SpeechDat_eu* (Hernáez et al., 2003), grabada sobre la red de telefonía fija a una frecuencia de muestreo de 8 kHz; *SpeechDat_eu_M*, similar para telefonía móvil, y una base de datos *Speecon-like*, grabada con varios tipos de micrófonos a varias distancias a una frecuencia de muestreo de 16 kHz, estas dos últimas creadas bajo la financiación del Gobierno Vasco y cedidas para su uso en investigación.

La base de datos *Speecon-like* contiene grabaciones de hablantes nativos y no-nativos, así como habla dialectal y estándar. Contiene señales de audio de 230 hablantes, grabadas en diferentes partes de Euskal Herria. En cada sesión, a cada hablante se le preguntaba por su nivel lingüístico, a elegir entre las opciones "nativo/a", "nivel alto" o "nivel bajo". El subcorpus de hablantes nativos está compuesto por 149 hablantes, el subcorpus de hablantes con nivel alto por 56 hablantes, y el de hablantes con nivel bajo por 25 hablantes.

3.2 Pruebas iniciales de reconocimiento de la voz de referencia

Los modelos ocultos de Markov o HMMs utilizados en reconocimiento automático del habla por el laboratorio están entrenados utilizando los hablantes de las primeras 155 sesiones (dos tercios), ya que el tercio restante se reservó para el test. Dichos modelos se crearon para fonemas con contexto (trifonemas), utilizando vectores de 39 parámetros MFCC (*Mel Frequency Cepstral Coefficients*, coeficientes cepstrales en las frecuencias de Mel), coeficientes basados en la percepción auditiva humana. Con esos modelos se realizó una prueba de reconocimiento preliminar sobre la voz de referencia, para tener una primera idea de la validez de la voz de

referencia y una impresión preliminar sobre el funcionamiento del sistema de evaluación de la pronunciación.

Se realizaron dos pruebas con diferentes diccionarios y diferentes gramáticas para así tener una visión más global de los resultados:

a) En la primera prueba, se evaluaron todos los ficheros, tanto los creados para la parte segmental como para la suprasegmental, 255 en total. El diccionario del sistema se creó con el lexicón completo del currículum: 921 entradas léxicas (EL). La gramática utilizada en este caso era un bucle de palabras sin modelado de lenguaje, donde cada palabra sucede a la siguiente con la misma probabilidad. Entre dos palabras, cabía la posibilidad de existir un silencio, pero sin modelar la coarticulación entre palabras.

b) En la segunda prueba, se evaluaron sólo los ficheros (60 en total) que contenían pares de palabras (contrastes entre sibilantes), ya que la capacidad de discernir entre fonemas es más manifiesta en este tipo de grabaciones. Para el diccionario, se utilizaron las palabras correspondientes a dichos ficheros (118 palabras diferentes), y se diseñó una gramática simple que modelaba la secuencia de dos palabras con silencios opcionales.

Los índices de error de palabra (WER, *Word Error Rate*) obtenidos en las dos pruebas de reconocimiento pueden verse en la Tabla 1.

TEST	DICC.	Nº FICH.	WER
1	921 EL	255	64,49 %
2	118 EL	60	76,67 %

Tabla 1: Resultados de reconocimiento de la voz de referencia, con modelos globales.

A primera vista, ante el hecho de que los resultados en el test 2 no eran tan buenos como los esperados, se dedujo que los modelos que habían sido creados para reconocimiento no eran capaces de discernir bien entre sibilantes. Por tanto, se pensó en repetir las pruebas con nuevos modelos creados utilizando sólo las grabaciones de hablantes nativos de la zona este de Euskal Herria, ya que en la zona oeste el fonema /s`/ ha sido históricamente asimilado por el fonema /*s*/ (actualmente, los dos se pronuncian como /*s*/). En total, se utilizaron 76 hablantes para entrenar nuevos modelos acústicos, y, tras repetir las pruebas, los resultados pueden verse en la Tabla 2.

TEST	DICC.	Nº FICH.	WER
1	921 EL	255	65,01 %
2	118 EL	60	81,86 %

Tabla 2: Resultados de reconocimiento de la voz de ref., con modelos de hablantes nativos.

Aunque los resultados son mejores con los nuevos modelos, siguen siendo más bajos de lo esperado para tareas tan sencillas. Por tanto, se realizó una prueba de verificación para analizar cómo se comportan los modelos a la hora de discernir un fonema. Para ello, se obtuvieron las distribuciones de los parámetros GOP de cada fonema, en dos situaciones diferentes: cuando el fonema está correctamente pronunciado, y cuando el fonema no está correctamente pronunciado. Para obtener la distribución de los fonemas incorrectamente pronunciados, se realizó una simulación de pronunciación incorrecta, realizando cambios controlados en el diccionario; es decir, sustituyendo un fonema de una posición concreta de cada palabra por otro del mismo grupo fonético (vocales, plosivas, nasales, líquidas y sibilantes), de forma aleatoria.

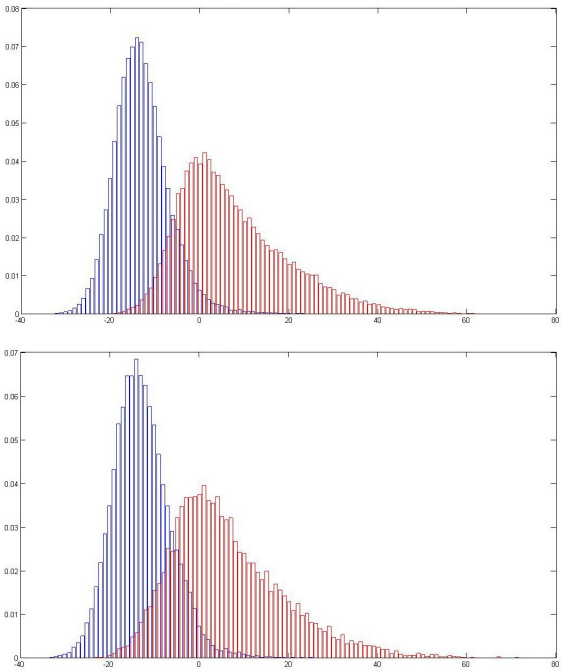


Figura 1: Pares de distribuciones (correctos: izda. vs. incorrectos: dcha.) de los GOP para el fonema /a/, con modelos globales del sistema de reconocimiento (arriba) y modelos creados con hablantes nativos (abajo).

El par de distribuciones (correcto - incorrecto) se calculó para cada fonema, con los modelos del sistema reconocimiento y con los modelos entrenados usando sólo las grabaciones de los hablantes nativos.

De este modo, se vio que, por ejemplo, los modelos acústicos del fonema /*a*/ dan como resultado dos pares de distribuciones casi idénticos (ver Figura 1), lo cual indica que las diferencias entre las realizaciones fonéticas de dicho fonema son casi iguales para los hablantes nativos de euskara y no-nativos de la base de datos. Además, los gráficos muestran pares de distribuciones parcialmente separados, con lo cual se deduce que, a priori, se puede establecer un umbral de decisión sin mucha dificultad y, por lo tanto, la detección de errores de pronunciación podría dar buenos resultados.

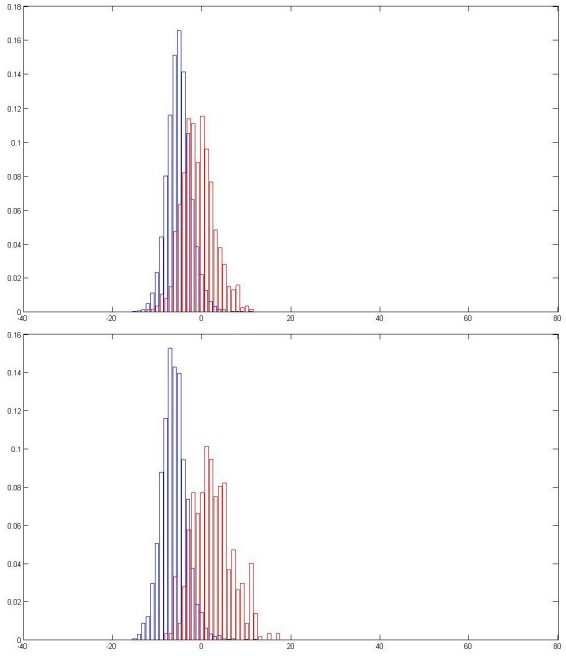


Figura 2: Pares de distribuciones (correctos: izda. vs. incorrectos: dcha.) de los GOP para el fonema /ts`/, con modelos globales del sistema de reconocimiento (arriba) y modelos creados con hablantes nativos (abajo).

Por otra parte, en cuanto al fonema /*ts`*/, el cual no existe en español, los pares de distribuciones experimentan un cambio para ambos grupos de modelos (ver Figura 2). Con los nuevos modelos, como cabía esperar, las distribuciones están más separadas, lo cual quiere decir que, en principio, se obtendrán mejores resultados en la verificación de fonemas.

Por último, se analizó la variación de los pares de distribuciones del fonema /*ts*/, el cual es problemático hoy en día en todo el territorio en donde se habla euskara, y se constató que los pares de distribuciones están muy solapados, tanto para los modelos de reconocimiento como para los creados para verificación (ver Figura 3). Este fonema se neutralizó a /ts`/ en la zona oeste de Euskal Herria, mientras que en la parte central y este parece que está siendo neutralizada actualmente a /*tS*/.

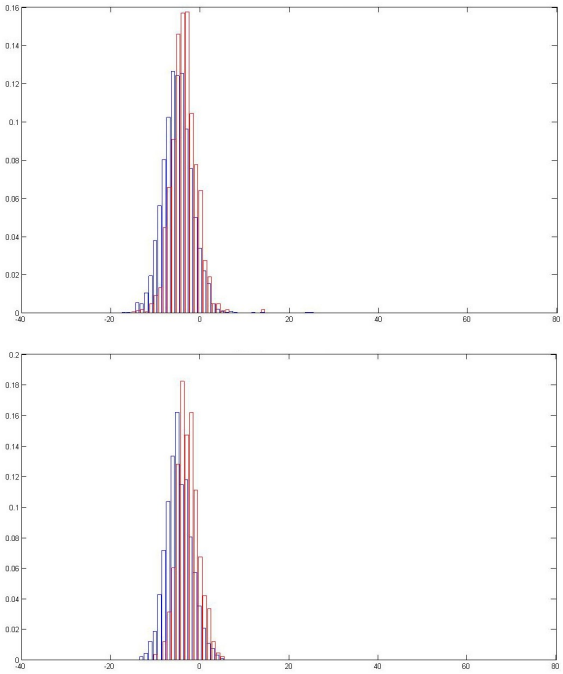


Figura 3: Pares de distribuciones (correctos: izda. vs. incorrectos: dcha.) de los GOP para el fonema /ts/, con modelos globales del sistema de reconocimiento (arriba) y modelos creados con hablantes nativos (abajo).

Para ver si realmente este modelo causa un empeoramiento de la tasa de error en nuestro sistema, se eliminó del currículum la parte donde se encontraba este fonema, y se repitieron las pruebas 1 y 2. Los resultados se muestran en la Tabla 3.

TEST	DICC.	Nº FICH.	WER
1	871 EL	235	65,70 %
2	100 EL	50	90,00 %

Tabla 3: Resultados de reconocimiento de la voz de ref., con modelos de hablantes nativos, y eliminando el fonema /ts/ del currículum.

Tras analizar las distribuciones del fonema /*ts*/, se concluye que hace falta entrenar mejor los modelos acústicos correspondientes a dicho fonema, ya que las señales que se disponen en la base de datos no tienen una buena correspondencia con sus transcripciones fonéticas, con pronunciaciones relativamente alejadas de las teóricamente esperadas.

3.3 Medida de la calidad de la segmentación de la voz de referencia

Las señales de audio se introdujeron en el sistema de reconocimiento y, por medio del procedimiento de alineamiento forzado, se obtuvo una segmentación a nivel de fonema. Para evaluar la calidad de la segmentación automática, los ficheros de audio se segmentaron también manualmente, a nivel de fonema.

La evaluación se realizó trama a trama. En total se analizaron un total de 584.612 tramas (todos los ficheros de la parte segmental), y se obtuvo un porcentaje de coincidencia de etiquetas del 88,08 %.

A primera vista, se observó que las mayores diferencias suceden en los finales de palabra, donde el sistema de reconocimiento necesita una cantidad de tramas mayor para salir del último HMM cuando éste no es el correspondiente al modelo de silencio. Esto es debido a ecos y reverberaciones que, aunque sean mínimas en las grabaciones de estudio, existen y tienen cierta presencia en la señal que el ser humano clasificaría como silencio.

Para solventar este problema, se pueden crear modelos de silencio más robustos. Sin embargo, se puede considerar que la segmentación de los fonemas es de muy buena calidad.

3.4 La entonación

Tal y como se ha explicado anteriormente, el análisis y evaluación de la prosodia se realizará a dos niveles: a nivel de palabra y a nivel de frase.

Como primer paso para la integración de la evaluación automática de la prosodia en el sistema AzAR, se ha creado una representación gráfica de las curvas de f0 de la señal de referencia y de la que el alumno ha grabado, alineadas y normalizadas. De esta forma, además de tener una referencia auditiva, el alumno tiene también una referencia visual que le permita evaluar mejor las diferencias entre las dos señales y aplicar así las correcciones necesarias.

4 Trabajo futuro

En la actualidad, el sistema de evaluación de la pronunciación para el euskara está en la fase de integración. La interfaz del software AzAR, desarrollada por el instituto IAE de la TU de Dresden, está siendo adaptada y modificada, y el sistema verificador de fonemas y de extracción de parámetros desarrollado por el grupo Aholab de la UPV/EHU está siendo integrado en el sistema AzAR.

Tras la integración, se realizarán varias pruebas de evaluación con estudiantes de euskara (L2) cuyo L1 es el español. En la actualidad se están diseñando los ejercicios y pruebas pertinentes para dicha tarea. La evaluación se realizará comparando los resultados obtenidos de forma automática con los proporcionados por un experto lingüista.

En publicaciones anteriores (Odriozola et al., 2012) se vio que la tasa de acierto (SA, *Scoring Accuracy*) del sistema, tanto en la evaluación de fonemas bien pronunciados como mal pronunciados, está alrededor del 80 %. Ese sistema proporciona sólo dos niveles a la salida, dependiendo de un umbral: si el parámetro de confianza GOP está por encima del umbral, se considera que el fonema está bien pronunciado; de lo contrario, está mal. Los sistemas actuales utilizan una salida con varios umbrales que permiten puntuar o evaluar una realización fonética con varios niveles de corrección, por lo general 3 ó 5. Este tipo de sistemas son más efectivos a la hora de interactuar con el usuario, ya que no sólo consideran que un fonema está "bien" o "mal" pronunciado (cosa que incluso para un humano puede llegar a ser una tarea compleja), sino que se tienen en cuenta graduaciones intermedias. El *feedback* suele darse con un sistema de colores donde, generalmente, el color verde corresponde a un fonema correctamente pronunciado, el color rojo a un fonema incorrectamente pronunciado, y colores intermedios para niveles intermedios. La impresión que un usuario puede llegar a percibir de un sistema puede mejorar notablemente si se implementa dicha característica.

5 Conclusiones

En este artículo se han presentado los primeros pasos para la integración del euskara en el sistema tutor de la pronunciación AzAR. Se ha concluido que, para el euskara, hacen falta dos tipos de análisis: la verificación de fonemas (para el entrenamiento de la parte segmental) y el análisis de la prosodia (para la parte suprasegmental).

La parte segmental está muy desarrollada en el sistema AzAR, pero necesita de modelos acústicos adecuados para cada lengua. En cuanto a la parte suprasegmental, se han dado los primeros pasos para implementar un sistema de evaluación automática de la prosodia en AzAR, ya que actualmente no cuenta aún con ningún *feedback* creado de forma automática. Esto constituye toda una línea de investigación que será desarrollada a corto plazo entre los dos grupos de investigación implicados en el proyecto del presente artículo.

Bibliografía

Alvarez, J.L. (Txillardegi). 1986. Proposamen bat azentuari buruz. *Euskera XXXI*, páginas 341–348 (en euskera).

Demenko, G., A. Wagner, N. Cylwik. 2010. The Use of Speech Technology in Foreign Language Pronunciation Training. *Archives of Acoustics*, 35(3), páginas 309–329.

Demenko, G., A. Wagner, N. Cylwik, O. Jokisch. 2009. An audiovisual feedback system for acquiring L2 pronunciation and L2 prosody. En *Proc. of 2nd ISCA Workshop on Speech and Language Technology in Education*, *SLaTE*, Wroxall Abbey Estate (Reino Unido).

Ding, H., H. Mixdorff, O. Jokisch. 2010. Pronunciation of German syllable codas of Mandarin Chinese speakers. En *Proc. Konferenz Elektronische Sprachsignal-verarbeitung*, *ESSV*, páginas 281–287, Berlin (Alemania).

Hernáez, I., I. Luengo, E. Navas, M. Zubizarreta, I. Gaminde, J. Sanchez. 2003. The Basque speech_dat (II) database: a description and first test recognition results. En *Proc. of Eurospeech-2003*, páginas 1549–1552. Ginebra (Suiza).

Hualde, J. I. 1991. *Basque Phonology*. Routledge, London & New York.

Jokish, O., U. Koloska, D. Hirschfeld y R. Hoffman. 2005. Pronunciation learning and foreign accent reduction by an audiovisual feedback system. En *Proc. of 1st Intern. Conf. on Affective Computing and Intelligent Interaction*, *ACII*, páginas 419–425, Pekín (China).

Odriozola, I., E. Navas, I. Hernaez, I. Sainz, I. Saratxaga, J. Sánchez, D. Erro. 2012. Using an ASR database to design a pronunciation evaluation system in Basque. En *Proc. of 8th Inter. Conf. on Language Resources and Evaluation*, *LREC*, Estambul (Turquía), páginas 4122–4126.

University of the Basque Country (UPV/EHU), Aholab Signal Processing Laboratory's website: http://aholab.ehu.es/sampa_basque.htm

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 109-116 recibido 05-05-12 revisado 04-06-12 aceptado 07-06-12

Técnicas de post-procesado de resultados en un sistema de diarización de locutores

Post-processing techniques for a speaker diarization system

David Tavarez UPV/EHU Alda. Urquijo s/n 48013 Bilbao david@aholab.ehu.es

Eva Navas UPV/EHU Alda. Urquijo s/n 48013 Bilbao eva@aholab.ehu.es

Daniel Erro UPV/EHU Alda. Urquijo s/n 48013 Bilbao derro@aholab.ehu.es

Ibon Saratxaga UPV/EHU Alda. Urquijo s/n 48013 Bilbao ibon@aholab.ehu.es

Inma Hernaez UPV/EHU Alda. Urquijo s/n 48013 Bilbao inma@aholab.ehu.es

Resumen: Este artículo presenta las técnicas de postprocesado diseñadas para mejorar los resultados de un sistema de diarización de locutores. Se han propuesto tres técnicas de mejora: el refinado de la segmentación voz/no voz, la asimilación de los segmentos cortos y la fusión de los clusters del mismo locutor. Las técnicas se han implementado en un módulo que se aplica como etapa de postprocesado y que ha mejorado un 22.3 % el resultado del sistema base. El módulo se ha aplicado sin realizar ningún ajuste sobre otro sistema de diarización de arquitectura similar al sistema base con una mejora del 21 % y sobre uno con arquitectura muy diferente sin conseguirse mejoras. Asimismo se ha utilizado con otra base de datos y se ha conseguido mejorar el DER un 17 %. Esto demuestra la validez de las técnicas desarrolladas para la mejora de los resultados de la diarización.
Palabras clave: Diarización de locutores, segmentación, transcripción enriquecida

Abstract: This paper presents the post-processing techniques designed to improve the results of a speaker diarization system. Three different techniques are proposed: refinement of speech vs. non speech segmentation, assimilation of short speech segments and fusion of clusters from the same speaker. These techniques have been implemented in a post-processing module that improves the result of the baseline system by 22.3 %. The same module has been applied to another speaker diarization system with a similar architecture to that of the baseline system with a DER improvement of 21 % and to another one with a very different architecture where no improvement has been achieved. It has also been used with another database with an improvement of 17 %. These experiments prove the validity of the techniques developed.
Keywords: Speaker diarization, segmentation, rich transcription

1. Introducción

La diarización de audio consiste en dividir un flujo de audio en regiones homogéneas de acuerdo a sus fuentes de audio específicas (Cettolo, Vescovi, y Rizzi, 2005). Estas fuentes pueden incluir tanto el tipo de audio (voz, música, ruido de fondo, etc.), como la identidad del locutor y las características del canal (Reynolds y Torres-Carrasquillo, 2005). La diarización de locutor se puede considerar un subtipo de diarización de audio que consiste en segmentar de forma automática una grabación de audio en regiones homogéneas relativas a la identidad del locutor, sin información a priori sobre la identidad y número de locutores presentes (Tranter y Reynolds, 2006). Adicionalmente, es posible llegar a identificar cada uno de ellos si se dispone de la información necesaria. Para cumplir este cometido deben combinarse varios algoritmos con diferentes finalidades que, en la mayoría de los sistemas, suelen ejecutarse de forma secuencial, es decir, cada uno se aplica a la señal completa antes de empezar con la tarea siguiente (Anguera, 2006). Comúnmente estas tareas incluyen la detección de voz, la detección de cambios de turno entre locutores, la agrupación de locutores y la resegmentación de la señal de audio (Anguera et al., 2012).

ISSN 1135-5948

Para determinar de manera objetiva la validez de los algoritmos desarrollados se organizan campañas competitivas de evaluación, como la NIST Rich Transcription[1] y la de Albayzin (Zelenák, Schulz, y Hernando, 2010). 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.

El sistema de diarización de locutores desarrollado en el grupo fue presentado a la campaña de evaluación Albayzin 2010 (Luengo et al., 2010). El algoritmo aplicado se basa en una implementación eficiente de un detector de cambio de turno basado en BIC (Chen y Gopalakrishnan, 1998) que utiliza únicamente los segmentos sonoros de voz y una agrupación de locutores off-line mediante proceso de agrupación o clustering jerárquico acumulativo de abajo arriba (Tavarez et al., 2012). Este sistema obtuvo buenos resultados en la evaluación, aunque carecía de etapa de resegmentación de la señal de audio. Para diseñar un módulo de post-procesado de los resultados que mejorara el rendimiento global del sistema base se ha llevado a cabo un detallado estudio de los errores cometidos y se han planteado distintas estrategias para paliar cada uno de ellos. En este artículo se presenta este módulo de post-procesado, describiendo las estrategias implementadas y los resultados obtenidos.

La sección 2 del artículo presenta brevemente la base de datos con la que se ha desarrollado el sistema. En la sección 3 se describe el análisis de los errores cometidos y en la sección 4 las técnicas de post-procesado propuestas para mejorar los resultados. La sección 5 detalla los experimentos realizados y los resultados obtenidos y finalmente en la sección 6 se exponen las conclusiones del trabajo.

2. Base de datos

En la campaña de evaluación Albayzin 2010 se proporcionó una base de datos de voz en catalán de programas de noticias emitidos por el canal de televisión 3/24 (Zelenák, Schulz, y Hernando, 2010). Fue grabada por el grupo de investigación TALP de la UPC y etiquetada por Verbio Technologies. Consta de un total de 24 grabaciones o sesiones en las que el número de locutores que interviene varía desde 30 hasta 250. La base de datos contiene unas 87 horas de audio con la siguiente distribución: 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, donde se engloba todo el material que no pertenece a las cuatro clases anteriores, incluyendo el ruido.

Para la campaña de evaluación la base de datos se dividió en dos partes: 16 sesiones para entrenamiento y desarrollo y las restantes 8 sesiones para pruebas.

3. Análisis de los errores

Los resultados obtenidos por el sistema base se han calculado de acuerdo con los criterios definidos por el NIST. La principal medida de error es el error total de diarización (DER, overall Diarization Error Rate) que está formado por la suma de los siguientes errores: voz de locutores no detectada (MST, Missed Speaker Time), segmentos de "no voz"marcados como voz de locutores (FAST, False Alarm Speaker Time) y errores de etiquetado incorrecto de locutores (SET, Speaker Error Time).

Con el fin de reducir el DER final, se ha llevado a cabo un análisis de los errores cometidos por el sistema base. Para ello, se han comparado las marcas de tiempo obtenidas con las marcas de referencia porporcionadas por la organización de Albayzin, para identificar la naturaleza de los distintos errores y diseñar estrategias que permitan tratar cada caso adecuadamente.

El DER obtenido por el sistema base en las señales de prueba es del 30.11 %, donde 2.8 % corresponde al MST, 2.2 % al FAST y 25.1 % al SET. Debido a la importancia del SET en los resultados obtenidos, se ha estudiado en detalle la aparición de errores que contribuyen en este aspecto, como son:

- segmentos cortos de un locutor que el proceso de agrupación de locutores asigna a otro diferente cuando la segmentación detecta un cambio que en realidad no existe.
- segmentos correspondientes a locutores con intervenciones de corta duración que el proceso de agrupación de locutores asigna a locutores ya existentes en lugar de crear un nuevo cluster.

[1] http://www.itl.nist.gov/iad/mig/tests/rt/

- distintas intervenciones de un mismo locutor que el proceso de clustering asigna a diferentes clusters, aumentando de forma errónea el número total de locutores identificados.

Tanto el MST como el FAST se deben a un mal funcionamiento del módulo de detección de voz. Debido a este incorrecto funcionamiento, segmentos de música que deberían ser eliminados pasan al módulo de detección de cambio de locutor y de agrupación de locutores y generalmente se les asigna una etiqueta de locutor nueva.

4. Módulo de postprocesado

Una vez analizados los errores cometidos por el sistema de diarización base se ha desarrollado un módulo de postprocesado con el fin de tratar cada uno de forma adecuada y reducir el DER final. A continuación se describe cada una de las partes que componen dicho módulo y que como puede verse en la Figura 1 consisten en un refinado de la segmentación voz/no voz, seguida de una etapa de asimilación de los segmentos de corta duración y por último de una fusión de los clusters correspondientes a un mismo locutor.

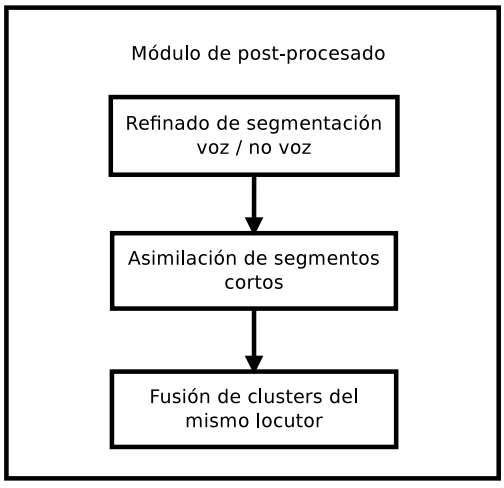


Figura 1: Diagrama del módulo de post-procesado

4.1. Etapa 1: Refinado de la segmentación voz/no voz

La primera etapa del módulo de postprocesado tiene como objetivo el refinado de los errores cometidos en el bloque de detección de voz. Para ello, en primer lugar, se entrena un modelo GMM (Modelo de Mezclas Gaussianas) para cada uno de los clusters obtenidos a la salida del sistema base y un modelo GMM para el silencio a partir de las marcas de referencia de las sesiones de entrenamiento designadas como "otros". A continuación, para cada segmento de voz marcado por el sistema base, se realiza una segmentación de Viterbi que incluye únicamente dos modelos GMM, el entrenado para el silencio y el del locutor marcado originalmente en dicho segmento. Por último, los silencios con una duración inferior a 750ms son eliminados.

De esta forma silencios, música, ruidos y el resto de posibles eventos acústicos detectados se marcan como "no voz", por lo que se consigue una reducción del FAST. Del mismo modo, intervenciones de otros locutores que han sido incluidas de forma errónea en cada uno de los diferentes clusters pueden ser marcadas en esta etapa como "no voz". Esto puede provocar un aumento del MST, sin embargo, también consigue aumentar la pureza de los clusters y con ello, mejorar el funcionamiento de las etapas posteriores.

4.2. Etapa 2: Asimilación de segmentos cortos

La segunda etapa del módulo de postprocesado tiene como objetivo eliminar los segmentos cortos marcados de forma errónea cuando se produce una intervención de larga duración de uno de los locutores. Para ello, en primer lugar, se localizan los segmentos sospechosos de estar erróneamente marcados en función de su duración y la del segmento que le precede. Dichas duraciones se han establecido de forma manual para optimizar el funcionamiento de esta etapa en la parte de desarrollo. A continuación se entrena un modelo GMM (G_x) usando todos los datos disponibles para el locutor marcado originalmente en el segmento sospechoso excepto los correspondientes a ese segmento. Además se entrena el modelo GMM del locutor anterior al segmento sospechoso (G_a), usando todos los datos disponibles en la grabación para él. Finalmente, si el segmento sospechoso queda mejor modelado por G_a que por G_x, es asimilado al cluster del locutor adyacente.

4.3. Etapa 3: Fusión de clusters

La tercera etapa tiene como objetivo la fusión de clusters pertenecientes al mismo locutor. Para ello, en primer lugar se generan modelos

GMM para cada uno de los clusters identificados por el sistema base con 60 segundos del material contenido en ellos. A continuación, para cada uno de los diferentes clusters, se extrae un segmento (con 60 segundos de la información que no ha sido utilizado para realizar el modelo GMM) y se calcula la diferencia de verosimilitudes en dicho segmento para el modelo marcado originalmente y cada uno de los restantes. Valores bajos obtenidos en la diferencia indican que los clusters contienen información similar, por lo que en función de dichos valores y las diferencias relativas obtenidas para cada uno de los diferentes clusters podemos tomar la decisión de combinar dos o más clusters en uno solo. El umbral de decisión para determinar si dos clusters deben fusionarse o no se ha establecido empíricamente optimizando los resultados de esta etapa en la parte de desarrollo de la base de datos.

En la Figura 2 se pueden observar las diferencias obtenidas para el cluster 2, en escala logarítmica. En este caso, los clusters 25 y 41, obtienen valores muy bajos de diferencia en relación al resto de clusters, por lo que parece lógico asumir que los clusters 2, 25 y 41 contienen información del mismo locutor, y por lo tanto podemos fusionarlos en uno solo.

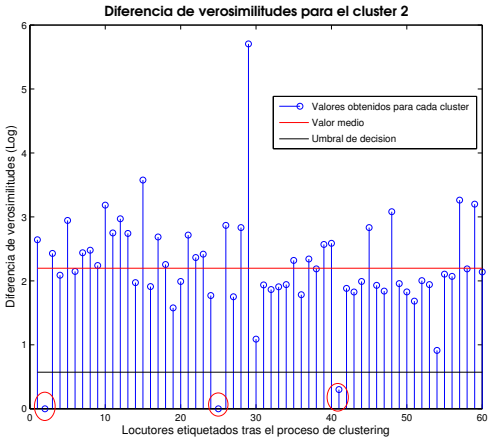


Figura 2: Diferencias de verosimilitudes obtenidas para el cluster 2

5. *Experimentos realizados*

Con el fin de probar el módulo de post-procesado diseñado se han realizado diversos experimentos. Primeramente se han aplicado las tres etapas a los resultados del sistema base, tanto en las sesiones de entrenamiento con las que se optimizaron los parámetros de configuración como a las sesiones de prueba. Posteriormente, y para comprobar la capacidad de generalización del módulo desarrollado se han hecho experimentos para aplicarlo a otros sistemas de diarización con diferentes arquitecturas y al mismo sistema base aplicado a una base de datos diferente.

5.1. Experimentos sobre el sistema base

En las tablas 1 y 2 se muestra el resultado obtenido después de aplicar el módulo de post-procesado a las marcas proporcionadas por el sistema de diarización base. En la Tabla 1 se recoge el DER obtenido para cada una de las sesiones de entrenamiento, tanto para las marcas originales del sistema base como a la salida de cada una de las etapas. Además se muestra en la última línea el valor de DER obtenido en la parte de entrenamiento de la base de datos.

S	DER	E1	E2	E3
1	22.17 %	21.83 %	21.54 %	29.49 %
2	24.58 %	24.46 %	24.38 %	13.10 %
3	23.10 %	23.01 %	22.92 %	18.11 %
4	27.47 %	27.67 %	27.50 %	27.50 %
5	14.15 %	12.94 %	12.93 %	9.89 %
6	21.22 %	21.40 %	21.32 %	16.21 %
7	24.84 %	24.86 %	24.89 %	27.72 %
8	27.26 %	27.38 %	27.38 %	19.90 %
9	28.92 %	28.28 %	28.60 %	26.80 %
10	34.75 %	34.54 %	35.26 %	26.80 %
11	27.94 %	27.70 %	27.90 %	15.91 %
12	27.42 %	27.22 %	27.22 %	25.54 %
13	31.92 %	32.13 %	31.86 %	32.34 %
14	41.16 %	40.87 %	41.00 %	25.84 %
15	32.50 %	32.73 %	32.62 %	27.25 %
16	32.06 %	32.09 %	32.02 %	24.18 %
All	28.25 %	28.14 %	28.16 %	23.33 %

Tabla 1: Resultado de las etapas de post-procesado en las sesiones de desarrollo

En la Tabla 2 se muestra el resultado obtenido para las sesiones de prueba. Al igual que en el caso anterior, se recogen los valores del DER tanto para las marcas originales del sistema base como a la salida de cada una de las etapas, así como el valor del DER obtenido en el conjunto de las sesiones de prueba.

Podemos observar en las tablas anteriores cómo se ha conseguido reducir el error en prácticamente la totalidad de las sesiones, tanto en la parte de entrenamiento como en

S	DER	E1	E2	E3
17	34.92 %	34.69 %	34.62 %	33.97 %
18	31.35 %	30.77 %	30.82 %	19.36 %
19	27.14 %	27.47 %	27.46 %	21.05 %
20	34.72 %	34.57 %	34.76 %	25.52 %
21	34.20 %	34.24 %	34.14 %	18.38 %
22	33.06 %	33.36 %	33.33 %	29.81 %
23	24.92 %	25.05 %	25.18 %	19.48 %
24	22.99 %	22.96 %	23.11 %	17.76 %
All	30.11 %	30.08 %	30.13 %	23.40 %

Tabla 2: Resultado de las etapas de post-procesado en las sesiones de prueba

la de prueba. Al aplicar el módulo de post-procesado se ha conseguido una reducción del DER del 17.4 % en las sesiones de entrenamiento y un 22.3 % en las sesiones de prueba, lo que prueba la validez de las etapas propuestas. En el caso de las dos primeras etapas, apenas se consigue reducción del error, sin embargo, permiten aumentar la pureza de los clusters y contribuyen a mejorar el funcionamiento de la tercera etapa.

5.2. Experimentos sobre otros sistemas de diarización

Una vez comprobado su buen funcionamiento para el sistema base y con el fin de comprobar si el módulo desarrollado puede ser de utilidad en otros sistemas de diarización, se han realizado experimentos para aplicarlo a sistemas de diarización con diferentes arquitecturas. Para ello se han utilizado las marcas proporcionadas por dos sistemas diferentes para la misma base de datos usada en Albayzin 2010. En todos los casos se han mantenido los parámetros de configuración obtenidos en las sesiones de entrenamiento para el sistema base. En primer lugar, se han utilizado las marcas proporcionadas por un sistema de arquitectura similar al sistema base, pero que trabaja de forma online (Luengo et al., 2010). La Tabla 3 recoge los resultados obtenidos al aplicar el módulo de post-procesado propuesto a este sistema.

S	DER	E1	E2	E3
E	26.77 %	26.72 %	26.76 %	21.38 %
P	27.17 %	27.18 %	27.32 %	21.45 %

Tabla 3: Resultado de las etapas de post-procesado sobre sistema online

Podemos observar cómo los resultados obtenidos han mejorado de forma similar a los del sistema de diarización base. El módulo de post-procesado elimina la característica de funcionamiento online de este sistema, pero reduce el DER en un 20.1 % en la parte de entrenamiento y un 21 % en la parte de prueba.

A continuación se ha aplicado el módulo de postprocesado a las marcas proporcionadas por un sistema de diarización desarrollado por el grupo GTM de la Universidad de Vigo (Docio, Lopez, y Garcia, 2010). Las distintas etapas que componen el módulo han sido diseñadas para corregir los errores observados en el sistema de diarización base. El objetivo de este experimento es comprobar el funcionamiento de dichas etapas en un sistema a priori diferente, cuyos errores no han sido analizados. Las tabla 4 recoge los resultados obtenidos.

S	DER	E1	E2	E3
E	25.48 %	25.54 %	25.31 %	25.91 %
T	25.62 %	25.62 %	25.26 %	27.00 %

Tabla 4: Resultado de las etapas de post-procesado sobre sistema GTM

Se puede observar cómo en este caso no se consigue mejora del error. Las dos primeras etapas diseñadas, de carácter menos específico, presentan unos resultados similares a los obtenidos en los casos anteriores, sin embargo la etapa de fusión de clusters alterna sesiones con mejora significativa y sesiones con mayor error. Esta etapa se ha diseñado analizando los errores cometidos por el sistema de diarización base y un sistema con una arquitectura diferente puede no compartir dichos errores, por lo que el resultado obtenido al aplicar esta etapa puede ser contrario al que se busca. Sin embargo, cabe recordar que en ningún momento se han modificado los parámetros de configuración del módulo de postprocesado, por lo que una optimización de los mismos podría conseguir una reducción significativa del error.

5.3. Experimentos sobre otra bases de datos

Por último, se ha propuesto estudiar la independencia del módulo de post-procesado de la base de datos utilizada, por lo que se ha utili-

zado el sistema de diarización base para marcar una pequeña base de datos creada a partir de señales proporcionadas por la Radiotelevisión Vasca (EiTB). Las señales corresponden a una colección de clips de noticias emitidas por EiTB en castellano y euskera durante el año 2010. En los ficheros de audio, además de la voz de los periodistas que narran las noticias y que aparecen repetidos en distintos ficheros, se incluyen también entrevistas y habla doblada sobre el audio original. Parte de estos clips de audio se ha concatenado para formar dos sesiones con diferentes características. La primera es una sesión de 20 minutos de duración en la que aparecen 9 locutores diferentes intercalados con largas intervenciones, en condiciones de bajo ruido. Esto favorece en principio el funcionamiento de las dos primeras etapas de módulo. La segunda es una sesión de 25 minutos en la que 40 locutores alternan intervenciones, que incluye segmentos en entornos con ruido y música de fondo, por lo que el funcionamiento de la tercera etapa debería tener mayor relevancia. Para establecer la referencia se ha llevado a cabo un marcado manual de las sesiones. El resultado obtenido se recoge en la tabla 5.

S	DER	E1	E2	E3
1	35.65 %	34.65 %	32.30 %	32.30 %
2	26.83 %	26.78 %	26.78 %	20.53 %
All	30.26 %	29.84 %	28.93 %	25.11 %

Tabla 5: Resultado de las etapas de postprocesado sobre el sistema base y la base de datos de EiTB

Podemos observar cómo los resultados obtenidos han mejorado de forma similar a los de la base de datos usada en Albayzin 2010. En este caso se ha obtenido una reducción del DER del 17 %. Con estos resultados se comprueba que el módulo desarrollado es válido para otras bases de datos de diferentes caracterśticas.

6. Conclusiones

Se han descrito diferentes técnicas de postprocesado diseñadas para mejorar los resultados de un sistema de diarización de locutores. Se han propuesto tres técnicas para tratar cada tipo de error cometido por dicho sistema: el refinado de la segmentación voz/no voz, la asimilación de los segmentos cortos y la fusión de los clusters del mismo locutor. Estas técnicas se han implementado y se han optimizado los parámetros de configuración utilizando la parte de desarrollo de la base de datos. Se ha dessarrollado un módulo de postprocesado para plicar las distintas técnicas al sistema de diarización base consiguiendo una mejora del 22.3 % en las sesiones de prueba. Con el fin de comprobar si el módulo desarrollado puede ser de utilidad en otros sistemas de diarización, se ha aplicado sin realizar ajustes sobre otro sistema de diarización de arquitectura similar al sistema base con una mejora del 21 % y sobre uno con arquitectura muy diferente sin conseguirse mejoras, aunque se plantea la posibilidad de obtener mejoras optimizado los parámetros de configuración. Por último, se ha comprobado que el módulo desarrollado es válido para otras bases de datos obteniendo una reducción del DER del 17 % utilizando las grabaciones de la Radiotelevisión Vasca (EiTB).

7. Agradecimientos

Los autores quieren agradecer a Iker Luengo el desarrollo del sistema base de diarización de locutores, al grupo GTM de la Universidad de Vigo el acceso a los resultados de su sistema de diarización y a la Radiotelevisión Vasca (EiTB) el uso de sus grabaciones.

Este trabajo ha sido financiado parcialmente por la UPV/EHU (Ayudas para la Formación de Personal Investigador), el Gobierno Vasco (proyecto BerbaTek, IE09-262) y el Ministerio de Ciencia e Innovación (Proyecto Buceador, TEC2009-14094-C04-02).

Bibliografía

Anguera, Xavier. 2006. *Robust Speaker Diarization for meetings.* Ph.D. tesis, Universitat Politecnica de Catalunya.

Anguera, Xavier, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, y Oriol Vinyals. 2012. Speaker Diarization: A Review of Recent Research. *IEEE Transactions on Audio, Speech and Language Processing*, 20(2):356–370.

Cettolo, Mauro, Michele Vescovi, y Romeo Rizzi. 2005. Evaluation of BIC-based algorithms for audio segmentation. *Computer Speech & Language*, 19(2):147–170, Abril.

Chen, S. S. y P. S. Gopalakrishnan. 1998. Speaker, environment and channel change detection and clustering via the bayesian information criterion. En *DARPA speech recognition workshop*, volumen 6, páginas 127–132.

Docio, L., P. Lopez, y C. Garcia. 2010. The uvigo-gtm speaker diarization system for the albayzin'10 evaluation. En *VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop, (FALA 2010)*, páginas 401–404, November.

Luengo, I., E. Navas, I. Saratxaga, I. Hernáez, y D. Erro. 2010. AhoLab Speaker Diarisation System for Albayzin 2010. En *FALA 2010*, páginas 393–396, Vigo.

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.

Tavarez, David, Eva Navas, Daniel Erro, y Ibon Saratxaga. 2012. Strategies to Improve a Speaker Diarisation Tool. En *LREC*, páginas 4117–4121, Estambul.

Tranter, S. E. y D. A. Reynolds. 2006. An overview of automatic speaker diarization systems. *IEEE Trans. on Audio, Speech and Laguage processing*, 14(5):1557–1565.

Zelenák, M., H. Schulz, y J. Hernando. 2010. Albayzin 2010 evaluation campaign: Speaker diarization. En *VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop*, páginas 301–304, Vigo, Spain, November.

Desarrollo de Recursos y Herramientas Lingüísticas

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 119-128 recibido 04-05-12 revisado 27-05-12 aceptado 05-06-12

Revisión de técnicas para la construcción de WordNets mediante estrategia de expansión*

A revision of techniques for WordNet construction following the expand model

Antoni Oliver, Salvador Climent, Marta Contreras
Universitat Oberta de Catalunya
Avda. Tibidabo 39-43 08035 Barcelona
aoliverg,scliment,mcontrerasf@uoc.edu

Resumen: Este artículo ofrece una revisión de métodos para la construcción de WordNets siguiendo la estrategia de expansión, es decir, mediante la traducción de las *variants* inglesas del Princeton WordNet. En el proceso de construcción se han utilizado recursos libres disponibles en Internet. El artículo presenta también los resultados de la evaluación de las técnicas en la construcción de los WordNets 3.0 para el castellano y catalán. Estas técnicas se pueden utilizar para la construcción de WordNets para otras lenguas.
Palabras clave: WordNet, recursos léxicos, semántica

Abstract: This paper presents a review of methods for building WordNets following the expand model, that is, by translating the English variants of the Princeton WordNet. Only free resources available online have been used. The paper also presents the evaluation of the techniques applied in the construction of Spanish and Catalan WordNets 3.0. These techniques can be also used for other languages.
Keywords: WordNet, lexical resources, semantics

1. Introducción

WordNet (Fellbaum, 1998) es una base de datos léxica del inglés donde las palabras pertenecientes a categorías abiertas (substantivos, verbos, adjetivos y adverbios) se organizan en conjuntos de sinónimos denominados *synsets*. El WordNet original se construyó para el inglés en la Universidad de Princeton (en el resto del artículo lo denominaremos PWN - *Princeton WordNet*). Siguiendo el modelo de PWN se han construido WordNets para muchas otras lenguas. Vossen (1998) distingue dos estrategias generales para la construcción de WordNets: (1) estrategia de combinación (*merge model*), en la que se genera una ontología propia para la lengua de llegada y posteriormente se generan las relaciones con PWN; y (2) estrategia de expansión (*expand model*) en la que se traducen las *variants* en inglés utilizando diversas estrategias. En esta segunda estrategia no es necesario establecer relaciones interlingüísticas.

La principal dificultad para la construcción de WordNets mediante la estrategia de expansión es la polisemia. Si todas las *variants* fuesen monosémicas (es decir, que estuviesen asignadas a un único *synset*) el problema sería simple, ya que únicamente tendríamos que encontrar una o más traducciones para la *variant* inglesa. Al tener la palabra inglesa un único sentido la traducción de ésta seria la *variant* correcta en la lengua de llegada.

En la tabla 1 podemos observar el número de *variants* que tienen asignadas un número determinado de *synsets*. Las *variants* que tienen asignado un único *synset*, son palabras monosémicas en inglés (al menos, según PWN). Así, el 82.32 % de las *variants* del PWN son monosémicas.

Otro aspecto interesante es obser-

* Este trabajo se ha llevado a cabo dentro del proyecto Know2 *Language understanding technologies for multilingual domain-oriented information access* (MICINN, TINN2009-14715-C04-04)

Núm. synsets	variants	%
1	123.228	82.32
2	15.577	10.41
3	5.027	3.36
4	2.199	1.47
5+	3.659	2.44

Tabla 1: Número de *variants* que tienen asignados un número determinado de *synsets*

var cuántas de estas *variants* monosémicas están escritas con alguna letra en mayúsculas (y corresponderán probablemente a nombres propios) y cuántas están en minúscula. En la tabla 2 podemos observar estos valores

	variants	%
minúscula	84.714	68.75
mayúscula	38.514	31.25

Tabla 2: Número de *variants* monosémicas del PWN según se encuentren en minúculas o con alguna letra en mayúsculas

Nuestro proyecto se enmarca en la creación de los WordNets 3.0 del español y catalán. Las versiones anteriores de los WordNets en castellano y catalán diferían en su licencia: mientras que el catalán disponía de una licencia libre, el castellano se distribuía con una licencia propietaria. Los WordNets 3.0, tanto para el castellano como para el catalán, se distribuirán bajo una licencia libre.

2. *La construcción de los WordNets del español y el catalán*

Tanto el WordNet castellano (Atserias et al., 1997) como el catalán (Benítez et al., 1998) se construyeron utilizando la estrategia de expansión y de una manera prácticamente idéntica. En primer lugar se desarrollaron manualmente una serie de conceptos considerados base. Estos conceptos son los considerados más importantes y que además son comunes en todas las lenguas involucradas en el proyecto EuroWordNet (Vossen, 1999).

Para desarrollar la parte correspondiente a los substativos de las primeras versiones de estos WordNets se utilizó una metodología basada en la consulta a diccionarios bilingües. Para llevar a cabo la evaluación de este método se consideraron dos tipos de relaciones: la relación palabra inglesa - *synset*, dada por el PWN; y la relación palabra castellana[1] - palabra inglesa, dada por los diccionarios bilingües utilizados.

Atendiendo a la relación palabra inglesa - *synset* se consideraron dos grupos:

- Las palabras inglesas monosémicas, es decir, las asignadas a un único *synset*.
- Las palabras inglesas polisémicas, es decir, las asignadas a más de un *synset*

Por otro lado, atendiendo a la relación palabra castellana - palabra inglesa, se distinguieron cuatro grupos:

- Grupo 1: palabras castellanas que tienen una única traducción a una palabra inglesa, y que a su vez esta palabra inglesa sólo tiene una traducción a la misma palabra castellana.
- Grupo 2: palabras castellanas que tienen más de una traducción a diversas palabras inglesas, pero todas estas palabras inglesas se traducen únicamente por dicha palabra castellana.
- Grupo 3: palabras castellanas que se traducen por una única palabra inglesa, pero esta palabra inglesa se traduce a más de una palabra castellana.
- Grupo 4: palabras castellanas que se traducen por más de una palabra inglesa, y a su vez, cada una de estas palabras inglesas se traducen a más de una palabra castellana.

Como resultado, se obtuvieron 8 grupos de conjuntos de tres elementos (palabra castellana - palabra inglesa -*synset*). La asignación de palabra castellana a *synset* se realizó directamente en todos los grupos, pero la división permitió realizar

[1] o catalana

una evaluación para cada grupo y asignar las precisiones obtenidas como *scores* a las *variants* castellanas.

La parte verbal se desarrolló siguiendo una metodología totalmente manual dada la gran polisemia verbal y el número relativamente pequeño de conceptos verbales. Los adjetivos y adverbios no se desarrollaron en las primeras versiones de los WordNets del español y del catalán.

3. Babelnet

El objetivo de Babelnet (Navigli y Ponzetto, 2010) es crear un red semántica de grandes dimensiones combinando el conocimiento lexicográfico de WordNet con el conocimiento enciclopédico de la Wikipedia. Así pues BabelNet ofrece una relación entre las entradas de la Wikipedia y los *synsets* de WordNet. A continuación podemos observar un fragmento del archivo que se distribuye con el proyecto BabelNet:

```
Adobe_brick adobe_brick%1:06:00::
02681392n
Fuselage fuselage%1:06:00:: 03408054n
Hearse hearse%1:06:00:: 03506880n
Merida_(Yucatan) merida%1:15:00::
08740367n
```

Para poder relacionar las dos fuentes toman de WordNet todos los posibles sentidos de una determinada palabra y todas las relaciones semánticas de los *synsets*. De la Wikipedia toman todas las entradas y las relaciones dadas por los enlaces de hipertexto de las páginas. Estas relaciones pueden ser de diferentes tipos y no están especificadas desde el punto de vista semántico. Para establecer un *mapping* entre los dos recursos utilizan lo que los autores denominan *contextos de desambiguación*. Estos contextos, para los artículos de la Wikipedia están formados por las etiquetas de sentido que tienen algunas de las entradas, los enlaces de hipertexto y las categorías. En el caso de WordNet estos contextos están formados por todos los sinónimos, hiperónimos e hipónimos, los lemas de las categorías abiertas de la glosa y las *variants* asociadas a los *synsets* hermanos, es decir, los que tienen un hiperónimo directo en común. Para establecer los *mappings* aplican los siguientes criterios:

- Para todas las páginas de la Wikipedia que tengan un título monosémico tanto en la Wikipedia como en WordNet, se enlaza directamente la página con el *synset*.
- Para el resto de página se calcula la intersección de los contextos de desambiguación para todos los sentidos de la Wikipedia y de WordNet.

Dado que las páginas de la Wikipedia disponen de enlaces interlingüísticos, esta relación se puede establecer para todas las lenguas que dispongan de la entrada correspondiente.

4. Uso de corpus paralelos

Se han publicado diversos trabajos sobre el uso de corpus paralelos en tareas relacionadas con la creación de WordNets y de otras ontologías similares. En (Kazakov y Shahid, 2009) se describe una metodología para la adquisición de *variants* asociadas a *synsets* a partir de un corpus paralelo multilingüe. Las *variants* se obtienen comparando las palabras alineadas en diversas lenguas. Si una determinada palabra en una determinada lengua se traduce por más de una palabra en varias lenguas distintas, probablemente quiere decir que la palabra dada tiene más de un significado. Esta suposición funciona también al revés. Si dos palabras de una determinada lengua se traducen por una única palabra en algunas otras lenguas, probablemente quiere decir que las dos palabras comparten un significado común. En Ide, Erjavec, y Tufis (2002) se presenta una idea similar junto a una implementación práctica.

En Fišer (2007) se presenta la construcción del WordNet esloveno utilizando un corpus paralelo multilingüe, un algoritmo de alineación de palabras y WordNets existentes para otras lenguas. Se obtiene un diccionario multilingüe mediante el algoritmo de alineación de palabras y se asignan todos los *synsets* de los WordNets disponibles. Algunas de las palabras en alguna de las lenguas son polisémicas de manera que tienen asignadas más de un

synset. En algunos casos, una palabra es monosémica en al menos una de las lenguas y por lo tanto tiene un único *synset* asignado. Este *synset* se utiliza para desambiguar y asignar un único *synset* al resto de lenguas, incluído el Esloveno. En Sagot y Fišer (2008) se utiliza un método muy parecido para el francés, junto a métodos basados en diccionarios.

En algunos trabajos anteriores presentamos una metodología de construcción de WordNets basadas en el uso de corpus bilingües paralelos. Estos corpus necesitaban estar anotados semánticamente en la parte inglesa. Como este tipo de corpus no está fácilmente disponible exploramos dos estrategias distintas para poder construir automáticamente los corpus necesarios:

- Mediante traducción automática de corpus anotados semánticamente (Oliver y Climent, 2011), (Oliver y Climent, 2012a)
- Mediante anotación semántica automática de corpus bilingües paralelos (Oliver y Climent, 2012b)

5. Estrategias evaluadas

En nuestos experimentos, hemos utilizado y evaluado tres estrategias distintas:

- Uso de diccionarios bilingües
  - Diccionarios bilingües generales
  - Diccionarios enciclopédicos
  - Diccionarios terminológicos
- Uso de Babelnet
- Uso de corpus paralelos

La primera estrategia, la basada en diccionarios bilingües la dividimos en tres grupos, dependiendo del tipo de diccionario utilizado. En esta primera estrategia obtenemos *variants* únicamente para *synsets* cuyas *variants* en inglés son monosémicas. Es decir, traducimos mediante diferentes tipos de diccionarios (generales, enciclopédicos o terminológicos) palabras inglesas monosémicas (asignadas a un único *synset*) y asignamos este *synset* a la correspondiente palabra o palabras castellanas o catalanas dadas por el diccionario. Es decir, tratamos los cuatro grupos monosémicos comentados en la sección 2 simultáneamente. Las dos últimas estrategias tratan tanto los casos monosémicos como polisémicos.

En todos los casos utilizamos una evaluación automática, comparando los resultados obtenidos con las versiones preliminares de los WordNets 3.0 para el castellano y catalán. De esta manera, no todos los resultados obtenidos pueden ser evaluados, ya que algunos *synsets* no tienen *variants* asignadas en estas versiones preliminares. Las *variants* obtenidas para un determinado *synset* que coincidan con las *variants* registradas en la versión preliminar para el mismo *synset* serán consideradas correctas. Si la *variant* obtenida no coincide con ninguna de las registradas en la versión preliminar, será considerada incorrecta. Si no tenemos registrada ninguna *variant* para el *synset* en cuestión, no se contabilizará para el cálculo de la precisión. Hay que tener en cuenta que los resultados considerados errores por la evaluación automática pueden ser correctos en realidad y que simplemente se trate de una *variant* no registrada en la versión preliminar. La evaluación de cada experimento nos produce tres listas: los resultados correctos, que en principio no requieren verificación posterior; los resultados incorrectos, que pueden no serlo y que en algunos casos se han verificado manualmente; los resultados que no se han podido evaluar y que no intervienen en el cálculo de la precisión calculada automáticamente. En algunos experimentos también se ha revisado manualmente estos últimos resultados.

En los próximos apartados presentamos las tres estrategias utilizadas y sus evaluaciones.

6. Uso de diccionarios bilingües

6.1. Uso de diccionarios bilingües generales

6.1.1. Metodología

En este experimento hemos utilizado un diccionario bilingüe para traducir las *variants* inglesas monosémicas escritas minúsculas en el PWN 3.0. Como podemos observar en la tabla 2 éstas son el

68.75 % de las *variants* inglesas monosémicas. De éstas, la mayoría corresponden a substantivos (74.26 %), seguidas de adjetivos (16.48 %). En mucho menor porcentaje se encuentran verbos (5.35 %) y adverbios (3.91 %).

6.1.2. Recusos utilizados

El único recurso necesario es un diccionario bilingüe que tenga una buena cobertura. En nuestros experimentos hemos obtenidos diccionarios bilingües inglés-castellano e inglés-catalán a partir de los diccionarios de transferencia de Apertium (Forcada, Tyers, y Ramírez-Sánchez, 2009) y a partir del Wiktionary[2]. En la tabla 3 podemos observar el número de entradas de cada uno de los diccionarios, así como el número de entradas una vez combinados ambos diccionarios para cada par de lenguas.

Diccionario	**eng-spa**	**eng-cat**
Apertium	20.366	29.154
Wiktionary	23.196	7.393
Total	34.600	32.921

Tabla 3: Número de entradas de los diccionarios bilingües.

6.1.3. Evaluación

Para el castellano podemos obtener un total de 12.676 *variants* de las cuales 7.401 son correctas, 2.997 incorrectas (según la evaluación automática) y no podemos evaluar 2.278. La precisión para el castellano, según la evaluación automática, es del 71.2 %. Se han revisado manualmente todos los resultados considerados incorrectos por la evaluación automática. Esto nos ha permitido calcular una nueva precisión, que ahora asciende al 93.95 %.

Para el catalán obtenemos un total de 8.335 *variants*, de las cuales 4.223 son correctas, 1.083 incorrectas (según la evaluación automática) y no podemos evaluar 3.029. La precisión para el catalán, según la evaluación automática, es del 79.6 %. Del mismo modo que para el castellano, hemos revisado los resultados incorrectos y hemos podido calcular una nueva precisión que asciende al 96.36 %.

[2]http://www.wiktionary.org

6.2. Uso de diccionarios enciclopédicos

6.2.1. Metodología

En este experimento se ha utilizado un diccionario enciclopédico para traducir las *variants* inglesas monosémicas escritas con la primera letra en mayúscula. Éstas constituyen el 31.25 % de las *variants* monosémicas, como se puede ver en la tabla 2. De estos la inmensa mayoría (99.17 %) son substantivos.

6.2.2. Recusos utilizados

Se ha creado un diccionario bilingüe inglés-castellano y uno inglés-catalán a partir de la Wikipedia inglesa, utilizando los enlaces interlingüísticos. De esta manera se han obtenido 59.659 entradas para el inglés-castellano y 22.205 entradas para el inglés-catalán.

6.2.3. Evaluación

Para el castellano podemos obtener un total de 10.356 *variants* de las cuales 4.722 son correctas, 1.916 incorrectas (según la evaluación automática) y no podemos evaluar 3.718. La precisión para el castellano, según la evaluación automática, es del 71.1 %. Si revisamos los casos dados por incorrectos y recalculamos la precisión, ésta sube hasta el 89.74 %. Algunos casos incorrectos lo son por pequeños detalles tipográficos fácilmente subsanables manualmente. Si consideramos este último grupo también como correcto, la precisión asciende hasta el 90.37 %.

Para el catalán obtenemos un total de 7.083 *variants*, de las cuales 2.642 son correctas, 1.278 incorrectas (según la evaluación automática) y no podemos evaluar 3.163. La precisión para el catalán, según la evaluación automática, es del 67.4 %. Después de la revisión manual de los clasificados como incorrectos, la precisión aumenta hasta el 90.94 %, y considerando los que requieren poca modificación manual ésta asciende hasta el 98.34 %.

6.3. Uso de diccionarios terminológicos

6.3.1. Metodología

En este experimento hemos utilizado un conjunto de diccionarios terminológicos multilingües para traducir las *variants*

inglesas monosémicas, tanto las que están en minúsculas como las que están en mayúsculas, ya que los diccionarios terminológicos incluyen a ambas.

6.3.2. Recusos utilizados

El único recurso necesario es un diccionario terminológico que contenga las lenguas que nos interesan y que tenga una buena cobertura. En nuestros experimentos hemos obtenido diccionarios terminológicos bilingües inglés-castellano e inglés-catalán a partir de todos los diccionarios terminológicos que ofrece Termcat en su apartado Terminologia Oberta[3]. De esta manera se ha confeccionado un diccionario terminológico inglés-castellano de 46.761 entradas y uno inglés-catalán de 46.653 entradas.

6.3.3. Evaluación

Para el castellano obtenemos un total de 10.456 *variants*, de las cuales 4.180 son correctas, 3.346 incorrectas (según la evaluación automática) y no podemos evaluar 2.930. La precisión para el castellano, según la evaluación automática, es del 55.5 %. Este resultado es muy bajo, por lo que decidimos revisar manualmente tanto las evaluadas automáticamente como incorrectas, como las no evaluadas. Muchas de las evaluadas automáticamente como incorrectas eran en realidad correctas y muchas de las no evaluadas también lo eran. Con estos nuevos datos, hemos calculado la precisión que es ahora del 98.57 %.

Para el catalán podemos obtener un total de 9.890 *variants* de las cuales 3.007 son correctas, 2.614 incorrectas (según la evaluación automática) y no podemos evaluar 4.269. La precisión para el catalán calculada automáticamente es del 53.5 %. Procedemos de igual manera que para el castellano y obtenemos una nueva precisión del 98.36 %.

7. Uso de Babelnet

7.1. Metodología

Para obtener los WordNets español y catalán a partir de BabelNet utilizaremos el fichero babel-to-wordnet-3.0.txt cuyo contenido hemos mostrado en la sección 3. Este fichero relaciona los *synsets* de WordNet con sus correspondientes entradas de la Wikipedia. Los títulos de las entradas serán las *variants* inglesas. A partir de los enlaces interlingüísticos de la Wikipedia podemos relacionar los títulos de las entradas inglesas con los de la española y catalana, obteniendo de este modo las correspondientes *variants* españolas y catalanas. De hecho, en este experimento utilizamos también los diccionarios enciclopédicos descritos en 6.2, ya que éstos se han obtenido a partir de los enlaces interlingüísticos de la Wikipedia.

7.2. Recusos utilizados

Como ya hemos comentado, utilizamos el fichero babel-to-wordnet-3.0.txt del proyecto BabelNet y los diccionarios enciclopédicos descritos en 6.2.

7.3. Evaluación

Para el castellano obtenemos un total de 26.209 *variants*, de las cuales 14.614 son correctas, 5.065 incorrectas (según la evaluación automática) y no podemos evaluar 6.530. La precisión para el castellano, según la evaluación automática, es del 74.3 %. Revisamos manualmente tanto las evaluadas automáticamente como incorrectas, como las no evaluadas. Una vez realizada la revisión manual podemos calcular un nuevo valor de precisión que es de 81.02 %. Si observamos la revisión manual, nos damos cuenta que una cantidad considerable de errores lo son por pequeños detalles ortotipográficos o por faltar o sobrar algún carácter. Si consideramos estos casos como correctos, la precisión aumenta hasta el 89.24 %.

Para el catalán podemos obtener un total de 18.366 *variants* de las cuales 9.044 son correctas, 3.548 incorrectas (según la evaluación automática) y no podemos evaluar 5.774. La precisión para el catalán, según la evaluación automática, es del 61 %. Hemos procedido de la misma manera que para el castellano y hemos calculado una nueva precisión del 80.91 %, que sube hasta el 97.43 % si consideramos correctas las que han sufrido pequeñas modificaciones.

[3] http://www.termcat.cat/productes/toberta.htm

8. Uso de corpus paralelos

8.1. Metodología

La metodología que presentamos se basa en la alineación a nivel de palabras de un corpus paralelo formado por el original en inglés etiquetado semánticamente y la traducción al castellano o catalán. En este artículo presentaremos únicamente los resultados obtenidos para el castellano. El corpus en inglés está etiquetado semánticamente utilizando los *synsets* de WordNet como etiquetas. En realidad la alineación que nos interesa es la alineación *synset* - palabra castellana, que es directamente deducible del corpus. A continuación podemos observar un ejemplo:

```
English:
Then he noticed that the dry wood of the wheels
had swollen.
Sense Tagged English:
00117620r he 02154508v that the 02551380a
15098161n of the 04574999n had 00256507v .
Spanish Translation:
Entonces se dio cuenta de que la madera seca
de las ruedas se había hinchado.
```

De este fragmento del corpus podremos obtener las siguientes relaciones mediante el algoritmo de alineación de palabras:

```
00117620r - entonces
02154508v - darse cuenta
02551380a - seco
15098161n - madera
```

8.2. Recusos utilizados

Como hemos comentado, exploramos dos estrategias para obtener el corpus necesario. Una se basa en la traducción automática de corpus etiquetados semánticamente. En este caso utilizamos el corpus Semcor[4] (Miller et al., 1993) y el Princeton WorNet Gloss Corpus (PWGC)[5] y el traductor automático Google Translate[6]. La segunda estrategia se basa en el etiquetado semántico automático de corpus paralelos. En este caso utilizamos en corpus European Parliament Proceedings Parallel Corpus (Europarl)[7] (Koehn, 2005) y Freeling (Padró et al., 2010) como etiquetador. En todos los casos se ha utilizado el Berkeley Aligner (Liang, Taskar, y Klein, 2006) como algoritmo para la alineación de palabras.

[4]http://www.cse.unt.edu/∽rada/downloads.html
[5]http://wordnet.princeton.edu/glosstag.shtml
[6]http://translate.google.com
[7]http://www.statmt.org/europarl/

8.3. Evaluación

La evaluación en este caso se ha llevado a cabo de manera automática y de una manera acumulativa empezando por el *synset* más frecuente en el corpus y siguiendo un orden descendiente en frecuencia. Los resultados se presentan de manera gráfica donde los valores del eje y representan la precisión acumulada y los valores del eje x el número de *synsets* extraídos (es decir, el número de *variants* asociadas a *synsets*).

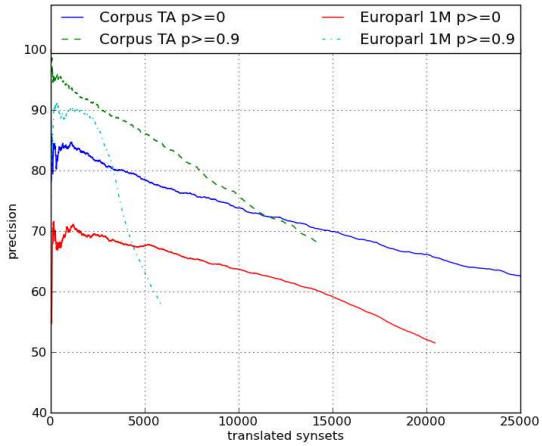


Figura 1: Comparación de los métodos basados en corpus paralelos.

En la figura 1 podemos observar los valores correspondientes al uso del corpus obtenido por traducción automática (Corpus TA) y los correspondientes a un fragmento de 1 millón de oraciones del corpus Europarl. Para ambos corpus ofrecemos dos gráficos, uno tomando todas las posibles alineaciones ($p \geq 0$) y otro tomando únicamente las alineaciones que tienen una probabilidad de 0.9 o superior ($p \geq 0{,}9$). Los mejores resultados se obtienen para el corpus obtenido mediante traducción automática y para $p \geq 0{,}9$, ya que podemos obtener unas 5.000 *variants* con una precisión del 85 % o superior.

9. Conclusiones

En este artículo hemos presentado una revisión de métodos de construcción de WordNets mediante la estrategia de expansión. Hemos presentado valores de evaluación de estos métodos para el castellano y catalán. Los valores obtenidos son similares a los obtenidos por Atserias et al. (1997) y Benítez et al. (1998). Los valores de precisión de nuestros métodos basados en diccionarios son similares a los valores obtenidos por los cuatro criterios monosémicos (que son de entre el 85 % y el 92 % para el castellano y entre el 93.3 % al 97.6 % para el catalán). La estrategia basada en BabelNet y en corpus paralelos no son fácilmente comparables con los resultados de Atserias y Benítez ya que no hemos realizado una separación de los resultados según el grado de monosemia-polisemia.

Un aspecto importante que se deduce del artículo es que la evaluación automática no es suficientemente fiable. En algunos casos hemos pasado de precisiones del 53.5 % al 98.36 % una vez revisados manualmente todas las *variants* obtenidas.

Así pues, los métodos presentados se pueden considerar de gran utilidad para la creación de WordNets. La validación final de los resultados es importante, pero esta validación es mucho más rápida que la creación de WordNets manualmente desde cero.

Como trabajo futuro pretendemos mejorar los resultados obtenidos mediante el uso de corpus paralelos. Por un lado, si utilizamos corpus etiquetados semánticamente y traducción automática, tenemos un problema de precisión dado por la calidad de la traducción automática (especialmente en la tarea de selección léxica) y un problema de cobertura ya que en el extremo podremos obtener únicamente las *variants* correspondientes a los *synsets* presentes en el corpus. Por este motivo, y dada la relativa facilidad de obtener corpus paralelos multilingües de gran tamaño, preferimos trabajar con esta segunda estrategia. En este caso, la precisión de la tarea de etiquetado semántico automático no es suficiente. Por este motivo queremos aprovechar la información presente en los corpus paralelos multilingües para llevar a cabo una desambiguación a través del etiquetado de diversas lenguas que dispongan de WordNets, tal y como se explica en (Ide, Erjavec, y Tufis, 2002).

Por otro lado, en los experimentos llevados a cabo hasta el momento con corpus paralelos hemos extraído los WordNets partiendo de cero. Como que ya disponemos de versiones de los WordNets para el castellano y catalán, podemos utilizar toda esta información para simplificar la tarea de extracción de nuevas *variants*.

Los WordNets 3.0 para el castellano y catalán desarrollados en este proyecto se pueden descargar libremente de http://adimen.si.ehu.es/web/MCR . Los WordNets del castellano y catalán tienen una licencia Creative Commons, concretamente la Attribution 3.0 Unported (CC BY 3.0) license[8].

Bibliografía

Atserias, J., S. Climent, X. Farreres, G. Rigau, y H. Rodriguez. 1997. Combining multiple methods for the automatic construction of multi-lingual WordNets. En *Recent Advances in Natural Language Processing II. Selected papers from RANLP*, volumen 97, página 327–338.

Benítez, Laura, Sergi Cervell, Gerard Escudero, Mònica López, German Rigau, y Mariona Taulé. 1998. Methods and tools for building the catalan WordNet. En *In Proceedings of the ELRA Workshop on Language Resources for European Minority Languages*.

Fellbaum, C. 1998. *WordNet: An electronic lexical database*. The MIT press.

Fišer, D. 2007. Leveraging parallel corpora and existing wordnets for automatic construction of the slovene wordnet. En *Proceedings of the 3rd Language and Technology Conference*, volumen 7, página 3–5.

Forcada, M. L, F. M Tyers, y G.írez-Sánchez@. 2009. The apertium machine translation platform: five years on.

[8] http://creativecommons.org/licenses/by/3.0/

En *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, página 3–10.

Ide, N., T. Erjavec, y D. Tufis. 2002. Sense discrimination with parallel corpora. En *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions-Volume 8*, página 61–66.

Kazakov, D. y A.R. Shahid. 2009. Unsupervised construction of a multilingual WordNet from parallel corpora. En *Proceedings of the Workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography, and Language Learning*, página 9–12.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. En *MT summit*, volumen 5.

Liang, Percy, Ben Taskar, y Dan Klein. 2006. Alignment by agreement. En *Proceedings of the HLT-NAACL '06.*

Miller, George A, Claudia Leacock, Randee Tengi, y Ross T Bunker. 1993. A semantic concordance. En *Proceedings of the workshop on Human Language Technology*, HLT '93, página 303–308, Stroudsburg, PA, USA. Association for Computational Linguistics. ACM ID: 1075742.

Navigli, Roberto y Simone Paolo Ponzetto. 2010. BabelNet: building a very large multilingual semantic network. En *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, página 216–225, Stroudsburg, PA, USA. Association for Computational Linguistics. ACM ID: 1858704.

Oliver, A. y S. Climent. 2011. Construcción de los wordnets 3.0 para castellano y catalán mediante traducción automática de corpus anotados semánticamente. En *Proceedings of the 27th Conference of the SEPLN, Huelva Spain.*

Oliver, A. y S. Climent. 2012a. Building wordnets by machine translation of sense tagged corpora. En *Proceedings of the Global WordNet Conference, Matsue, Japan.*

Oliver, A. y S. Climent. 2012b. Parallel corpora for wordnet construction. En *Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (Cicling 2012). New Delhi (India).*

Padró, L., S. Reese, E. Agirre, y A. Soroa. 2010. Semantic services in freeling 2.1: Wordnet and UKB. En *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010).*

Sagot, B. y D. Fišer. 2008. Building a free french wordnet from multilingual resources. En *Proceedings of OntoLex 2008*, Marrackech (Morocco).

Vossen, P. 1998. Introduction to Eurowordnet. *Computers and the Humanities*, 32(2):73–89.

Vossen, P. 1999. EuroWordNet a multilingual database with lexical semantic networks. *Computational Linguistics*, 25(4).

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 129-138 recibido 05-05-12 revisado 31-05-12 aceptado 05-06-12

Labeling Semantically Motivated Clusters of Verbal Relations

Etiquetado de clusters de relaciones verbales motivados semanticamente

Gabriela Ferraro
Dept. of Information and Communication Technologies, Universitat Pompeu Fabra
gabriela.ferraro@upf.edu

Leo Wanner
ICREA and Dept. of Information and Communication Technologies, Universitat Pompeu Fabra
leo.wanner@upf.edu

Resumen: El clustering de documentos es un campo de investigación popular en los ámbitos del Procesamiento del Lenguaje Natural, la Minería de Datos y la Recuperación de información (RI). El problema de agrupar unidades léxicas mediante clustering ha sido menos estudiado y menos aún, el problema de etiquetar los clusters. Sin embargo, en nuestra aplicación que trata sobre la extracción de tuplas de relaciones para ser usadas como entrada a programas para dibujar diagramas de bloques o mapas conceptuales, este problema es fundamental. La valoración de varias estrategias de etiquetado de clusters de documentos nos revela que algunas de estas técnicas pueden ser también aplicadas para etiquetar nuestros clusters, compuestos por verbos semánticamete similares. Para confirmar esta suposición, llevamos a cabo una serie de experimentos y evaluamos su rendimiento contra *baselines* y un *golds-tandard* de clusters etiquetados.
Palabras clave: Etiquetado de clusters, clustering, clasificación de relaciones

Abstract: Document clustering is a popular research field in Natural Language Processing, Data Mining and Information Retrieval. The problem of lexical unit (LU) clustering has been less addressed, and even less so the problem of labeling LU clusters. However, in our application that deals with the distillation of relational tuples from patent claims as input to block diagram or a concept map drawing programs, this problem is central. The assessment of various document cluster labeling techniques lets us assume that despite some significant differences that need to be taken into account some of these techniques may also be applied to verbal relation cluster labeling we are concerned with. To confirm this assumption, we carry out a number of experiments and evaluate their outcome against baselines and gold standard labeled clusters.
Keywords: cluster labeling, clustering, relation classification

1. Introduction

Clustering is a popular field of research in Natural Language Processing, Data Mining and Information Retrieval. Most often, the goal is to group the documents in a given document collection with respect to their semantic similarity (Hearst and Pedersen, 1996; Zhu et al., 2006). Some works also address the problem of grouping lexical units (LUs) according to specific semantic criteria. For instance, (Yang and Powers, 2005) group object nouns with respect to their proximity in a taxonomy. According to their approach, *peach*, *pear*, *apricot*, *strawberry*, *banana*, *melon*, etc. form a single cluster and so do *birch*, *fir*, *oak*, etc. (Sekine, 2005; Schulte im Walde, 2006; Korhonen, Krymolowski, and Collier, 2006; Davidov and Rappoport, 2008) cluster verbal relations into classes such as {*compress*, *reduce*, *minimize*, *trim*, *cut*, etc.}, again in accordance with predefined semantic criteria. However, surprisingly little work has been done so far on labeling the obtained LU clusters; the few proposals made on cluster labeling at all nearly exclusively refer to document clusters. This is despite the fact that an ideal cluster label not only reflects the semantic commonalities shared by all members of a given cluster, but also uniquely differentiates this cluster from other clusters in the collection. It could thus be used in any term generalization task.

In our application, we face such a term ge-

ISSN 1135-5948

neralization task. We aim at distilling relational tuples from functional descriptions such as patent claims in order to provide input to block diagram or a concept map drawing programs. This implies that we directly face the problems of verbal relation clustering and relation cluster labeling.[1] A straightforward use of verbal relation names extracted from a functional description (as, e.g., *comprise* between *automatic focusing device* and *an objective lens* or *include* between *astigmatic optical system* and *optical element*) as done, e.g., by (Cascini and Russo, 2007), is not appropriate: block diagrams and concept maps are conceptual representations. They must achieve a sufficient abstraction over concrete terms. Thus, *comprise*, *include*, *contain*, *have*, etc. are sufficiently similar to be considered the same relation in a block diagram and thus should be captured by the same concept—for instance, 'part-of' and named the same. In the same vein, *cause*, *lead to*, *result in*, etc. should be captured by a single concept—'cause'.

In what follows, we focus on the problem of verbal relation cluster labeling. Section 2 describes the problem of cluster labeling in general. Section 3 outlines the verbal relation cluster labeling experiments we carried out to assess how labeling strategies inspired by document cluster labeling perform on the LU cluster labeling task, and Section 4 presents the evaluation of these experiments. In Section 5, we summarize the related work on cluster labeling. In Section 6, we draw some conclusions from our experiences and outline some lines of future work along which the cluster labeling strategies we experimented with can be improved.

2. The problem of cluster labeling

As already mentioned in Section 1, cluster labeling proposals focused so far mainly on document cluster labeling. Document clustering is a key technique in cluster-based search, *scatter-gather*-based document browsing, opinion mining, data mining, etc. (Muresan and Harper, 2004; Pirolli, 2007). Cluster labeling is used in connection with clustering to make the results of clustering more transparent to the user (Osinski and Weiss, 2005; Mika, 2005). As cluster labels, sentences, phrases or simply lists of terms that are assumed to characterize well the clusters in question are taken.

In the clustering literature, two main strategies of document cluster labeling can be identified: (i) internal cluster labeling and (ii) differential cluster labeling. In internal cluster labeling, the label of a given cluster is chosen drawing solely on the content of the cluster itself. For instance, (Chen and Liu, 2004; Cutting, Karger, and Pedersen, 1993) suggest to pick as label a linguistic construction or a sequence thereof (e.g., the title of one of the documents in the cluster, a list of terms, a phrase, etc.) that proves to be closest to the cluster's *centroid* according to measures such as cosine. (Cutting et al., 1992; Osinski and Weiss, 2005) propose frequency-based internal labeling strategies which select as label the term or a list of terms that are most frequent in the given cluster. Internal labeling strategies have the advantage of being simple.

In differential labeling, the label of a cluster is chosen by contrasting this cluster with the other available clusters. Often, statistical measures such as Mutual Information (MI), Information Gain and the χ^2 test are applied, which calculate the statistical dependence of a candidate label on the cluster in question (relatively to the other clusters in the collection): if the candidate label is dependent on the cluster (more than on the other clusters), it is considered a good label for it; see, e.g., (Pellegrini, Maggini, and Sebastiani, 2006; Carmel, Roitman, and Zwerdling, 2009).

Our task is different from the task of document cluster labeling. As already mentioned, we face the problem of labeling clusters of semantically similar verbal LUs such that they can be used as relation labels in a conceptual map-like representation. In such a setting, we need to choose as label a single lexical element (a phrase or a list of terms are not appropriate). Still, the general ideas underlying the internal and differential clustering strategies seem to stay valid: we can choose a label of a cluster either by drawing solely on the members of this cluster or by exploring the influence of the other clusters as well.

When choosing the label, we can either pick one of the lexemes of the cluster in question or choose an abstract label that captu-

[1] Obviously, we also face the problem of relation extraction. However, we cannot delve into this topic here. Interested readers are asked to consult (Ferraro and Wanner, 2011).

res, in a sense, all members of the cluster. To obtain the most suitable abstract term in a cluster, we can look up the members of the cluster in a taxonomy or to look for a common hyperonym among the members using an external resource such as Wordnet (WN). The problem with using WN might be that in contrast to its nominal hierarchies which tend to be rich and deep, the verbal hierarchies in WN are relatively flat and poor.

Another possibility is to enrich clusters by lexemes retrieved from thesauri since thesauri group lexemes according to their similarity of meaning. Our intuition is that enriching clusters by semantically related lexemes retrieved from a thesaurus increases the possibility of finding a common abstract label. This intuition is based on two observations: (i) we can look for the most frequent common thesaurus term in the clusters, avoiding the restriction of assigning as label a lexeme from the cluster itself, (ii) we can further apply statistical tests, such as Mutual Information, which are best suited for clusters that contain overlapping terms.

In what follows, we carry out a number of experiments in order to assess to what extent the possible approaches sketched above lead to successful verbal relation cluster labeling, and to be able to choose the best one for our applications.

3. Cluster labeling experiments

The input to all cluster labeling strategies described in this section is a set of verb clusters, which have been grouped automatically according to their semantic similarity in a previous step of our application. We have experimented with seven different cluster labeling strategies. Three strategies are internal cluster labeling techniques and four are differential cluster labeling techniques.

3.1. Internal cluster labeling

We experimented with the following internal cluster labeling strategies:

Frequency-oriented labeling (Freq): Choose as cluster label the member of the cluster with the highest frequency in the reference corpus. This strategy is motivated by classic cluster labeling techniques that choose one of the members of the cluster as its label (Osinski and Weiss, 2005; Chen and Liu, 2004). It has the advantage of being simple. Thus, for the cluster:

> C_i = {*bound:63, limit:74, restrain:21, inhibit:101, fasten:49, fix:53, secure:13, lock:28*}

this strategy suggests *inhibit* as cluster label (the suffix ':X' denotes the frequency of the corresponding member in the reference corpus).

Verb hyperonym-oriented labeling (VHyp): Choose as cluster label the most frequent hyperonym of the cluster as it appears in the WN verb hierarchy. To implemente this strategy, first, for each member of a cluster, all its WN hyperonyms are retrieved and the most frequent hyperonym synset is selected. Then, from this synset, the most frequent lexeme in the corpus is chosen as the cluster label. For example, for the above cluster, the most frequent hyperonym synset is:

> $C_i(hyper)$ = {*bound3, check4, confine1, limit1, restrain2, restrict3, throttle1, trammel2, decide1, decide upon1, determine4, make-up one's mind1* }

From this hyperonym synset, the most frequent hyperonym found in the reference corpus is *limit:1* (in this case, the suffix ':X' stands for the WN sense). Therefore, *limit* is chosen as the cluster label. This strategy is motivated by the fact that the cluster label should be more abstract to ensure that it captures all members of the cluster.

Thesaurus Freq (ThesFreq): Choose as cluster label the most frequent lexeme found in a cluster populated by LUs from the Open-Office Thesaurus. To populate a cluster by LUs from the thesaurus, for each of the members of the cluster, the verbal lexemes related to it via the different semantic relations are retrieved from the thesaurus. For instance, the following verbal lexemes are associated with the member *lock* of the cluster C_i introduced above:

> *thesaurus(lock)* = {*fasten, fix, secure, lock up, lock up, engage, mesh, operate, move, displace, engage, interlock, interlace, hold, take hold, interlock, embrace, hug, bosom, squeeze, overwhelm, overpower, sweep over, whelm, overcome, overtake, lock in, lock away, put away, shut up, shut away, lock up, confine, pass, go through, go across, construct, build, make*}

The most frequent among them is *fix*. It is thus chosen as cluster label.

Cuadro 1: Examples of the performance of the internal cluster labeling strategies

Gold Standard Clusters	GS	Freq	VHyper	ThesFreq
{comprise, contain, have, include}	contain	comprise	comprise	get
{bound, limit, restrain, inhibit, fasten, fix, secure, lock}	limit	inhibit	limit	fix
{compress, trim, reduce, minimize}	reduce	reduce	cut	lessen
{extract, pull-out}	extract	extract	remove	take-out
{remove, cut, delete, erase, exclude}	remove	remove	remove	take-out
{enter, insert, interpose, introduce, enclose}	insert	insert	connect	introduce
{apply, feed, provide, give, use, supply, render}	produce	provide	provide	give
{hold , maintain, retain, support, prevent}	keep	support	maintain	hold
{accord, allow, let, permit}	let	accord	have	permit

Table 1 displays a sample of the results of the application of the internal cluster labeling strategies to a number of gold standard clusters.

3.2. Differential cluster labeling

The differential cluster labeling strategies have been implemented using the MI and the χ^2 measures.

VHyp MI-oriented labeling (VHyp-MI): Choose as cluster label the hyperonym with the highest MI value. First, for each member of the cluster, its WN hyperonyms are retrieved (as already in the internal VHyp-oriented labeling strategy). Then, for a given cluster, we calculate the MI of each hyperonym and select as label that hyperonym which shows the highest MI value. Consider, for illustration, Table 2, where the MI values of the label candidates for the cluster C_i are displayed.

The hyperonym with the highest MI value turns out to be *moderate*. Therefore, it is chosen as the cluster label.

Cuadro 2: Examples of the MI values of the candidate labels for C_i

Label candidate	MI value
moderate	1391.80
restrict	1394.56
throttle	1394.56
restrain	1389.33
put restrictions on	1388.25
check	1387.05

VHyp χ^2-oriented labeling (VHyp-χ^2): Choose as cluster label the hyperonym with the highest χ^2 value. The procedure is the same as above, only that instead of MI, the χ^2 measure is applied. Table 3 displays the χ^2 values for the different candidate labels for C_i. Since *throttle* shows the highest χ^2 value, it is chosen as label.

Cuadro 3: Examples of the χ^2 values of the candidate labels for C_i

Label candidate	χ^2 value
throttle	90.30
confine	82.29
hold-in	76.31
restrain	65.83
check	57.68
fasten with a lock	34.82

Thesaurus MI-oriented labeling (ThesMI): Choose as cluster label the thesaurus lexeme with the highest MI value. The clusters are populated with the thesaurus matches as in the ThesFreq strategy. An example of the experiment run is shown in Table 4. For the cluster C_i, the term with the highest MI value is *restrict*.

Cuadro 4: Examples of the Thesaurus-MI values for the candidate labels for C_i

Label candidate	MI value
restrict	1392.68
interlock	1388.25
stick	1384.86
tie	1377.48
fix	1374.42
lessen	1374.34

Thesaurus χ^2-oriented labeling

(Thesχ^2): Choose as cluster label the thesaurus lexeme with the highest χ^2 value. Again, the clusters are populated with its thesaurus matches as in the ThesFreq strategy. An example of the results of an experiment run is shown in Table 5. For C_i, the term with the highest χ^2 value is *restrict*.

Cuadro 5: Examples of the Thesaurus χ^2 values for some candidate labels for C_i

Label candidate	**χ^2 value**
restrict	76.91
trammel	76.90
curb	76.31
hold in	65.83
control	34.48
fasten	20.70

Table 6 presents some examples of the performance of the application of the differential cluster labeling strategies.

3.3. Fallback strategies

Sometimes, differential labeling strategies come up with several candidate labels with the same weight. Since we have to decide which of them to choose, we use two different simple *fallback strategies*. The first of them chooses as label the candidate with the highest frequency in the reference corpus. The second picks the label randomly among the candidates with the same weight.

4. Evaluation

We carried out a qualitative evaluation of the implemented cluster labeling strategies, resorting to human judges. For the evaluation, we use a gold standard of 54 verb clusters as the list of clusters to name. The 54 clusters were presented to three judges, together with the labels assigned to each of the clusters by our system and by a human collaborator (the gold standard labels), such that the judges did not know the origin of a label. For each cluster, the judges were asked to qualify all the labels as 'correct' (corr), 'partially correct' (pcorr) or 'incorrect' (incorr). Table 7 shows the evaluation results of the internal labeling strategies and Table 8 the results of differential labeling.

Table 7 reveals that the Freq strategy, which chooses as the label of a cluster its most frequent member, achieves with 78 % of correctness the best results. This is somewhat surprising since one would expect that a label that abstracts over the individual members of a cluster would be more appropriate. However, the VHyp strategy shows significantly worse results than Freq, achieving only 43 % of correctness. We assume that this is largely because the hyperonyms in WN tend to be too abstract to serve as a label of their hyponyms—as is, e.g., also the case with *move* for the cluster {*disperse*, *propagate*}. The ThesFreq strategy shows acceptable results, achieving a 58 % of correctness and 26 % of partially correctness. The weakness of the ThesFreq strategy is that it uses all semantic relations in the thesaurus to retrieve candidate labels. The use of synonymy and hyperonymy only appears more promising and will be tested in the future. A baseline strategy that arbitrarily chooses a member of a given cluster as the label of this cluster reaches a 31 % match with the gold standard labels.

Cuadro 7: Internal clustering labeling strategies evaluation.

	% Corr	**% Pcorr**	**% Incorr**
Gold st.	77 %	17 %	7 %
Freq	78 %	20 %	2 %
VHyp	43 %	25 %	32 %
ThesFreq	58 %	26 %	16 %

Cuadro 8: Differential clustering labeling strategies evaluation.

	%Corr	**%Pcorr**	**%Incorr**
Gold st.	77 %	17 %	7 %
VHyp-MI	50 %	45 %	5 %
VHyp χ^2	60 %	27 %	13 %
ThesMI	70 %	25 %	5 %
Thes χ^2	67 %	22 %	11 %

Table 8 shows the results of the differential cluster labeling strategies. As in internal labeling strategies, the strategies that used the thesaurus perform better than the ones that use verb hyperonyms from WN. The best differential strategy is ThesMI, achieving a 70 % of correctness. The Thesχ^2 strategy has a slightly lower score, achieving a 67 % of correctness.

Cuadro 6: Examples of the performance of the differential cluster labeling strategies

Gold Standard Clusters	**GS**	**VHyp-MI**	**VHyp χ^2**	**ThesMI**	**Thesχ^2**
{comprise, contain, have, include}	contain	comprise	incorporate	incorporate	incorporate
{bound, limit, restrain, inhibit, fasten, fix, secure, lock}	limit	moderate	throttle	restrict	restrict
{compress, trim, reduce, minimize}	reduce	trim down	thin-out	find-out	minify
{extract, pull-out}	extract	move forcibly	pull-up	pull-up	press-out
{remove, cut, delete, erase, exclude}	remove	erase	kill	cancel	take-out
{enter, insert, interpose, introduce, enclose}	insert	shut-it	enclose	pull-in	pull-in
{apply, feed, provide, give, use, supply, render}	produce	administer	furnish	furnish	furnish
{hold , maintain, retain, support, prevent}	keep	hold on	hold on to	defend	defend
{accord, allow, let, permit}	let	grant	grant	consent	consent

According to the qualitative evaluation, the performance of one of the internal cluster labeling strategies, namely 'Freq', is the one that is most similar to the performance of our human judge, while the strategies that are based on WN hyperonyms, perform significantly poorer—although in the literature, WN hyperonym hierarchies are most commonly used for lexical labeling. This is partly due to the fact that most of the works on lexical labeling target the labeling of nominal rather than verbal clusters and WN, which is used as reference resource, has, in general, very flat verbal hierarchies.

Differential strategies that use the thesaurus as an external resource show competitive results, as they are close to the human judgements. A weakness of differential labeling is that sometimes labels are low frequency terms and appear somewhat questionable. For instance, in the ThesMI strategy, the cluster {*become*, *convert*, *turn*} is labeled by the term *metamorphose*, which is judged as 'partially correct'. Even if this term reflects the meaning of the cluster it is considered inappropriate to be used as cluster label in the technical domains of our corpus.

5. Related work

Although the focus of cluster labeling research has been on document cluster labeling, some proposals exist also for LU cluster labeling. In what follows, we focus on those proposals. Thus, the proposal by (Pantel and Ravichandran, 2004), which addresses the problem of labeling clusters of semantically similar nouns, is an example for internal cluster labeling. The input of their system are semantic classes (clusters of nouns) and the output is a ranked list of label names for each semantic class. First, for each member of a cluster, grammatical signatures that capture its prototypical semantic context in different occurrences are computed. In other words, each word of a cluster is represented by a feature vector where each feature corresponds to a context in which the word occurs. As context, the grammatical functions (such as *subject*, *direct object*, etc.) computed by the Minipar (Klein and Manning, 2003) parser are used. For example, "catch —" represents a verb object context. If the word *wave* would occur in this context, the context would thus include the feature of *wave*. Then, among these signatures, simple hyperonymy patterns, such as "Noun–apposition–Noun" (e.g., H1N1, the disease) are searched. At last, the mutual information scores for each hyperonymy candidate are calculated and the highest scoring hyponym is chosen as the name of the cluster. Further similar proposals of internal labeling include (Carmel, Roitman, and Zwerdling, 2009; Manning, Raghavan, and Schütze, 2008).

The proposal by (Dias et al., 2009) is, in principle, a proposal on document cluster labeling because it addresses the problem of clustering of webpage results and the subsequent labeling of the obtained clusters. However, since it chooses as label of a given cluster

a noun or a noun compound it is worth to be mentioned here. It is an example for differential cluster labeling in that the chosen label (i) occurs in most of the URLs of the cluster in question, (ii) discriminates the cluster sufficiently well from the other clusters.

The more complex problem of labeling nodes in a hierarchy (which requires distinguishing more general labels for parents from more specific labels for children) is tackled by (Glover et al., 2002) and (Treeratpituk and Callan, 2006). Some clustering algorithms attempt to find a set of labels first and then build (often overlapping) clusters around the labels; see, e.g., (Osinski and Weiss, 2005; Zamir and Etzioni, 1999; Mika, 2005)—even if, as pointed out by (Manning, Raghavan, and Schütze, 2008), no comprehensive study that compares the quality of such *label-based* clustering with the classic clustering algorithms is known.

As far as labeling clusters of similar verbs is concerned, i.e., the problem addressed in this paper, to the best of our knowledge, no work has been dedicated to this problem as yet.

6. Conclusions and future work

In the context of semantic verb clustering, differential labeling strategies seem more suitable since they take into account the panorama of all clusters. This is coherent with the evaluation results obtained so far: differential labeling strategies outperform nearly all internal labeling strategies; the exception is the internal labeling based on frequency, which performs better.

The results also shows that internal cluster labeling strategies are efficient, but since they do not distinguish terms that are frequent in the collection in general from those that are frequent only in the cluster, they may assign the same label to more then one cluster. With respect to differential labeling, we need to take into account that very low frequency terms should be omitted as label candidates as they would not be the best in representing a whole cluster. So far, we did not apply any frequency filters, such that all terms are considered as label candidates. In the future, we plan to experiment with a hybrid labeling technique that combines internal and differential methods and to take the context of the verbal relations into account. Furthermore, we plan to experiment with other external lexical resources for enriching clusters for the purpose of labeling—among them, synonym dictionaries. Some work has been done in the past on grouping semantically similar nouns and semantically similar adjectives (Rooth et al., 1999; Boleda, Schulte im Walde, and Badia, 2008). Given that verb nominalizations and adjectives are also frequently used in patent claims, both word categories need to be considered in our future work as well.

Bibliografía

Boleda, G., S. Schulte im Walde, and T. Badia. 2008. An analysis of human judgements on semantic classification of catalan adjectives. *Research on Language and Computation*, 6:247–271.

Carmel, D., H. Roitman, and N. Zwerdling. 2009. Enhancing cluster labeling using wikipedia. In *Proceedings of the 32nd international ACM SIGIR Conference*, SIGIR '09, pages 139–146, New York, NY, USA. ACM.

Cascini, G. and D. Russo. 2007. Computer-aided analysis of patents and search for triz contradictions. *International Journal of Product Development*, 4(1):52–67.

Chen, Keke and Ling Liu. 2004. Clustermap: Labeling clusters in large datasets via visualization. In *Proc. of ACM Conf. on Information and Knowledge Mgt. (CIKM*, pages 285–293.

Cutting, D. R., D. R. Karger, and J. O. Pedersen. 1993. Constant interaction-time scatter/gather browsing of very large document collections. In *Proceedings of the 16th annual international ACM SIGIR Conference*, SIGIR '93, pages 126–134. ACM.

Cutting, D. R., J. O. Pedersen, D. Karger, and J. W. Tukey. 1992. Scatter/gather: a cluster-based approach to browsing large document collections. In *Proceedings of the 15th annual international ACM SIGIR Conference*, SIGIR '92, pages 318–329. ACM.

Davidov, D. and A. Rappoport. 2008. Unsupervised discovery of generic relationships using pattern clusters and its evaluation by automatically generated sat analogy questions. In *Meeting of the Association*

for Computational Linguistics, pages 692–700.

Dias, G., S. Pais, F. Cunha, H. Costa, H. Machado, T. Barbosa, and B. Martins. 2009. Hierarchical soft clustering and automatic text summarization for accessing the web on mobile devices for visually impaired people. In *Proceedings of the FLAIRS Conference*, pages 231–236.

Ferraro, G. and L. Wanner. 2011. Towards the derivation of verbal content relations from patent claims using deep syntactic structures. *Knowledge-Based Systems*, 24:1233 – 1244.

Glover, E. J., K. Tsioutsiouliklis, S. Lawrence, D. M. Pennock, and G. W. Flake. 2002. Using web structure for classifying and describing web pages. In *Proceedings of the 11th international conference on World Wide Web*, pages 562–569. ACM.

Hearst, M. A. and J. O. Pedersen. 1996. Reexamining the cluster hypothesis: scatter/gather on retrieval results. In *Proceedings of the 19th annual international ACM SIGIR Conference*, SIGIR '96, pages 76–84, New York, NY, USA. ACM.

Klein, D. and C. Manning. 2003. Accurate unlexicalized parsing. *Proceedings of the 41st meeting of the Association for Computational Linguistics*.

Korhonen, A., Y. Krymolowski, and N. Collier. 2006. Automatic classification of verbs in biomedical texts. In *Proceedings of the 21st ACL*, ACL-44, pages 345–352. Association for Computational Linguistics.

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

Mika, K. 2005. Findex: Search Result Categories Help Users when Document Ranking Fails. In *CHI '05: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 131–140, New York, NY, USA. ACM.

Muresan, G. and D. J. Harper. 2004. Topic modeling for mediated access to very large document collections. *Journal of the American Society for Information Science and Technology*, 55:892–910.

Osinski, S. and D. Weiss. 2005. A concept-driven algorithm for clustering search results. *IEEE Intelligent Systems*, 20:48–54, May.

Pantel, P. and D. Ravichandran. 2004. Automatically labeling semantic classes. In *HLT - NAACL*, pages 321–328.

Pellegrini, M., M. Maggini, and F. Sebastiani. 2006. M.: Cluster generation and cluster labelling for web snippets: A fast and accurate hierarchical solution. Technical report, In Proceedings of the 13th SPIRE 2006.

Pirolli, P. 2007. *Information Foraging Theory: Adaptive Interaction with Information*. Oxford University Press.

Rooth, M., S. Riezler, D. Prescher, G. Carroll, and F. Beil. 1999. Inducing a semantically annotated lexicon via em-based clustering. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 104–111. Association for Computational Linguistics.

Schulte im Walde, S. 2006. Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics*, 32(2):159–194.

Sekine, S. 2005. Automatic paraphrase discovery based on context and keywords between ne pairs. In *International Workshop on Paraphrase*.

Treeratpituk, P. and J. Callan. 2006. Automatically labeling hierarchical clusters. In *Proceedings of the 2006 international conference on Digital government research*, dg.o '06, pages 167–176, New York, NY, USA. ACM.

Yang, D. and D. Powers. 2005. Measuring semantic similarity in the taxonomy of wordnet. In *ACSC '05: Proceedings of the Twenty-eighth Australasian conference on Computer Science*, pages 315–322. Australian Computer Society, Inc.

Zamir, O. and O. Etzioni. 1999. Grouper: A dynamic clustering interface to web search results. pages 1361–1374.

Zhu, Y. H., G. Z. Dai, B. C. M. Fung, and D. J. Mu. 2006. Document clustering method based on frequent co-occurring

words. In *Proc. of the 20th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, pages 442–445, Wuhan, China, November.

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 139-146 recibido 04-05-12 revisado 01-06-12 aceptado 05-06-12

A Hybrid Approach to Treebank Construction

Una aproximación híbrida a la construcción de treebanks

Montserrat Marimon
Dpt. de Lingüística General
Universitat de Barcelona
Barcelona, Spain
montserrat.marimon@ub.edu

Lluís Padró
TALP Research Center
Universitat Politècnica de Catalunya
Barcelona, Spain
padro@lsi.upc.edu

Resumen: Este artículo describe investigación sobre los efectos de la desambiguación morfosintáctica usada como un preproceso de un analizador sintáctico profundo basado en HPSG, en el contexto del desarrollo de un *treebank* del español de código abierto, en el entorno de DELPH-IN. La anotación treebank se realiza manualmente tomando las decisiones apropiadas entre las opciones propuestas por el sistema y ordenadas por un módulo estadístico. Los experimentos presentados muestran que el uso de un etiquetador reduce la ambigüedad de las frases, y contribuye a limitar la cantidad de frases cuyo análisis sobrepasaa el límite de tiempo, y ayuda a al módulo estadístico a clasificar el árbol correcto entre los n mejores. Por un lado, nuestros resultados validan los beneficios ya reportados en la literatura de tal preproceso de análisis profundo con respecto a la velocidad, cobertura y precisión. Por otro lado, proponemos una estrategia basada en existentes herramientas de código abierto y recursos para desarrollar con alta consitencia treebanks de sintaxis profunda para idiomas con limitada disponibilidad de recursos lingüísticos.
Palabras clave: Anotación sintáctica profunda de corpus, análisis HPSG, desambiguación morfosintáctica.

Abstract: This paper describes research on the effects of PoS tagging as a preprocess for HPSG-based deep parsing in the context of an open-source Spanish treebank development in the DELPH-IN framework. The treebank annotation is performed by hand selecting the proper decisions among the choices proposed by the system and ranked by a statistical module. The presented experiments show that the use of a tagger lowers the ambiguity of the sentences, both reducing the amount of sentences that reach time-out before the entire parse forest is built, and helping the ranker to place the right tree among the n-best trees. On the one hand, our results validate the benefits –already reported in the literature– of such preprocess to deep parsing with regard to speed, coverage, and accuracy. On the other hand, we propose a strategy based on existing open-source tools and resources to develop highly-consistent deep-annotated treebanks for languages with limited availability of linguistic resources.
Keywords: Deep syntax treebank annotation, HPSG parsing, PoS tagging.

1 Introduction

Linguistically interpreted natural language texts constitute a crucial resource both for theoretical linguistic investigations about language use and for practical NLP purposes. Thus, in recent years, there has been an increasing interest in the construction of treebanks and, nowadays, both theory-neutral and theory-grounded treebanks have been developed for a great variety of languages.[1]

While first efforts in treebank building used manual annotation, recent significant advances in the development of large-scale robust efficient grammars and hybrid statistical/symbolic approaches for resolving ambiguities have made it possible to use sophisticated linguistic handcrafted deep-syntax frameworks (such as HPSG or LFG) to support the annotation task (Riezler et al., 2002; Prins and van Noord, 2003; Toutanova et al., 2005).

However, a drawback of these approaches is that their detailed granularity produces a huge ambiguity, creating efficiency problems to the parser machinery and making the effective use of the results difficult. Ambiguity not only slows down processing, but it also impoverishes the grammar performance in terms of coverage due to time-out problems when parsing long sen-

[1] Some of these treebanks are presented in (Hinrichs and Simov, 2004).

ISSN 1135-5948

tences. Besides, it also leads to negative effects on parsing accuracy, caused by the combinatorial explosion of the search space. In the case of treebank development, this may represent a severe slow down –with consequent cost increase– and that some sentences can not be parsed and must be excluded from the resource. Thus, strategies must be devised to extend the coverage and the efficiency of deep parsers used in treebank development. Such strategies should rely on basic existing state-of-the-art resources (e.g. a PoS tagger) in order to be potentally appliable to the development of deep syntax annotated corpus for a wide range of languages.

Many research lines have been pursued to improve the performance of deep parsers, most of them relying on hybrid systems that combine shallow and deep NLP paradigms. The prime motivation for most of the published hybrid directions was to improve the efficiency of the parsers (Bangalore et al., 1997; Bangalore and Joshi, 1999; Ciravegna and Lavelli, 1997; Watanabe, 2000; Grover and Lascarides, 2001; Marimon, 2002; Crysmann et al., 2002; Prins and van Noord, 2003; Daum, Foth, and Menzel, 2003; Frank et al., 2003; Clark and Curran, 2004; Zhang, Matsuzaki, and Tsujii, 2009). Besides, some of the integrated shallow-deep processing also showed improvements in the robustness (Marimon, 2002; Crysmann et al., 2002; Daum, Foth, and Menzel, 2003; Adolphs et al., 2006) and the precision (Prins and van Noord, 2003; Daum, Foth, and Menzel, 2003; Sagae, Miyao, and Tsujii, 2007) of rule-based symbolic grammars.

As for the level of shallow information that the hybrid architectures integrate to achieve their goals, it ranges from simple morphological information and PoS information to different shallow syntactic analysis.

The works by Grover and Lascarides (2001) and by Prins and van Noord (2003) are two examples of the benefits of using basic PoS information. Grover and Lascarides (2001) interface PoS tag information with the existing lexicon of the Alvey Natural Language Tools system: if a word exists in the lexicon, the PoS tag is used as a filter, accessing only those entries of the appropriate category, if the word is unknown to the system, a basic underspecified entry for the PoS tag is used as its lexical entry. An experiment with 200 sentences shows how performance improves a 37.5%, with a precision of 30.5%. Prins and van Noord (2003) show how a HMM n-gram PoS tagger can be used to filter unlikely lexical categories to increase the speed of a parsing system based on a wide-coverage HPSG for Dutch. Experimental results with a test set of 216 sentences show that the use of the tagger greatly reduces parsing time, and, in addition, yields an increase of parsing precision.

Other proposals have extended the integrated information and exploit shallow syntactic analysis as produced by different shallow tools. In this line, Bangalore et al. (1997) present a system which applies a statistical disambiguation technique prior to parsing in the LTAG framework. A (trigram) disambiguation model is used to disambiguate so-called supertags, tags that represent the syntactic behavior of words and have a 1-to-1 mapping with the grammar lexical types. The task of the parser is thus reduced to establish the dependency links, with a parsing speed-up of about factor 30, with a tag accuracy of 68%. Later experiments reported in (Bangalore and Joshi, 1999) improve the tag accuracy to 92% by using much larger amount of training data and adding some smoothing techniques. The benefits of suppertagging in parsing speed has also been demonstrated in other lexicalised formalisms like CCG (Clark and Curran, 2004) and HPSG (Zhang, Matsuzaki, and Tsujii, 2009; Dridan, 2009). In another line of research, Ciravegna and Lavelli (1997) propose to use text chunking for controlling an agenda-based bottom-up chart parser; preliminary text chunking allows them to focus directly on the constituents that seem more likely, reducing the spurious ambiguity. The chunking process is done via finite state automaton, taking the output of a PoS tagger. They claim that experiments show a reduction of about 68% of constituents generated and of 78% of time consumed. Frank et al. (2003) combine macro-structural constraints derived from a probabilistic topological field parser for German with a constraint-based HPSG parser and report a performance gain of factor 2.25 on a set of 5060 sentences. Watanabe (2000) describes an algorithm for accelerating the CFG-Parsing process by using dependency information provided by stochastic parsers, interactive systems and linguistic annotations added in the source text. Reported reduction of processing time is about 45% and 15%. And, more recently, Sagae, Miyao, and Tsujii (2007) combine dependency and HPSG parsing and report a 1% absolute improvement in precision and recall of predicate-argument identification in HPSG parsing over a strong baseline.

The contribution of more than one shallow

component to the performance of deep analysis has also been investigated, though to a smaller extend. Marimon (2002) integrates a cascade of shallow components performing PoS tagging and chunk recognition as a pre-processing module of a HPSG-based grammar of Spanish implemented in the ALEP system. Experimental results show that the efficiency of the overall analysis improves an average of 65% and that the system also provides robustness to the linguistic processing, while maintaining both the recall and the precision of the grammar. The same approach is used by Crysmann et al. (2002) within the LKB system, where they use partial analyses from shallow processing to guide the deep parser to identify relevant candidates for deep processing. Also, Daum, Foth, and Menzel (2003) investigate the contributions of both taggers and chunkers to the performance of a deep syntactic parser with a Weighted Constraint Dependency Grammar of German and report to achieve a high degree of lexical robustness, reduced run time requirements, and a considerably improved parsing accuracy on a set of 1845 sentences.

This paper describes research on the effects of a state-of-the-art PoS tagger in deep parsing of unrestricted Spanish text, carried out in the context of on-going work for the creation of a new open-source resource for Spanish –an HPSG-based treebank called Tibidabo–. We focus on investigating to what extend using a tagger affects the system results both in terms of *coverage* (measured as the percentage of sentences for which it produces an output in the allocated time) and *accuracy* (measured as the percentage of sentences for which the right parse tree is ranked among the best ones). Additionaly, our research contributes to validate the benefits of a PoS tagger on parsing speed already reported in the literature.

Note that, being our goal to build a treebank, the preprocess must rely on existing state-of-the-art tools and we can not resort to more sophisticated techniques –e.g. supertagging– due to the lack of training material.

The following two sections summarize the set-up and motivation of our research. Section 3.1 describes experiments on the influence of tagging on deep parsing of unrestricted Spanish text, and section 5 presents some conclusions and some directions for future work.

2 The Annotation Environment

As we have already mentioned, the research we describe in this paper is carried out in the context of on-going work for the construction of a new open-source language resource for Spanish: an HPSG-based treebank.[2]

Our investigation uses the DELPH-IN open-source tools for writing and processing HPSG grammars and the DELPH-IN publicly available Spanish Resource Grammar.[3]

The treebanking environment in the DELPH-IN framework is based on the selection[4] of the correct analysis among all the analyses that are produced by a symbolic grammar, instead of using only human annotation. It also provides a Maximum Entropy (ME) based stochastic learner (Toutanova et al., 2005) that observes decisions taken by the annotators and applies the same in unseen parses to reduce the outputs generated by the grammar and, therefore, the manual annotation effort in treebanking even with long sentences.

Nevertheless, some sentences still can not be included in the treebank due to: (a) the parser can not build the complete parse forest in the allocated time and exits with a time-out, or (b) the parser generates a large number of possible analysis and the right one is not ranked between the solutions offered to the annotator.

We will study whether the use of a PoS tagger reduces the timed-out sentences and whether it increases the number of sentences for which the right analysis is present among those ranked best by the statistical component.

2.1 Parser and Grammar

The Spanish Resource Grammar is a broad-coverage precise grammar for Spanish that aims at full parsing of unrestricted text.

The grammar is implemented on the *Linguistic Knowledge Builder* (LKB) system –an interactive grammar development environment for typed feature structure grammars– (Copestake, 2002).

The Spanish Resorce Grammar is grounded in the theoretical framework of HPSG (Pollard and Sag, 1987; Pollard and Sag, 1994), a constraint-based lexicalist approach to grammatical theory where all linguistic objects (i.e., words and phrases) are represented as typed feature structures, and they use the *Minimal Recursion Semantics* (MRS) semantic representation (Copes-

[2]The current treebank version is already publicly available within the DELPH-IN framework.

[3]See http://www.delph-in.net/.

[4]Selection is done by rejecting (or, alternatively, selecting) the lexical items and grammar rules that originate the multiple parses to incrementally disambiguate the sentence until a single analysis is left.

take et al., 2006). Using unification of typed feature structures, the MRS representation assigns a syntactically flat semantic representation to linguistic expressions which offers, by means of labeling of arguments and their co-indexation, a list of semantic relations and a set of syntactic limitations on possible scope relations among them.

The Spanish Resource Grammar has a full coverage lexicon of closed word classes (pronouns, determiners, prepositions and conjunctions) and it contains about 50,000 lexical entries for open word classes.[5] These lexical entries are defined by a set of about 500 lexical types that represent the type of words in the lexicon. Following well-established theoretical HPSG proposals, these lexical types are organized into a multiple inheritance type hierarchy (i.e., subtypes may inherit properties from more than one supertype higher in the hierarchy) allowing for lexical generalizations shared by several subtypes to be captured only once. The grammar also has 70 lexical rules to perform valence changing operations on lexical items (e.g. movement and removal of complements) which reduces the number of lexical entries to be manually encoded in the lexicon, and 230 phrase structure rules to combine words and phrases into larger constituents and to compositionally build up the semantic representation.

The Spanish Resource Grammar deals with a wide range of constructions in Spanish, including: main clauses with canonical word order surface and word order variations, valence alternations, determination, agreement, null-subject, compound tenses and periphrastic forms, raising and control, passives, (basic) comparatives and superlatives, all types of relative clauses, unbounded dependency constructions, cliticization phenomena, constructions with *se*, coordination, and nominal and verbal ellipsis.

2.2 PoS tagger

In our system, before parsing input sentences, raw text is pre-processed by FreeLing, an open-source language analysis tool suite performing shallow processing functionalities (Padró et al., 2010).[6]

FreeLing receives a sentence, morphologically annotates each word by dictionary look-up, and performs state-of-the-art HMM disambiguation, with an estimated accuracy around 97%. The morphological analysis step includes the application of a cascade of specialized processors that annotate punctuation symbols, multi-words, numerical expressions, date/time expressions, ratios, percentages, monetary amounts, and proper nouns.[7]

The integration of FreeLing is done using the LKB *Simple PreProcessor Protocol* (SPPP) which maps PoS tags into partial feature structures.[8] This SPPP interfacing module allows the definition of some adaptation rules aiming to ensure the smooth integration of both tools and to provide the best balance between parsing efficiency and accuracy. For instance, a list of words or tags causing ambiguities not solved with high reliability by the HMM tagger (like the ambiguity pronoun-conjunction of the word *que* (that), or proper names at sentence beginning) can be specified. For those words and tags, the PoS tagger decisions will be ignored (no analysis will be discarded) when found at the specified position, passing all possibilities to the deep parsing to be resolved by the symbolic grammar.

Also, this interfacing module can be configured with a list of substitutions of certain categories in FreeLing output by the category expected by the grammar. In this way, we avoid parsing failures due to discrepancies in the FreeLing tagset and the lexical categories assumed by the Spanish Resource Grammar (this is the case, for instance, of deictic adverbs like *here*, *there*, *today*, *tomorrow*, etc., which FreeLing tags as adverbs while the grammar lexicon encodes them as pronominal signs).

2.3 Target corpus

To create the treebank Tibidabo we chose newspaper text we borrowed from the corpus AnCora, a corpus of 528,000 words (17,363 sentences) (Taulé, Martí, and Recasens, 2008). Table 1 shows the number of sentences and ratio distributed along the sentence length.

Although the AnCora corpus already provides syntactic annotation, semantic roles, coreference, and other linguistic markup similar to what a deep analysis framework as HPSG and MRS can

[5] The grammar also includes a set of generic lexical entry templates for open classes to deal with unknown words for virtually unlimited lexical coverage.

[6] The FreeLing toolkit may be downloaded from: http://nlp.lsi.upc.edu/freeling.

[7] FreeLing also includes a guesser to deal with words which are not found in the lexicon by computing the probability of each possible PoS tag given the longest observed termination string for that word.

[8] SPPP assumes that a pre-processor runs as an external process to the LKB system and communicates with its caller through its standard input and output channels. See http://wiki.delph-in.net/moin/LkbSppp.

Sentence length	*# sentences*	*% of the corpus*
1-5	872	5.02
6-10	1,420	8.17
11-15	1,877	10.81
16-20	2,029	11.62
21-25	2,051	11.81
26-30	1,987	11.44
31-35	1,871	10.77
36-40	1,701	9.79
41-45	1,318	7.59
46-50	997	5.74
51+	1,246	7.17
Total	17,363	100

Table 1: Distribution of sentence lengths in the corpus.

offer, the annotations are hand created. Even if a thorough metodology and detailed criteria are used, human annotators are error-prone or may misinterpret the criteria. Any human-annotated resource unavoidably suffers from a certain degree of error or inconsistent criteria due to this fact.

We believe that providing a corpus consisting of the same text annnotated under a different paradigm –where the annotation criteria are enforced by a deep analysis lexical grammar instead of human annotators– may be a valuable resource for research. Such corpus can be useful in studying the variability of human annotation, the ability of machine learing algorithms to capture the structures annotated in each approach, the study of how different linguistic criteria can be mapped to each other, among many other possibilities.

3 Experimental Setting

A rough idea of the coverage of the current version of the grammar may be drawn from the fact that about 30.4% of the sentences of up to 50 words receive at least a full parse.[9]

Parsing failures in the remaining 70% of sentences are basically due to two reasons. First, the processing components –as any other complex software in development stage– certainly show some deficiencies –lack of coverage, errors and unanticipated interactions, lack of robustness– that are responsible for 12.2% of the parsing failures. Second, 57.4% of the input sentences reach time-out limit set in the parsing engine (which was set at 60 seconds per sentence), because they get a too large number of analyses. The failure ratio due to time-out limit increases considerably with longer sentences (see table 3), which clearly shows up the need for improving the efficiency of the system to enable parsing of unrestricted Spanish text.

The 30% of sentences up to 50 words that receive at least a full parse get, in fact, an average of 5,040 parses/sentence. This amount of possible trees requires too many reject/select decisions by the human annotator, increasing the difficulty of the task and dramatically slowing down the treebank construction. To palliate this, the stochastic ranker in the DELPH-IN framework is trained and used to select a reduced number of parse trees to be presented to the annotator, thus reducing the number of decisions needed to disambiguate the sentence. Nevertheless, a huge amount of possible parses poses a more difficult challenge to the ranker, and the right tree may not always be among those selected.

Both the time-out problem and the large number of trees the ranker has to deal with reduce the number of sentences that can be annotated and included in the treebank. Thus, overcoming these issues is a crucial step to build a complete and useful resource.

Since lexical ambiguity is a cause shared by both problems, our approach is to use a PoS tagging preprocessor that reduces the ambiguity the parser has to deal with. In the following section we present two experiments that measure the influence of tagging on both the efficiency of the parser –which assigns (multiple) analyses to input sentences– and the accuracy of the ranking model –which chooses the best ones among them.

3.1 Corpus Ambiguity

Before reporting the results of our experiments, we present some statistics on the morphological, lexical, and syntactic ambiguities in the corpus.

We denote as *morphological ambiguity* the PoS ambiguity that is typically addressed by a tagger. Table 2 shows a summary of the morphological ambiguities (tags per word) in the corpus.

	Ambig. words		*All words*	
	# words	*tg/w*	*# words*	*tg/w*
open-class	83,000	2.30	235,000	1.46
closed-class	144,000	2.62	293,000	1.80
Total	227,000	2.46	528,000	1.63

Table 2: Morphological ambiguity profiles of the corpus.

[9]Longer inputs can not be parsed within established time-out limits.

The Spanish grammar implemented in the DELPH-IN system is grounded in the theoretical framework of HPSG, a heavily lexicalist approach to grammatical theory where words are assigned many lexical classes that differ, for example, in the valence frame.[10] We denote as *lexical ambiguity* the average number of lexical classes per word that the parser takes into account. In the case of our corpus, it is 7.0 lexical classes per word.

Given an input sentence, the parser considers, for each word, all lexical classes matching the valid PoS tags for that word. Then, all possible parses consistent with those possibilities are built, producing a large amount of full syntactic analyses. We denote as *syntactic ambiguity* the average amount of possible full parses per sentence generated by the parser.

The average syntactic ambiguity for the 30% of sentences up to 50 words that get some analysis, is 5,040 parses/sentence.

4 Experiments

4.1 Experiment 1: Influence on Coverage

To investigate the effects of the PoS tagger on the efficiency of the Spanish grammar we parsed the whole corpus with and without the tagger and compared the system performance.

Table 3 shows the ratio of sentences that received at least a full parse, as well as the percentage of sentences for which the parser timed out, distributed along the sentence length.[11]

Not surprisingly, due to tagging errors, PoS tagging caused a small loss in the number of short sentences receiving an analysis: A 3% less of sentences under 10 words were analyzed, but since there are relatively few sentences in that range, this represents a loss of only 0.4% over the whole corpus. However, the tagger certainly had a positive impact on longer sentences –with lengths between 11 and 40 words– where the observed coverage increase was 7.2%. Note that sentences in this length range constitute two thirds of the whole corpus. Thus, the overall ratio of sentences in the whole corpus that received an analysis increased in 6.9%.

The PoS tagger reduced the morphological ambiguity from 1.63 to 1.03 tags/word,[12] which reduced from 7.0 to 4.7 lexical classes per word

Sent. length	% of corpus	*Parsed sentences*		*Timeout ratio*	
		no tag	*tag*	*no tag*	*tag*
1-5	5.0	91.4	87.4	0	0
6-10	8.2	89.2	86.6	2.6	0.8
11-15	10.8	73.8	74.3	12.6	4.6
16-20	11.6	49.9	61.2	34.3	10.6
21-25	11.8	25.2	38.0	58.7	34.8
26-30	11.5	10.3	21.1	74.4	50.3
31-35	10.8	3.5	9.2	82.0	61.6
36-40	9.8	1.2	3.4	86.6	71.5
41-45	7.6	0.5	1.0	88.8	75.2
46-50	5.7	0.2	0.2	89.6	77.3
51+	7.2	0	0	100.0	100.0
Total	100.0	30.4	37.3	57.4	42.6

Table 3: Percentages of parsed and timed-out sentences.

the lexical ambiguity the parser has to deal with. This caused the parser to build less constituents not contributing to the final parse, making it possible to parse 7% more sentences, for which the parser timed-out before. The syntactic ambiguity when using PoS tagging slightly increased (from 5,040 to 5,434 analysis/sentence) due to the fact that longer sentences –which are more ambiguous– that were not parsed before are now included in the count.

As we expected, morphological disambiguation also had a positive impact on parsing time and reduced average processing time from 38.4 to 30.4 sec/sentence (even when longer sentences are now included in the count).

4.2 Experiment 2: Influence on Accuracy

To evaluate the impact of tagging on the accuracy of the ME ranking model, we calculated the ratio of sentences for which the parse in the gold standard is ranked among the n best by the stochastic model. Note that an output analysis includes both a phrase structure tree and a MRS semantic representation, and that exact match is required.

The corpus used in this experiment was a small part of the whole treebank, consisting of 2,570 sentences of lengths up to 16 words. The experiment was performed using 5-fold cross-validation.

Table 4 shows the accuracy of the parse selection model (percentage of sentences for which the right full parse tree was ranked by the model

[10] For example, the average numbers of entries per verb is 1.84, however, some verbs have as many as 8 lexical entries.

[11] The corpus was parsed with a Quad-Core 2.83GHz with 8Gb RAM.

[12] In a few cases where the tagger has large error rates, such as the conjunction/relative ambiguity for the word *que*, the tagger output is ignored and the ambiguity maintained.

among the n best) when the used treebank was parsed either without or with the tagger. Differences are significant at a 95% confidence degree according to a paired t-test for all values of n except $n = 10$.

n-best	no tagger	tagger
1	54.8%	56.3%
2	65.1%	66.9%
3	70.8%	73.0%
4	74.1%	77.0%
5	78.1%	79.5%
10	85.7%	85.6%
20	91.4%	89.6%
30	94.7%	91.6%

Table 4: Accuracy of the parse model selection model for n-best analyses with and without the tagger.

The reason why the use of a PoS tagging improves parsing accuracy is that it reduces the number of candidate analyses, largely reducing the search space that the selection model has to deal with. For this set of sentences, the grammar assigned an average of 7.7 lexical classes per word and produced an average of 1,235 parse trees per sentence. If PoS tagging is used, these figures are reduced to 4.3 classes per word and 606 analyses per sentence.

5 Conclusions and Future Work

This paper describes research that shows the usefulness of PoS tagging for improving speed, coverage, and accuracy of HPSG-based deep parsers used in treebank development. A first experiment shows that, by improving parsing speed, PoS tagging increases parsing coverage for long sentences. The second experiment shows a statistically significant improvement in parsing accuracy when using a Maximum-Entropy based model to rank the n-best analysis for each sentence.

Presented results show a 14.8% decrease of timed-out sentences (from 57.4% to 42.6%) consisting of a 6.9% of coverage increase, plus a 7.9% of sentences that no longer time out but are not yet analyzed due to tagger errors or to the lack of the appropriate rules in the grammar.

They also show that the use of a PoS tagger yields a significant increase in the percentage of sentences for which the right tree is ranked among the best ones by the statistical module.

The presented results are most informative in order to design an optimal annotation strategy aiming to maximize the annotation speed while maintaining high levels of accuracy: about 50% of the sentences in the corpus can be annotated using the tagger and setting the ranker to select a small number of trees. Given this reduced forest, the annotation process is very fast, since each sentence requires only a few annotator decisions to be disambiguated. The 20% of sentences that have not been annotated (e.g. because the right tree was not among those proposed) can be processed again with a higher number of candidate trees. The remaining sentences can be annotated at a slower rate without the tagger. Finally, a higher time-out can be set to have the parser analyze sentences where it timed out before, and repeat the process. This strategy is based on existing open-source tools and resources[13], and makes it possible to develop highly-consistent deep-annotated treebanks for languages with limited availability of linguistic resources.

Future work will include the extension of the grammar coverage, the extension of the treebank. We will also study the viability of training a high-precision ranker that allows the automatic annotation of a large number of sentences in the treebank.

Acknowledgments

This work has been partially funded by the European Union through project X-LIKE (FP7-ICT-2011-288342), by the Spanish Government through the programe *Ramón y Cajal* and the project KNOW2 (TIN2009-14715-C04-03/04), and by the Catalan Government via the mobility programe *Beques per a estades per a la recerca fora de Catalunya*.

References

Adolphs, P., S. Oepen, U. Callmeier, B. Crysmann, D. Flickinger, and B. Kiefer. 2006. Some fine points of hybrid natural language parsing. In *Proceedings of the 5th Internatinal Conference LREC*, Genoa, Italy.

Bangalore, S., C. Doran, B.A. Hockey, and A. Joshi. 1997. An approach to robust partial parsing and evaluation metrics. In *Proceedings of the 5th International Workshop on Parsing Technologies*, Boston, MA.

Bangalore, S. and A. Joshi. 1999. Supertagging: An approach to almost parsing. *Computational Linguistics*, 2(25):237–265.

[13] All the used DELPH-IN resources –the software, the SRG grammar, and the treebank– as well as the FreeLing toolkit are licensed under GPL or LGPL.

Ciravegna, F. and A. Lavelli. 1997. Controlling bottom–up chart parsers though text chunking. In *Proceedings of the 5th International Workshop on Parsing Technologies*, Boston, MA.

Clark, S. and J.R. Curran. 2004. The importance of supertagging for wide-coverage CCG parsing. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*, Geneva, Switzerland.

Copestake, A. 2002. *Implementing Typed Feature Structure Grammars*. CSLI Publications, Stanford.

Copestake, A., D. Flickinger, C.J. Pollard, and I.A. Sag. 2006. Minimal recursion semantics: an introduction. *Research on Language and Computation*, 3(4):281–332.

Crysmann, B., A. Frank, B. Kiefer, S. Müller, G. Neumann, J. Piskorski, U. Schäfer, M. Siegel, H. Uszkoreit, F. Xu, M. Becker, and H.U. Krieger. 2002. An integrated architecture for shallow and deep processing. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Pittsburgh.

Daum, M., K.A. Foth, and W. Menzel. 2003. Constraint based integration of deep and shallow parsing techniques. In *Proceeding of the 10th Conference of the EACL*, Budapest.

Dridan, R. 2009. Using lexical statistics to improve HPSG parsing. Master's thesis, Saarland University, Sarbrücken, Germany.

Frank, A., M. Becker, B. Crysmann, B. Kiefer, and U. Schäfer. 2003. Integrated shallow and deep parsing: Topp meets HPSG. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*.

Grover, C. and A. Lascarides. 2001. XML-based data preparation for robust deep parsing. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, Toulouse, France.

Hinrichs, E.W. and K. Simov, editors. 2004. *Research on Language and Computation*, volume 2(4). Kluwer Academic Publishers.

Marimon, M. 2002. Integrating shallow linguistic processing into a unification-based spanish grammar. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan.

Padró, L., M. Collado, S. Reese, M. Lloberes, and I. Castelón. 2010. Freeling 2.1: Five years of open-source language processing tools. In *Proceedings of 7th Language Resources and Evaluation Conference (LREC'10)*, La Valletta, Malta.

Pollard, C.J. and I.A. Sag. 1987. *Information-Based Syntax and Semantics, Volume 1: Fundamentals*. CSLI Lecture Notes, Stanford University, CA.

Pollard, C.J. and I.A. Sag. 1994. *Head-driven Phrase Structure Grammar*. The University of Chicago Press and CSLI Publications, Chicago.

Prins, R. and G. van Noord. 2003. Reinforcing parser preferences through tagging. *Special issue on Evolutions in Parsing of the journal Traitement Automatique des Langues 44(3)*, pages 121–139.

Riezler, S., T.H. King, R.M. Kaplan, R. Crouch, J.T. Maxwell, and M.Johnson. 2002. Parsing the wall street journal using a lexical-functional grammar and discriminative estimation techniques. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, Philadephia, PA.

Sagae, K., Y. Miyao, and J. Tsujii. 2007. HPSG parsing with shallow dependency constraints. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic.

Taulé, M., M.A. Martí, and M. Recasens. 2008. AnCora: Multilevel annotated corpora for catalan and spanish. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC-2008*, Marrakech, Morocco.

Toutanova, K., C.D. Manning, D. Flickinger, and S. Oepen. 2005. Stochastic HPSG parse disambiguation using the redwoods corpus. *Journal of Logic and Computation*.

Watanabe, H. 2000. A method for accelerating CFG-parsing by using dependency information. In *Proceedings of the 18th International Conference on Computational Linguistics*, Saarbrücken, Luxembourg, Nancy.

Zhang, Y.-Z., T. Matsuzaki, and J. Tsujii. 2009. HPSG supertagging: A sequence labeling view. In *Proceedings of the 11th International Conference on Parsing Technology (IWPT'09)*, Paris, France.

Aprendizaje Automático en PLN

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 149-156 recibido 04-05-12 revisado 03-06-12 aceptado 06-06-12

Detección de la polaridad en citas periodísticas: una solución no supervisada *

A non supervised method for sentiment polarity detection on reported speech from news

A. Montejo-Ráez, E. Martínez-Cámara, M. T. Martín-Valdivia, L. A. Ureña-López
Departamento de Informática
Universidad de Jaén
Las Lagunillas s/n, Jaén - 23071
{amontejo,emcamara,maite,laurena}@ujaen.es

Resumen: El presente trabajo expone los resultados alcanzados mediante un método no supervisado para la detección de la polaridad en textos relativos a citas aparecidas en noticias en inglés, correspondientes al corpus 2010_JRC_1590_quotes. Este método, basado en la obtención de un subgrafo de WordNet obtenido mediante el algoritmo Page Rank y su ponderación mediante los valores correspondientes en SentiWordNet, propone una solución no supervisada que ofrece unos resultados competitivos sobre algunas técnicas actuales.
Palabras clave: Análisis de emociones, clasificación de la polaridad, SentiWordNet, Page Rank

Abstract: This papers shows the results obtained by a non supervised method in the task of sentiment polarity detection on news quotations (reported speech) from the 2010_JRC_1590_quotes corpus. This method, which obtains a subgraph of WordNet by applying Page Rank over it, weights such a graph with corresponding SentiWordNet polarity scores, offering a non-supervised solution comparable to state-of-the-art techniques.
Keywords: Sentiment Analysis, polarity classification, SentiWordNet, Page Rank

1. Introducción

El análisis de opiniones (*Sentiment Analysis)* estudia el tratamiento de opiniones en textos. Este estudio, enmarcado dentro del Procesamiento del Lenguaje Natural (PLN), está suscitando un interés creciente en la comunidad investigadora por diversas razones. Principalmente, el análisis de la ingente cantidad de información que genera la Web 2.0, donde cada usuario en Internet es un generador potencial de información a través de comentarios, bitácoras, envíos por microblogging (como Twitter) y una larga lista de posibles escenarios en los que expresar una opinión con fuerte carga subjetiva. Poder analizar de forma automática estas publicaciones permite "tomar la temperatura" de un amplio grupo de usuarios acerca de temáticas diversas, o determinados productos. Es por esto que el análisis de opiniones (AO) es objeto de interés por parte de los investigadores como de las empresas, pues permite abrir una ventana a la opinión en la Red, mas allá de las encuestas explícitas.

La presencia de opiniones no solo se restringe a las plataformas 2.0 que pueblan Internet, sino que también se encuentran opiniones en ámbitos más profesionales, como puede ser el periodístico o crítica especializada (literaria, cinematográfica, gastronómica ...). El tratamiento de estas opiniones también interesa tanto al mundo académico como al empresarial, ya que debido a su carácter profesional pueden crear opinión, que luego se transmite a través de las redes sociales.

Dentro del AO, la clasificación de *polaridad* simplifica este problema como sigue: dado un texto t, el objetivo es obtener una puntuación que nos indique si el texto expresa

* Esta investigación ha sido subvencionada parcialmente por el proyecto del Instituto de Estudios Giennenses Geocaching Urbano (RFC/IEG2010), el Fondo Europeo de Desarrollo Regional (FEDER), a través del proyecto TEXT-COOL 2.0 (TIN2009-13391-C04-02) por el gobierno español, y por la Comisión Europea bajo el Séptimo programa Marco (FP7 - 2007-2013) a través del proyecto FIRST (FP7-287607)

una opinión positiva o negativa, dentro de un rango donde 0 indicaría una carga subjetiva neutra, 1 una carga subjetiva positiva y -1 una carga subjetiva negativa:

$$p(t) \in [-1, 0, 1] \quad (1)$$

El método propuesto es capaz de calcular este valor a partir de un texto sin necesidad de utilizar conocimiento del dominio, conjunto de datos etiquetados y modelos previamente entrenados, sino ponderando el uso conjunto de las bases de conocimiento de WordNet y SentiWordNet mediante los valores obtenidos por el algoritmo Page Rank (Page et al., 1998).

El artículo se estructura como sigue. Primero mencionamos algunos de los trabajos más recientes en clasificación de la polaridad. El siguiente punto describe el método propuesto y sus fases. Después pasamos a detallar la configuración de los experimentos y a comentar los resultados obtenidos. Finalmente, cerramos el trabajo con unas reflexiones finales e información sobre los aspectos que darán continuidad a esta investigación.

2. Estado del arte en la clasificación de polaridad

En la resolución de un problema mediante aprendizaje automático se pueden seguir principalmente dos estrategias: supervisada y no supervisada. Ésta dicotomía se encuentra muy presente en la literatura sobre AO, aunque las técnicas supervisadas están más representadas que las no supervisadas. El trabajo que se suele tomar como referencia para los métodos supervisados es (Pang, Lee, y Vaithyanathan, 2002), el cual utiliza como característica la presencia o no de los términos para el cálculo de la polaridad. En cambio, (Martínez-Cámara et al., 2011) utilizando el algoritmo SVM como (Pang, Lee, y Vaithyanathan, 2002) pero sobre un corpus de críticas de cine en español, obtiene mejores resultados representando los unigramas con el valor TF-IDF.

En la estrategia supervisada no solo se han utilizado características léxicas, sino que también se ha intentado explotar la información sintáctica. Un ejemplo de ello son los trabajos (Mullen y Collier, 2004) y (Whitelaw, Garg, y Argamon, 2005), en los que se utilizan los adjetivos como características a la hora de la clasificación de la polaridad. El tratamiento de la negación también se ha tenido en cuenta en los experimentos supervisados, siendo (Wiegand et al., 2010) un buen resumen de lo publicado en el aprovechamiento de esta característica. No es el objetivo de este artículo ser un listado de todos los trabajos que siguen un enfoque supervisado, por lo que si el lector quiere una mayor información puede consultar (Pang y Lee, 2008), (Liu, 2010), (Tsytsarau y Palpanas, 2012).

Los métodos supervisados adolecen de la necesidad de disponer de conjuntos de datos etiquetados, y de la dificultad de adaptación de las técnicas entre dominios distintos. Éstas son las principales razones que llevan a los investigadores a estudiar métodos no supervisados o semi-supervisados, que no requieren de datos etiquetados para la construcción de un modelo de clasificación, y que su aplicación a distintos dominios sea mucho más sencilla.

Los métodos no supervisados en AO suelen fundamentarse en la detección de identificadores de subjetividad u opinión en los textos, y posteriormente calculan la polaridad aplicando alguna función basada en los indicadores encontrados. Para dicha identificación se suelen utilizar conjuntos de vocablos etiquetados por su polaridad. Existen tres estrategias para la elaboración de estos lexicones: manual, basado en diccionario y basado en corpus. El enfoque manual es bastante costoso, por lo que se suele aplicar como último paso de los enfoques automáticos a modo de revisión. El método basado en diccionario construye un lexicón etiquetado por medio de la combinación de un conjunto de palabras de semilla, y el uso de recursos léxicos, como es el caso de WordNet, para la aplicación de la lista inicial (Kim y Hovy, 2004), (Hu y Liu, 2004). Los métodos basados en diccionario generan listas de indicadores de opinión genéricas, que no dan buenos resultados en dominios muy específicos. Los métodos basados en corpus se adaptan mejor al dominio en el que trabajan, ya que se basan en las características propias del corpus para la ampliación del lexicón de palabras semilla inicial (Hatzivassiloglou y McKeown, 1997), (Kanayama y Nasukawa, 2006).

Los métodos anteriores más que no supervisados deberían considerarse semi-supervisados, debido a que necesitan de un cierto conocimiento inicial, es decir, una lista de indicadores de opinión etiquetada

para calcular la polaridad. Además, aunque el método basado en diccionario es el más genérico, tiene una cierta dependencia de los términos semilla que se hayan seleccionado. Reduciendo al mínimo el conocimiento inicial, y por lo tanto ganando capacidad de adaptación del dominio, se encuentran los métodos que basan la clasificación en cálculos estadísticos, y en las relaciones semánticas de los términos. Como referencia de este enfoque se encuentra el trabajo de Turney en 2002 (Turney, 2002). Actualmente, algunos métodos (Wu et al., 2010) intentan resolver el problema de la adaptación al dominio mediante un mapeado de conceptos comunes, buscando un espacio dimensional más allá del contenido léxico (Ji et al., 2011), explorando conexiones entre grafos (Wu, Tan, y Cheng, 2009) o mediante métodos probabilísticos (Tan et al., 2009).

La metodología seguida en la generación del corpus Emitoblog propone la generación de modelos de alta granularidad para el AO sobre distintos tipos de textos, propios de publicaciones coloquiales en el marco de la Web 2.0 (Boldrini et al., 2012). Este trabajo demuestra que es posible la transferencia de dominio de modelos entrenados sobre ciertos conjuntos de datos a otros textos de dominio diferente, gracias a la independencia relativa que permiten características detalladas acerca de los sentimientos. En cualquier caso, es un solución que necesita de entrenamiento, por lo que en mayor o menor medida seguirá viéndose afectada por las limitaciones en la transferencia de sentimientos según dominio.

Por otra parte, SentiWordNet (Baccianella, Esuli, y Sebastiani, 2008) es un recurso léxico construido a partir de otro recurso: WordNet (Fellbaum, 1998). Proporciona información acerca de la orientación semántica en cuanto a emoción de los "synsets". Un synset es el ítem de información básico en WordNet, y representa un "concepto", sin ambigüedad. La mayoría de las relaciones en este grafo léxico trata los synsets como nodos: hiperonimia, sinonimia, homonimia, antonimia... SentiWordNet devuelve, para cada synset, una tripleta de tres valores que miden la carga de "positividad", "negatividad" u "objetividad" del mismo. La última versión de SentiWordNet (3.0) se ha generado a partir de las anotaciones manuales de versiones previas, propagando sobre el grafo dichos valores de emoción mediante un algoritmo de tipo *random walk*. Este recurso ha sido utilizado en AO y representa una fuente de información independiente de dominio (Denecke, 2008; Ogawa, Ma, y Yoshikawa, 2011).

3. Expansión de conceptos y pesado SWN

El método propuesto permite la expansión de conceptos mediante un algoritmo de tipo Page Rank (Page et al., 1998) sobre el grafo de WordNet, multiplicando los valores obtenidos de los synsets por los pesos de polaridad de SentiWordNet. Con más detalle, el procesamiento es el siguiente:

1. Los textos son procesados, haciendo uso de la biblioteca NLKT[1], para extraer los lemas y el *part of speech* (POS) de los mismos.
2. Se construyen los "contextos" para cada frase que son pasados a la herramienta UKB[2] para el cálculo, mediante un algoritmo de tipo Random Walk(Agirre y Soroa, 2009), que al mismo tiempo que desambigua los términos, genera los synsets con mayor peso resultado del proceso iterativo de propagación propio de estos algoritmos.
3. Expandimos el texto reemplazando los términos por sus vectores PPV (*Personalized Page Rank vectors*), que consisten en una secuencia de synsets de WordNet.
4. Los synsets, con los pesos asociados, son mapeados a SentiWordNet, obteniendo para ellos la carga subjetiva.
5. Se realiza el cálculo final de la polaridad mediante la media de la suma de los productos entre el peso del synset tras Page Rank y la polaridad asociada en SentiWordNet.

Debido a que buscamos una combinación de los valores de SentiWordNet con los pesos obtenidos por Page Rank, es importante asegurarse que la fórmula final produce valores comparables. Para ello, los valores de Page Rank para los synsets se ajustan a una norma L_1, de forma que estos vectores de conceptos sumen la unidad como valor máximo

[1] http://www.nltk.org
[2] http://ixa2.si.ehu.es/ukb/

posible. La polaridad final es obtenida mediante la fórmula siguiente:

$$p(t) = \frac{\mathbf{r} \cdot \mathbf{s}}{|\mathbf{t}|} \quad (2)$$

donde p es el valor de polaridad final, $\mathbf{r}$ es el vector de synsets con los pesos obtenidos por Page Rank sobre WordNet, $\mathbf{s}$ es el vector de polaridades correspondientes de SentiWordNet y t es el conjunto de conceptos expandidos que se derivan de la frase.

4. Experimentación

4.1. El corpus JRC

El corpus sobre citas en noticias periodísticas con el que vamos a trabajar ha sido preparado por el Joint Research Center (Balahur et al., 2010) y puede ser descargado desde su web de recursos lingüísticos[3]. Consiste en 1,590 citas en inglés que se han extraído automáticamente desde diversas fuentes periodísticas *online* y anotadas manualmente con expresividad del sentimiento hacia entidades como personas y organizaciones mencionadas en dichas citas. Estas anotaciones han sido llevadas a cabo por cuatro expertos, a partir de los cuales se construyen juicios de polaridad positiva o negativa. De las 1,590 entradas, sólo 427 tienen carga subjetiva y con valor consensuado entre los anotadores.

Un ejemplo de texto en dicho corpus sería el siguiente (etiquetado como *negativo* por dos anotadores):

> *Elisabeth's imprisonment was premeditated; Fritzl had been planning her imprisonment in the dungeon cellar for two to three months before he actually lured his daughter down there.*

4.2. Evaluación

Hemos seleccionado de entre las entradas del corpus aquellas que son subjetivas y han sido consensuadas entre los anotadores, si bien también incluimos los extractos de noticias con una única anotación (ya sea positiva o negativa), lo que nos proporciona un total de 718 citas. Dado el carácter no supervisado de nuestro método, todas éstas son consideradas para la evaluación, calculando los valores de precisión, cobertura y medida F1 sobre este proceso de clasificación binaria.

El sistema, tal y como se ha planteado en secciones anteriores, puede configurarse en base a dos variables:

1. *Tamaño de la expansión.* Este valor hace referencia al número de synsets que vamos a asociar a cada término del texto. De esta forma, un valor 1 indicaría que nos quedamos con el synset asociado al término, por lo que nos limitaríamos a realizar un proceso de desambiguación mediante Page Rank sobre WordNet a partir de las semillas que se obtienen del resto de términos en el texto. Cualquier valor superior a 1 ya nos lleva a una expansión en el conjunto de synsets que vamos a asociar al texto. Es aquí donde reside la fortaleza del método, pues sentidos con carga emocional relevante pueden entrar en el cálculo final de la polaridad aunque no estén representando sentidos (conceptos) explícitamente reflejados en el texto. Trabajamos de esta forma con una solución que tienen en cuenta conceptos "implícitos" en el contexto de los términos de cada fragmento periodístico.

2. *Decisión sobre neutros.* En nuestra propuesta, actualmente los valores sin una orientación positiva o negativa clara son desechados. Pero también sería interesante analizar la tendencia del método a considerar neutros los textos que deberían ser positivos o negativos. Para ello se incorpora un segundo parámetro que fuerza un sesgo sobre el cálculo final, evitando una evaluación de la polaridad neutra (es decir, de valor cero).

4.3. Resultados

4.3.1. Tamaño de la expansión

Para estudiar el efecto que el tamaño de la expansión tiene sobre el método hemos explorado valores desde 1 (la mera desambiguación) a 20 (20 synsets asociados por término). Antes de entrar a analizar los valores obtenidos con el valor de expansión óptimo para esta colección de datos, podemos visualizar dicho efecto en la Figura 1.

De esta figura se extraen algunas observaciones interesantes. La primera es que el efecto que produce generalmente mejora la capacidad de clasificación en términos de F1 (siendo equivalente para precisión y cobertura). A medida que añadimos más términos el comportamiento es estable y podemos decir que en este caso añadir más de 9 conceptos por término no lleva a mejora alguna. En

[3] http://langtech.jrc.it/JRC_Resources.html

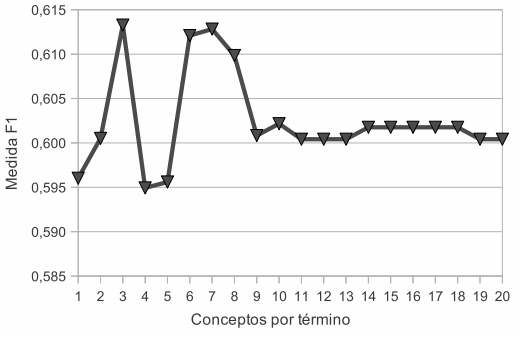


Figura 1: Efecto de la variación en el tamaño de la expansión de conceptos

cambio, sí es notable la variabilidad de la medida sobre los primeros valores. Creemos que dicho efecto tiene que ver con la inclusión de la antonimia y la polisemia como relaciones también consideradas en el grafo de WordNet. Dado que esta gráfica es sobre el total de las expansiones, es importante considerar un análisis futuro que estudie dicho efecto de forma más detallada, pues las expansiones pueden producir resultados muy dispares en función al término asociado.

Los resultados globales con el método propuesto sobre el corpus JRC Quotations quedan reflejados en el Cuadro 1, donde indicamos lo obtenido con expansiones de 1 a 10 conceptos por término (*vsize* en la tabla).

Precisión	Cobertura	F1	vsize
0,5991	0,5929	0,5960	1
0,6039	0,5971	0,6005	2
0,6199	0,6067	**0,6132**	3
0,6011	0,5888	0,5949	4
0,6022	0,5891	0,5955	5
0,6185	0,6057	0,6120	6
0,6192	0,6065	0,6128	7
0,6166	0,6031	0,6098	8
0,6075	0,5941	0,6008	9
0,6089	0,5955	0,6021	10

Cuadro 1: Resultados obtenidos

4.3.2. Tratamiento de valores neutros

En cuanto al tratamiento de los valores neutros, hemos considerado los tres escenarios posibles: considerar los neutros como positivos, considerar los neutros como negativos, o considerar los neutros como tales, no computando en las clases positiva o negativa. El Cuadro 2 resume los valores obtenidos con un valor de expansión de 3 (aquel que nos ha dado mejores resultados). Hemos observado que es mejor considerar el neutro como tal, o lo que es lo mismo, el método propuesto no introduce ningún tipo de sesgo hacia una clase determinada, por lo que no es necesario realizar corrección alguna. Este comportamiento se mantiene independiente del tamaño de la expansión.

Precisión	Cobertura	F1	Neutro
0,6103	0,6008	0,6055	negativo
0,6198	0,6050	0,6123	positivo
0,6199	0,6067	**0,6132**	neutro

Cuadro 2: Consideración de valores neutros

4.3.3. Pesos de Page Rank

Llegados a este punto podemos plantearnos la conveniencia o no de la fórmula planteada, pues re-evalúa los valores de SentiWordNet con los valores para los synsets obtenidos mediante el algoritmo Page Rank (mediante una aproximación basada en Random Walk). El diagrama mostrado en la Figura 2 arroja luz sobre esta cuestión. Podemos reconocer claramente la contribución que los pesos de Page Rank aportan al valor final de polaridad, no solo en cuanto a la mejor sino en cuanto a la estabilización que introduce en base al tamaño de la expansión. También es importante señalar que estos efectos no son tan marcados cuando la expansión no va más allá de uno o dos términos. Lo cual puede estar relacionado con la discusión anterior acerca de los efectos de synsets provenientes de relaciones no deseables en el grafo de WordNet.

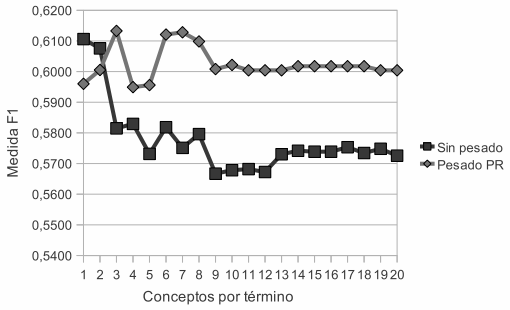


Figura 2: Efecto del pesado con los valores obtenidos por Page Rank

4.3.4. Comparación con otros trabajos

Existen dos trabajos destacados que hayan propuesto métodos de clasificación de

la polaridad evaluados con este corpus y que ya han sido mencionados anteriormente. En un primer trabajo (Balahur et al., 2010), precisamente donde se introduce este corpus a toda la comunidad científica, se estudian diversas variantes en base al uso de distintos recursos lingüísticos (SentiWordNet, Micro WordNet, WordNet Affect, etc.). Obtienen un valor de *accuracy* de 0,25 mediante el uso de SentiWordNet, mientras que nuestro valor máximo para esta misma medida es de 0,59 con una expansión de 6 conceptos con término. No hemos indicado los valores de *accuracy* pues consideramos más interesante para evaluar la clasificación de la polaridad los valores de precisión, cobertura y medida F. No obstante, hemos de resaltar que se logra un valor de 0,82 en esta misma medida usando determinadas bolsas de palabras específicas para este tipo de corpus.

En un trabajo posterior (Boldrini et al., 2012), encontramos otros resultados para la clasificación de este corpus, alcanzando un valor para F1 de 0,5340. Ya hemos indicado en el apartado anterior que el valor máximo obtenido en nuestro caso es de 0,6132. Con lo cual el sistema que presentamos obtiene una mejora de 7,92 %.

5. Conclusiones y trabajo futuro

El método propuesto plantea una solución muy prometedora para el AO. Los resultados obtenidos demuestran que es posible conseguir valores de precisión y cobertura similares o superiores a los de los métodos supervisados más extendidos, escapando a los problemas que plantea la transferencia de sentimiento entre distintos dominios o, dicho de otro modo, la validez de los modelos entrenados sobre un dominio para clasificar en otro. Además, aunque los experimentos han sido probados pare el inglés, el método es completamente adaptable a cualquier idioma diferente.

Dada las bondades del método, nuestros objetivos se centran, por un lado, en seguir comprobando su validez sobre otros corpora y, por otro lado, en aplicar dicho método sobre idiomas distintos al inglés. Esto último es un reto importante, pero afortunadamente creemos que a día de hoy disponemos de recursos como Freeling (Padró et al., 2010) o el Multilingual Central Repository (Atserias et al., 2003) que pueden hacer viable la multiligualidad de nuestra propuesta sin recurrir a traducciones (Denecke, 2008).

Entre otros retos, está el estudio y tratamiento de la negación, ya que las relaciones en el grafo de palabras de WordNet puede verse fuertemente afectado por un seguimiento incorrecto de las relaciones de antonimia, actualmente consideradas para calcular la expansión conceptual.

Bibliografía

Agirre, Eneko y Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. En *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 33–41, Morristown, NJ, USA. Association for Computational Linguistics.

Atserias, J., L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, y P. Vossen. 2003. The MEANING Multilingual Central Repository. En Sojka, P and Pala, K and Smrz, P and Fellbaum, C and Vossen, P, editor, *GWC 2004: SECOND INTERNATIONAL WORDNET CONFERENCE, PROCEEDINGS*, páginas 23–30. 2nd International Global WordNet Conference (GWC 2004), Brno, CZECH REPUBLIC, JAN 20-23, 2004.

Baccianella, Stefano, Andrea Esuli, y Fabrizio Sebastiani. 2008. Sentiwordnet 3.0 : An enhanced lexical resource for sentiment analysis and opinion mining. *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC10*, 0:2200–2204.

Balahur, Alexandra, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik Van Der Goot, Matina Halkia, Bruno Pouliquen, y Jenya Belyaeva. 2010. Sentiment analysis in the news. *English*, 10:2216–2220.

Boldrini, Ester, Alexandra Balahur, Patricio MartÃnez-Barco, y AndrÃ©s Montoyo. 2012. Using emotiblog to annotate and analyse subjectivity in the new textual genres. *Data Mining and Knowledge Discovery*, páginas 1–32. 10.1007/s10618-012-0259-9.

Denecke, K. 2008. Using sentiwordnet for multilingual sentiment analysis. En *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*, páginas 507 –512, april.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Hatzivassiloglou, Vasileios y Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. En *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98, páginas 174–181, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hu, Minqing y Bing Liu. 2004. Mining and summarizing customer reviews. En *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, páginas 168–177, New York, NY, USA. ACM.

Ji, Yang-Sheng, Jia-Jun Chen, Gang Niu, Lin Shang, y Xin-Yu Dai. 2011. Transfer Learning via Multi-View Principal Component Analysis. *JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY*, 26(1):81–98, JAN.

Kanayama, Hiroshi y Tetsuya Nasukawa. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. En *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, páginas 355–363, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kim, Soo-Min y Eduard Hovy. 2004. Determining the sentiment of opinions. En *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Liu, Bing. 2010. Sentiment analysis and subjectivity. En Nitin Indurkhya y Fred J. Damerau, editores, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN 978-1420085921.

Martínez-Cámara, Eugenio, M. Teresa Martín-Valdivia, José M. Perea-Ortega, y L. Alfonso Ure na López. 2011. Técnicas de clasificación de opiniones aplicadas a un corpus en español. *Procesamiento de Lenguaje Natural*, 47.

Mullen, Tony y Nigel Collier. 2004. Sentiment analysis using support vector machines with diverse information sources. En *EMNLP*, páginas 412–418. ACL. conf/emnlp/2004.

Ogawa, Tatsuya, Qiang Ma, y Masatoshi Yoshikawa. 2011. News Bias Analysis Based on Stakeholder Mining. *IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS*, E94D(3):578–586, MAR.

Padró, Lluís, Miquel Collado, Samuel Reese, Marina Lloberes, y Irene Castellón. 2010. Freeling 2.1: Five years of open-source language processing tools. En Nicoletta Calzolari (Conference Chair) Khalid Choukri Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Mike Rosner, y Daniel Tapias, editores, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Page, Lawrence, Sergey Brin, Rajeev Motwani, y Terry Winograd. 1998. The PageRank Citation Ranking: Bringing Order to the Web. Informe técnico, Computer Science Department, Stanford University.

Pang, Bo y Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, jan.

Pang, Bo, Lillian Lee, y Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. En *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, páginas 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tan, Songbo, Xueqi Cheng, Yuefen Wang, y Hongbo Xu. 2009. Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis. En Boughanem, M and Berrut, C and Mothe, J and SouleDupuy, C, editor, *ADVANCES IN INFORMATION RETRIEVAL, PROCEEDINGS*, volumen 5478 de *Lecture Notes in Computer Science*, páginas 337–349. Google; Mateixware Informat Serv; Microsoft Res;

Yahoo; Exalead; GDR 13; Univ Paul Sabatier; ARIA; Inforsid & Reg; Midi Pyrennees. 31st European Conference on Information Research, Toulouse, FRANCE, APR 06-09, 2009.

Tsytsarau, Mikalai y Themis Palpanas. 2012. Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24:478–514.

Turney, Peter 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. Association for Computational Linguistics.

Whitelaw, Casey, Navendu Garg, y Shlomo Argamon. 2005. Using appraisal groups for sentiment analysis. En *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, páginas 625–631, New York, NY, USA. ACM.

Wiegand, Michael, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, y Andrés Montoyo. 2010. A survey on the role of negation in sentiment analysis. En *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, NeSp-NLP '10, páginas 60–68, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wu, Qiong, Songbo Tan, y Xueqi Cheng. 2009. Graph ranking for sentiment transfer. En *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, páginas 317–320, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wu, Qiong, Songbo Tan, Miyi Duan, y Xueqi Cheng. 2010. A Two-Stage Algorithm for Domain Adaptation with Application to Sentiment Transfer Problems. En Cheng, PJ and Kan, MY and Lam, W and Nakov, P, editor, *INFORMATION RETRIEVAL TECHNOLOGY*, volumen 6458 de *Lecture Notes in Computer Science*, páginas 443–453. Natl Taiwan Univ; Natl Sci Council, Republ China; Minist Educ, Republ China. 6th Asia Information Retrieval Societies Conference, Taipei, TAIWAN, DEC 01-03, 2010.

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 157-162 recibido 09-05-12 revisado 04-06-12 aceptado 06-06-12

Learning a Statistical Model of Product Aspects for Sentiment Analysis*

Aprendizaje de un Modelo de Características de Productos para el Análisis de Opiniones

Lisette García-Moya
Universitat Jaume I
Castellon, Spain
lisette.garcia@uji.es

Rafael Berlanga Llavori
Universitat Jaume I
Castellon, Spain
berlanga@uji.es

Henry Anaya-Sánchez
Universitat Jaume I
Castellon, Spain
henry.anaya@uji.es

Resumen: En este artículo se introduce una nueva metodología para modelar características de productos a partir de una colección de opiniones de usuarios. La metodología propuesta se basa en modelos estadísticos de lenguajes y es aplicable a productos de dominio arbitrario. La metodología combina un kernel de palabras de opinión con un modelo de traducción de palabras para estimar el modelo de características. Se presenta además un método para modelar las opiniones vertidas sobre las características. Los experimentos realizados sobre diferentes colecciones de opiniones muestran resultados alentadores en el modelado tanto de características como de opiniones vertidas sobre éstas.
Palabras clave: Minería de Opiniones, Análisis de Opiniones, Modelado de Características de Productos

Abstract: In this paper, we introduce a new methodology for modeling product aspects from a collection of free-text customer reviews. The proposal relies on a language modeling framework and is domain independent. It combines both a kernel-based model of opinion words and a stochastic translation model between words to approach the aspect model of products. We also present a ranking-based methodology to model the sentiments expressed about the aspects. The experiments carried out over several collections of customer reviews show encouraging results in the modeling of product aspects and their sentiments even from individual customer reviews.
Keywords: Opinion Mining, Sentiment Analysis, Product Aspect Modeling

1. Introduction

With the increasing availability of user-generated contents, such as consumer opinion web sites, blogs, Internet forums and social networks, people have more opportunities to express their opinions and make them available to everyone. Publicly available opinions provide valuable information for decision-making processes based on a new collective intelligence paradigm designated as crowdsourcing. The consumer opinion web sites constitute an invaluable way of promotion in which satisfied customers tell other people how much they like a business, product, service, or event. It has became one of the most credible forms of advertising because people who do not stand to gain personally by promoting something put their reputations on the line every time they make a recommendation.

Therefore, in the last years the computational treatment of sentiment and opinions has been viewed as a challenging area of research that can serve different purposes.

One of the most relevant applications of sentiment analysis is the aspect-based summarization (Carenini, Ng, and Pauls, 2006; Yu et al., 2011). Broadly speaking, given a collection of opinion posts about a product or service, this task is aimed at obtaining the most relevant opined aspects (also called features) along with their most relevant

* This work has been partially funded by the "Ministerio de Economía y Competitividad" with contract number TIN2011-24147 and by the Fundació Caixa Castelló project P1-1B2010-49. Lisette García-Moya has been supported by the PhD Fellowship Program of the Universitat Jaume I (PREDOC/2009/12).

ISSN 1135-5948

sentiment information expressed by customers (usually an opinion word and/or a polarity score). Aspect-based summarization is usually composed of three main tasks: *aspect identification*, *sentiment classification*, and *aspect rating*. Aspect identification is focused on extracting the set of aspects concerning the product from the reviews. The word *aspects* is intended to represent both components and attributes. For example, given the sentence, "The bed was comfortable", the review is about the "bed" aspect and the opinion is positively expressed by mean of the opinion word "comfortable". The sentiment classification task consists in determining the opinions about the product aspects and/or their polarities, whereas aspect rating leverages the relevance of aspects and their opinions to properly present them to the users.

In this paper, we address the aspect-based summarization task by introducing a domain independent methodology for modeling product aspects from a set of free-text customer reviews. The proposal relies on a language modeling framework, which combines both a probabilistic model of opinion words and a stochastic self-translation model between words to approach the aspect model of products. From the proposed method, we also derive a ranking-based method to approximate the sentiments expressed about the aspects.

Our work extends the preliminary approach introduced in García-Moya et al. (2012). Specifically, in the present work we propose a more general methodology that effectively allows –for example– the use of dependency relations between words in the modeling of product aspects. We carried out experiments on a wider set of review collections, including a taxonomy-based opinion dataset that entails a harder aspect modeling task. Finally, we also provide an evaluation of the proposed ranking-based method.

As already shown in García-Moya et al. (2012), one strong point of our proposal is that it can effectively retrieve the product aspects even if we do not rely on NLP techniques. This is the main difference with respect to most of the approaches on sentiment analysis that consider the task of aspect identification (Wu et al., 2009; Qiu et al., 2009), as they strongly rely on dependency analysis.

2. Modeling Product Aspects

Given a collection of customer reviews about a specific product and a free-text document d –which can be either a subcollection of reviews or an individual review–, our goal is to obtain a probabilistic model for retrieving the product aspects from d.

Specifically, we consider modeling the set of aspects discussed in d as a statistical language model that assigns higher probability values to words defining aspects.

Let $V = \{w_1, \ldots, w_n\}$ represents the vocabulary of d. Let also $Q = \langle Q(w_1), \ldots, Q(w_n)\rangle^\top$ and $\mathcal{T} = \{\mathrm{p}(w_i|w_j)\}_{1\leq i\leq n, 1\leq j\leq n}$ be a vector-shaped model of opinion words and an n-by-n (column-wise) stochastic matrix representing an entailment-based self-translation model of words from d respectively.

Then, we propose to model unigram product aspects as follows:

$$P(w_i) \propto \left(\begin{pmatrix} \mathrm{p}(w_1|w_1) & \cdots & \mathrm{p}(w_1|w_n) \\ \vdots & \ddots & \vdots \\ \mathrm{p}(w_n|w_1) & \cdots & \mathrm{p}(w_n|w_n) \end{pmatrix}^k \cdot \begin{pmatrix} Q(w_1) \\ \vdots \\ Q(w_n) \end{pmatrix} \right)_i \quad (1)$$

$$= \left(\mathcal{T}^k \cdot Q\right)_i \quad (2)$$

where $k > 0$ is the number of times that the stochastic translation $\mathcal{T}$ is applied to the opinion model Q.

In the context of customer reviews, opinion words (e.g., "excellent", "terrible", etc.) are usually utilized to express sentiments about the different aspects of a product. This causes the review texts to reflect some entailment relationship from opinion words to aspect words. The idea behind the above model is that by applying successively the entailment model $\mathcal{T}$ to Q, we can capture such an entailment relationship between opinion and aspects, and define in this way a model of aspect words (García-Moya et al.,2012).

The unigram language model of aspects $P = \{P(w_i)\}_{1\leq i\leq n}$ can be extended to generate aspects of arbitrary length from the model:

$$P^*(s) = \prod_{t=1}^{r} P(w_{i_t})^{1/r} \quad (3)$$

where $s = w_{i_1} \ldots w_{i_r}$ $(r > 0)$.

In addition, we consider refining the unigram model P to avoid the assignment of high probability values to meaningless words

(e.g., prepositions, conjunctions, etc.). The refined unigram language model P' is obtained by performing an *Expectation Maximization* process aimed at maximizing the cross entropy:

$$-\sum_{i=1}^{n} P(w_i)\log(\lambda P'(w_i)+(1-\lambda)P_{bg}(w_i)) \quad (4)$$

where P_{bg} is a background language model of the source language of the reviews (e.g. English). Currently, we estimate P_{bg} from the COCA corpus (Davies, 2011).

The estimation of both the self-translation model $\mathcal{T}$ and the opinion model Q are described in the next sections.

3. Self-translation Model $\mathcal{T}$

For all $i, j \in \{1, \ldots, n\}$, we define $\mathrm{p}(w_i|w_j)$ to be proportional to the number of times word w_i occurs in a local context of words from d containing an occurrence of w_j. In this way,

$$\mathrm{p}(w_i|w_j) = \frac{\mathrm{p}(w_i, w_j)}{\mathrm{p}(w_j)} \quad (5)$$

where:

$$\mathrm{p}(w_i, w_j) \propto \sum_{l\in\mathcal{L}} \mathrm{p}(w_i|l)\cdot\mathrm{p}(w_j|l)\cdot\mathrm{p}(l) \quad (6)$$

$$\mathrm{p}(w_j) = \sum_{w_i\in V} \mathrm{p}(w_i, w_j) \quad (7)$$

$\mathcal{L}$ is the set of all local contexts of words contained in d, $\mathrm{p}(w_i|l) = |l|_{w_i}/|l|$ and $\mathrm{p}(l) = |\mathcal{L}|^{-1}$ ($|l|_{w_i}$ is the number of times w_i occurs in l, and $|l|$ is the number of words contained in l).

In this paper, we consider two alternatives for defining local contexts. The first one defines local contexts as the N-grams occurring in the sentences of d. Given a bag D of dependency relations observed among word occurrences in d, the second alternative defines local contexts as the word tuples of D.

4. Modeling Opinion Words

We rely on a kernel-based density estimation approach to define Q from a predefined set of (general-domain) opinion words $\{u_1, \ldots, u_m\}$. Thus, we define:

$$Q(w) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{K}(w, u_i) \quad (8)$$

where $w \in V$ and $\mathrm{K}(w, u_i)$ is the gaussian kernel:

$$\mathrm{K}(w, u_i) = \exp\left(-0{,}5\cdot h(g(w), g(u_i))^2/\sigma^2\right) \quad (9)$$

such that h represents the geodesic distance between distributions (Dillon et al., 2007), $g(v)$ is the posterior distribution of words $\{\mathrm{p}(w_i|v)\}_{1\le i\le n}$, and σ is a predetermined distribution width. In our experiments, we set $\sigma = 0{,}3$.

Relying on $\mathcal{T}$, we also propose to rank sentiment words with respect to the sequence $s = w_{i_1} \ldots w_{i_r}$ by regarding the score:

$$R(w) = \prod_{t=1}^{r} p(w_{i_t}|w)^{1/r} Q(w). \quad (10)$$

5. Evaluation

To evaluate our approach, we firstly rely on four collections of customer reviews each one corresponding to a product (Apex AD2600, Canon G3, Nokia 6610 and Norton).[1] These collections of reviews are manually annotated at the sentence level with the relevant product aspects referred to in the text (Hu and Liu, 2004; Ding, Liu, and Yu, 2008).

We compare several aspect language models obtained from our approach (by varying the value of k, and using either N-grams of different sizes or dependency relations to estimate the translation model) to the baseline language model obtained by replacing the model $\{P(w_i)\}_{1\le i\le n}$ by the MLE model of words from each product collection.

Since the goal is to measure the effectiveness of the statistical language models on the generation of products aspects, the performance of each language model is measured from the *log-likelihood* of generating the bag of aspects $S = \{s_1, \ldots, s_k\}$ that have been manually annotated in the collection. Specifically, we have considered the shifted likelihood:

$$\ell_{shift}(P^*, S) = \ell(P^*, S) - k\log n \quad (11)$$

where the likelihood $\ell(P^*, S)$ is defined as:

$$\ell(P^*, S) = \sum_{i=1}^{k} \log P^*(s_i) \quad (12)$$

[1] `http://www.cs.uic.edu/~liub/FBS`

Table 1: Language model performance for generating the bag of product aspects.

k	Model	Apex AD2600		Canon G3		Nokia 6610		Norton	
		review	product	review	product	review	product	review	product
	baseline	0.0028	6.2978	0.0065	4.327	0.0087	4.5503	0.0033	1.2247
	w2	0.0139	6.0104	0.0146	4.1133	0.0322	4.0846	0.009	1.4627
	w3	0.0141	7.1017	0.0156	4.572	0.0333	4.7469	0.0094	1.6359
	w4	0.0141	7.3233	0.0157	4.6456	0.0335	4.8683	0.0096	1.6923
5	**w5**	0.014	7.3554	0.0157	4.6598	0.0334	4.8742	0.0096	1.7083
	w6	0.014	7.3191	0.0157	4.6645	0.0333	4.8392	0.0096	1.7071
	drAll	0.2324	-2.6241	0.3045	1.2349	0.3744	-2.3262	0.2435	-2.2707
	drSelected	**0.8955**	**17.0883**	**1.1059**	**15.2788**	**1.5542**	**15.2324**	**0.997**	**13.3926**
	w2	0.0141	7.125	0.0154	4.6545	0.0332	4.7512	0.0094	1.6571
	w3	0.0142	7.428	0.0158	4.7432	0.0335	4.9571	0.0096	1.7134
	w4	0.0141	7.448	0.0158	4.7323	0.0334	4.97	0.0097	1.7294
10	**w5**	0.014	7.4237	0.0158	4.717	0.0332	4.9368	0.0096	1.7263
	w6	0.014	7.3736	0.0157	4.7037	0.0331	4.885	0.0096	1.7182
	drAll	0.2321	-2.5711	0.3038	1.3244	0.3732	-2.346	0.2434	-2.2561
	drSelected	**0.8951**	**17.2074**	**1.1047**	**15.3229**	**1.5518**	**15.0994**	**0.9969**	**13.3938**
	w2	0.0142	7.3454	0.0157	4.7474	0.0335	4.8923	0.0096	1.7059
	w3	0.0142	7.4502	0.0158	4.7613	0.0335	4.9797	0.0097	1.7254
	w4	0.0141	7.4523	0.0158	4.7381	0.0334	4.9837	0.0097	1.7322
15	**w5**	0.014	7.4254	0.0158	4.7203	0.0332	4.9522	0.0096	1.7274
	w6	0.014	7.3752	0.0157	4.7059	0.033	4.9004	0.0096	1.7188
	drAll	0.2323	-2.5589	0.3042	1.3386	0.3738	-2.3192	0.2435	-2.2467
	drSelected	**0.8932**	**17.2486**	**1.1057**	**15.3953**	**1.5537**	**15.2319**	**0.9971**	**13.4158**
	w2	0.0142	7.3952	0.0158	4.7737	0.0336	4.9356	0.0097	1.7236
	w3	0.0142	7.4518	0.0159	4.7642	0.0335	4.9837	0.0097	1.7281
	w4	0.0141	7.4524	0.0158	4.7386	0.0333	4.9873	0.0097	1.7324
20	**w5**	0.014	7.4254	0.0157	4.7206	0.0332	4.9588	0.0096	1.7274
	w6	0.014	7.3753	0.0157	4.706	0.033	4.9079	0.0096	1.7189
	drAll	0.2322	-2.5593	0.304	1.3445	0.3733	-2.3267	0.2434	-2.2461
	drSelected	**0.8952**	**17.2414**	**1.105**	**15.3912**	**1.5522**	**15.2217**	**0.9969**	**13.4179**

The greater the value of this measure, the better the performance of the language model P^*.

In Table 1, we show the performance of each language model. For each product, we include two columns: one measuring the average performance obtained by applying the method to each individual review (column labeled as *review*), and the other one measuring the performance obtained by applying the method to each (entire) product review collection. The label "wN" represents the model obtained by using N-grams as local contexts of words to build the translation model $\mathcal{T}$; whereas the labels "drAll" and "drSelected" refer to models that defines the local contexts from dependency relations. The model "drAll" uses all the dependency relations obtained with the Stanford dependency parser (De Marneffe, MacCartney, and Manning, 2006), while "drSelected" considers only the set of relations {"nn", "acomp", "advmod", "amod", "det", "dobj", "infmod", "iobj", "measure", "nsubj", "nsubjpass", "partmod", "prep", "rcmod", "xcomp", "xsubj"}.

Several observations can be made by analyzing Table 1. Firstly, it can be appreciated that models obtained by instantiating our approach clearly outperform the baseline (except in the case of drAll when applied to the entire collection, and in the case of $w2$ when $k = 5$ for some products). This shows that only relying on the frequency of words is not enough for modeling product aspects from a collection of customer reviews. Secondly, it can be seen that the models based on dependency relations outperforms those models based on N-grams when applied to individual customer reviews; being *drSelected* the best model. However, *drAll* surprisingly performs the worst at collection level. It seems that using arbitrary dependency relations increases the uncertainty associated to the transla-

Table 2: Performance obtained on the Taxonomy-Based Opinion Dataset.

Measure	Model	cars		headphones		hotels	
		review	category	review	category	review	category
$\ell_{\textbf{shift}}(\mathbf{P}^*, \mathbf{S})$	**baseline**	0.0046	0.1419	0.002	0.0344	0.0068	0.0621
	w2	0.012	0.1766	0.012	0.0616	0.0172	0.0888
	w3	0.0117	0.1757	0.0115	0.059	0.0168	0.0875
	w4	0.0116	0.1747	0.0112	0.0573	0.0165	0.0859
	w5	0.0114	0.1739	0.0111	0.056	0.0163	0.085
	w6	0.0113	0.1728	0.0109	0.0551	0.0161	0.0841
	drAll	0.0948	-0.3437	0.2619	0.2727	0.2965	0.2596
	drSelected	**0.8578**	**1.8708**	**0.9015**	**1.6247**	**1.2073**	**1.9802**
MAP	**w2**	0.5769	0.4436	0.6148	0.498	0.5432	0.416
	w3	0.591	0.4506	0.5841	0.4524	0.5331	0.4013
	w4	0.5523	0.4253	0.5631	0.4275	0.5093	0.3802
	w5	0.5231	0.3927	0.5382	0.4031	0.496	0.3763
	w6	0.501	0.3567	0.5261	0.3837	0.503	0.4002
	drAll	0.6994	0.5623	0.6609	0.5455	0.6574	0.5580
	drSelected	**0.7109**	**0.5988**	**0.6903**	**0.5889**	**0.6832**	**0.5942**

tion model $\mathcal{T}$ as d becomes larger. Finally, we can notice that using N-grams of size $3 \leq N \leq 5$ produce overall similar results.

5.1. Ranking Sentiment Words for Product Aspects

A second experiment was focused on evaluating the score function R proposed for ranking sentiment words for each product aspect.

In this case, we consider the Taxonomy-Based Opinion Dataset from Cruz et al. (2010). This dataset consists of three review collections, each one corresponding to a product category (namely, cars, headphones and hotels). For each collection, there is a set of customer reviews about different products in the category. The customer reviews have been manually annotated at the sentence level with the following elements: (i) the product aspects (explicitly or implicitly) referred in the sentence, (ii) the category of each aspect (based on a given taxonomy), and (iii) the sentiment or opinion word associated to each aspect.

To measure the quality of the ranking of sentiment words obtained for each aspect (in a review/category), we consider calculating the *Mean Average Precision* (MAP) of the obtained ranking with respect to the set of expressed sentiments about the aspect.

In Table 2, we show the average performance of drAll and drSelect (using $k = 15$) in both the generation of product aspects and the retrieval of sentiment words for each manually labeled aspect in each category of the Taxonomy-Based Opinion Dataset.

Similar to the previous experiments, the model drSelect performs the best in each case. The values of ℓ_{shift} (measured according to the entire categories) are relatively smaller in this dataset, since it entails a harder task (there is more than one product in each collection/category). This also justify that MAP values in the case of individual reviews are larger than the values obtained in the case of the entire categories.

6. Conclusion

In this paper, a new methodology for modeling product aspects from a collection of customer reviews has been presented. The proposed method is based on the language modeling framework and is both domain and language independent. We have also presented a ranking-based methodology to model the sentiments expressed about the aspects. The experiments carried out over several collections of customer reviews (with different degree of difficulty) have shown the usefulness of the proposal for properly modeling the product aspects and retrieving their sentiments even from individual reviews. As future work, we plan to develop a generative routine to produce a set of (multi-word) product aspects likely to be generated from both the language model of aspects and the language model of noun phrases from a set of customer reviews. We also consider to extend our methodology to include modeling the polarity of sentiment words.

References

Carenini, G., R. Ng, and A. Pauls. 2006. Multi-document summarization of evaluative text. In *Proc. of EACL 2006*, pages 305–312.

Cruz, F.L., J.A. Troyano, F. Enríquez, F.J. Ortega, and C.G. Vallejo. 2010. A knowledge-rich approach to feature-based opinion extraction from product reviews. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 13–20. ACM.

Davies, Mark. 2011. Word frequency data from the Corpus of Contemporary American English (COCA). Downloaded from http://www.wordfrequency.info on June 01, 2011.

De Marneffe, M.C., B. MacCartney, and C.D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454.

Dillon, J., Y. Mao, G. Lebanon, and J. Zhang. 2007. Statistical Translation, Heat Kernels, and Expected Distance. In *Proc. of the 23rd Conference on Uncertainty in Artificial Intelligence*.

Ding, Xiaowen, Bing Liu, and Philip S. Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the international conference on Web search and web data mining*, pages 231–240.

García-Moya, Lisette, Henry Anaya-Sánchez, and Rafael Berlanga-Llavori. 2012. Combining Probabilistic Language Models for Aspect-Based Sentiment Retrieval. In *Proceedings of the 34th European Conference on Information Retrieval*, volume 7224 of *Lecture Notes in Computer Science*, pages 561–564. Springer-Verlag.

Hu, M and B Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177. ACM Press, New York, NY.

Qiu, Guang, Bing Liu, Jiajun Bu, and Chun Chen. 2009. Expanding domain sentiment lexicon through double propagation. In *Proceedings of the 21st international jont conference on Artifical intelligence*, IJCAI'09, pages 1199–1204.

Wu, Yuanbin, Qi Zhang, Xuanjing Huang, and Lide Wu. 2009. Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, pages 1533–1541.

Yu, J., Z.J. Zha, M. Wang, and T.S. Chua. 2011. Aspect ranking: identifying important product aspects from online consumer reviews. In *Proc. of ACL 2011*, pages 1496–1505.

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 163-170 recibido 07-05-12 revisado 30-05-12 aceptado 05-06-12

A First Approach to the Automatic Detection of Zero Subjects and Impersonal Constructions in Portuguese

Primera aproximación para la detección automática de pronombres cero y construcciones impersonales en portugués

Luz Rello,[1,2] **Gabriela Ferraro**[2]
Web Research Group[1] & TALN, Centre for Autonomous Systems and Neuro-Robotics[2]
Dept. of Information and Communication Technologies
Universitat Pompeu Fabra
luzrello@acm.org, gabriela.ferraro@upf.edu

Iria Gayo
Grupo de Gramática del Español
Departamento de Lengua Española
Universidad de Santiago de Compostela
iria.delrio@usc.es

Resumen: Este trabajo constituye un primer intento de abordar la detección automática de sujetos elididos y de construcciones impersonales en portugués de Brasil, una tarea que no nos consta que se haya llevado a cabo previamente en esta lengua. Para ello, creamos un corpus que contiene más de 5.600 casos anotados con las clases que deben identificarse: sujetos explícitos, sujetos o pronombres omitidos y construcciones impersonales. Estos casos se clasificaron mediante aprendizaje automático basado en rasgos lingüísticamente motivados. Los resultados obtenidos son modestos, aunque prometedores, y proporcionan una orientación para futuros trabajos en este ámbito.
Palabras clave: elipsis de sujeto, construcción impersonal, pronombre cero, sujeto nulo, aprendizaje automático.

Abstract: In this paper we present a first approximation to the automatic detection of zero subjects and impersonal constructions in Brazilian Portuguese. To the best of our knowledge, this is the first attempt of approaching such task using machine learning in Portuguese. We compiled a corpus containing more than 5,600 instances annotated with the classes to be identified: explicit subjects, zero subjects or pronouns and impersonal constructions. We applied machine learning using linguistically motivated features to classify the instances. The results are modest but promising and provide guidance for future work.
Keywords: subject ellipsis, impersonal construction, zero pronoun, null subject, machine learning

1 Introduction

Portuguese is a pro-drop language (Chomsky, 1981) meaning that subject ellipsis is a highly recurring phenomenon. For instance, in our Brazilian Portuguese corpus, 21% of the subjects are elided.

Numerous natural language processing (NLP) tasks require the identification of subject ellipsis. For instance, the identification of zero pronouns is necessary for zero anaphora resolution (Mitkov, 2002; Chaves and Rino, 2008) and for co-reference resolution (Ng and Cardie, 2002). Also, it has been found to be helpful in a number of NLP applications such as machine translation (Peral and Ferrández, 2000) or text categorization (Yeh and Chen, 2003). However, to the best of our knowledge, the recognition of zero pronouns and non-referential impersonal constructions has not yet been addressed in Portuguese. The goal of this paper is to present a first approach of a method to accomplish this task.

The remainder of the paper is organized as follows. Section 2 provides a literature review while Section 3 describes the classes of Brazilian Portuguese subjects. Section 4 presents the creation and the annotation of the corpus and Section 5 discusses the features used and the preliminary results of the

ISSN 1135-5948

machine learning (ML) method. Finally, in Section 6, conclusions are drawn and plans for future work are discussed.

2 Related Work

The difficulty in detecting missing subjects and non-referential pronouns has been acknowledged since the first studies on computational treatment of anaphora (Bergsma, Lin, and Goebel, 2008; Mitkov, 2010).

Identification of zero pronouns and non-referential pronouns is a crucial step in coreference and anaphora resolution systems because the identification of zero anaphors first requires that they be distinguished from non-referential impersonal constructions (Mitkov, 2010). To approach these tasks we find both ruled-based and ML-based approaches. While machine learning methods are known to perform better than rule-based techniques for identifying non-referential expressions (Boyd, Gegg-Harrison, and Byron, 2005), the most favorable approach for detecting zero subjects is under debate (Ferrández and Peral, 2000; Rello, Baeza-Yates, and Mitkov, 2012).

Among the various computational methods for anaphora resolution in Portuguese, to the best of our knowledge, there is only a rule-based system for pronoun resolution which considers specifically zero subjects (Bick, 2010). This method reaches a f-measure of 70.6% (just 74 annotated zero pronouns) for the resolution of zero anaphora, which is the most difficult to approach according to the author (Bick, 2010). However, we found no evaluation for the task of the identification of zero subjects. The rest of the approaches for anaphora resolution in Portuguese do not consider specifically zero subjects and non-referential impersonal constructions. In (Chaves and Rino, 2008) Mitkov's algorithm is used for the resolution of third person pronouns in texts written in Brazilian Portuguese while in (Rocha, 1999), the classification used does not include zero subjects or empty categories.

Apart from Portuguese, there are other pro-drop languages on which related work about zero pronoun identification has been carried out, such as Japanese (Yoshimoto, 1988) – rule-based approach–, Spanish (Ferrández and Peral, 2000; Rello and Illisei, 2009) –rule-based approach–, Chinese (Zhao and Ng, 2007) –ML based– and Romanian –ML based– (Mihaila, Ilisei, and Inkpen, 2011). For the identification of explicit non-referential constructions we found ML based studies in English (Evans, 2001; Bergsma and Yarowsky, 2011), Spanish (Rello, Baeza-Yates, and Mitkov, 2012) and French (Danlos, 2005).

Our approach is mainly inspired by the ML-based methods and combines features which were useful in (Evans, 2001; Zhao and Ng, 2007; Rello, Baeza-Yates, and Mitkov, 2012).

3 Classification

Subject ellipsis is the omission of the subject in a sentence. We consider not only missing referential subject (zero subject) as manifestation of ellipsis, but also non-referential impersonal constructions.

Literature related to linguistic theory (Cunha and Cintra, 1984; Mateus et al., 2003; Rello and Gayo, 2011) has served as a basis for establishing the linguistically motivated classes and the annotation criteria of this work. The features into which Portuguese subjects were distinguished are: [± elliptic] and [± referential] subjects. These two features result in a ternary classification:

- Explicit subjects: [− elliptic, + referential].
- Zero subjects: [+ elliptic, + referential].
- Impersonals: [− elliptic, − referential].

In the examples explicit subjects are presented in italics, zero pronouns are marked by the symbol Ø and impersonal constructions are not explicitly indicated. In the English translations the subjects which are elided in Portuguese are marked with parenthesis and italics.

3.1 Explicit Subjects

Explicit Subjects are realized usually by a nominal group: noun, pronoun, noun phrase (a), free relatives, semi-free relatives or substantival adjectives (Cunha and Cintra, 1984). The syntactic positions of subjects can be pre-verbal or post-verbal. The occurrence of post-verbal subjects is not restricted by any conditions in Portuguese (Mateus et al., 2003). Projections of non-nominal categories such as clauses containing an infinitive or a conjugated verb, interrogative indirect clauses, or indirect exclamative clauses,

can function as subjects (Cunha and Cintra, 1984).

(a) *Este Decreto* dispõe sobre o exercício das funções de regulação, supervisão e avaliação de instituições de educação superior [...].[1]
This Ordinance disposes about the exercise of the functions of regulation, supervision and evaluation of institutions of superior education [...].

3.1.1 Zero Pronouns

An elliptic subject (b) is the result of a nominal ellipsis, where a realized lexical element –elliptic subject– which is needed for the interpretation of the meaning and the structure of the sentence, is omitted since it can be retrieved from its context.

(b) Ø Dispõe sobre as sanções aplicáveis aos agentes públicos nos casos de enriquecimento ilícito no exercício de mandato, cargo emprego [...].
(It) disposes about the applicable sanctions to the public agents in the cases of illicit enrichment in the mandate exercise, position used [...].

In Portuguese, the noun head can be omitted (Clara, 2008) when the subject of which it is a part fulfills some structural requirements and a definite article occurs, such as *os* in (c).

(c) Em o período do estudo, foram analisadas 7.956 ligações, sendo os usuários de drogas os Ø que mais procuraram atendimento, com 2.600 ligações.
During the study period, 7,956 phone calls were analysed, being the drug users *(those that)* looked for more attention, with 2,600 phone calls.

3.1.2 Impersonals

Impersonal constructions are both non-referential and elliptic (Cunha and Cintra, 1984; Mateus et al., 2003). The appearance of clauses containing zero pronouns is similar to impersonal constructions. This category is composed of impersonal constructions which are formed by impersonal verbs (d) and reflex impersonal clauses, impersonal clauses with *"se"* (e).

(d) Ainda não há consenso em relação a melhor sistemática a ser empregada para apresentação de um instrumento com equivalência transcultural.
(There) is still no consensus in relation to the best systematic to be used for presentation of an instrument with transcultural equivalence.

(e) Optou-se por uma abordagem qualitativa.
(It) was chosen a qualitative approach.

4 Corpus

The training data used in the learning process was obtained from a corpus created specifically for this work, the Explicit Subjects, Zero-Pronouns and Impersonal Constructions corpus (ESZIC) (Rello and Gayo, 2012).[2]

The corpus is composed of 17 documents, originally written in Brazilian Portuguese and belonging to two genres: legal and health.

The legal texts are extracted from the: Civil Code of United States of Brazil (until third book, title II), Brazilian Penal Code (until title VIII, chapter VII), Brazilian Constitution of 1988 (until title III, chapter V), Law of Administrative Dishonesty (whole text), Antitrust Law (until chapter III, article VII), Law no. 9,637 (whole text), Law no. 12,232 (whole text); and Decree no. 5,773 (until chapter II, section II).

The health texts are psychiatric papers taken from the digital journal of psychiatry: *Revista de Psiquiatria do Rio Grande do Sul*. All the papers were written from 2003 to 2009.[3]

The texts were parsed by *Palavras*[4] (Bick, 2000), a parser based on Constraint Grammar methodological paradigm (Karlsson, 1990; Karlsson, Voutilainen, and Anttila, 1995). *Palavras* returns morphological information (part of the speech (POS) and lemma), syntactic information (structure of constituents and their dependency values) and semantic information (semantic prototypes).

[1] All the examples provided are taken from our corpus.

[2] Publicly available at: `http://www.luzrello.com/Projects.html`.

[3] The full-text articles from *Revista de Psiquiatria do Rio Grande do Sul* are available online at: `http://www.scielo.br/scielo.php?script=sci_serial&pid=0101-8108&lng=en&nrm=iso`

[4] `http://beta.visl.sdu.dk/visl/pt/info/`

4.1 Corpus Annotation

The annotator was presented the sentences in which a verb or a group of verbs appear and prompted to classify the verb into one of classes: verb with an explicit subject, verb with a zero subject or verb with no subject (impersonal construction).

Our corpus contains 5,665 finite verbs, 77% have an explicit subject, 21% a zero pronoun and 2% are impersonal constructions (see Table 1). This fact is consistent with linguistic literature since some studies claim that Brazilian Portuguese is a partial pro-drop language, mainly due to the progressive decrease of zero pronouns usage (Kato and Negrão, 2000; Gayo and Rello, 2011).

N **of instances**	**Legal**	**Health**	**All**
Explicit subjects	1,891	2,462	4,353
Zero subjects	462	740	1,202
Impersonals	55	55	110
Total	**2,408**	**3,257**	**5,665**

Table 1: Number of instances per class.

4.2 Inter-annotator Agreement

To test the reliability, validity and stability of the annotations we have computed the inter-annotator agreement.

Among the possible metrics to measure inter-annotator reliability we chose Fleiss' Kappa statistical measure (Fleiss, 1971). This measure is a generalization of Scott's Pi statistic (Scott, 1955) and an extension of the Cohen's Kappa coefficient of agreement for nominal scales to measure agreement in ordinal scale data (Cohen, 1960). Whereas Cohen's Kappa works for only two raters, Fleiss' Kappa works for any number of raters giving categorical ratings, to a fixed number of items. Fleiss' Kappa determines the chance of agreement among arbitrary coders and do not treat all kinds of disagreements in the same manner.

Given the high cost of conducting inter-annotator studies, we choose a representative sample from our corpus to limit the scope of the analysis. We extracted 10% of the instances of each of the texts of the corpus covering the two genres. Two volunteer graduate students, native speakers of Portuguese, participated in the study.

There is no universally accepted interpretation of Kappa values. However, it is common practice among researchers in computational linguistics to consider 0.8 as a minimum value of acceptance (Artstein and Poesio, 2008).

Considering these factors, our results indicate that the annotation is reliable (see Table 2). There is a small number of categories but the Fleiss Kappa value is high. Therefore, our corpus can provide a reliable resource to study subject ellipsis in Portuguese.

Genre	2 Annotators	3 Annotators
Legal	0.8261	0.8255
Health	0.9589	0.8570

Table 2: Fleiss' Kappa coefficient for the inter-annotator agreement.

5 Detecting Subject Ellipsis

In this section we present the linguistically motivated features and the results of our ML-based method.

5.1 Features

We extracted nine features taken from previous studies (Evans, 2001; Rello, Baeza-Yates, and Mitkov, 2012) in order to classify instances according to the three classes defined in Section 3. The values of the features were derived from information provided by *Palavras* parser. These are:

- **Subject** The presence or absence of a subject in the clause, as identified by the parser.
- **Lemma** The lemma of the finite verb, that is, its infinitive form.
- **Number** The grammatical number of the verb (singular or plural).
- **Person** The grammatical person of the verb (first, second, or third).
- **Total Noun Phrases** The number of noun phrases in the clause that precede the verb.
- **Previous Noun Phrases** The total number of noun phrases in the clause.
- **Se** A binary feature encoding the presence or absence of the Portuguese particle *"se"* in the clause.
- **Previous POS** The POS of the four tokens preceding the instance.
- **Following POS** The POS of the four tokens following the instance.

Class	P	R	F
Explicit Subj.	88.4%	90.4%	89.4%
Zero Subj.	60.2%	55.9%	57.9%
Impersonals	81.7%	69.1%	74.9%
Weighted Avg.	82.6%	83.0%	82.8%

Table 3: LAD Tree performance (83.04% accuracy for ten-fold cross validation).

5.2 Preliminary Results

The training data is composed of 5,665 vectors. Each vector corresponds to one finite verb extracted from the corpus and is composed by the values of the features derived from the corpus. There is a training set but no explicit test set, therefore we use cross-validation.

To determine the most accurate algorithm for our classification task, we compared the learning algorithms implemented in Weka (Witten and Frank, 2005). Firstly, the classification was executed using the default values of the algorithms using ten-fold cross-validation. Secondly, the highest performing classifiers were compared modifying the number of iterations.

The decision tree learning classifier LAD Tree (Breiman, 1984) using 10 iterations was the best performing one with an overall accuracy[5] of 83.04%. Table 3 shows the results for each class using ten-fold cross validation.

We used as baseline the output of parser *Palavras*. It is not possible to make a fair comparison because both classes, zero subjects and impersonals are not distinguished by *Palavras*, so we included them in the same class (verbs with no subject identified by the parser) to compare both systems accuracies. Our method outperforms *Palavras* for identifying explicit subjects and impersonals. Although *Palavras* presents a higher accuracy for the identification of verbs with no explicit subjects, or method distinguishes between referential and non-referential elided subjects.

A confusion matrix show us that most of the wrongly identified instances are zero pronouns (36.10%) classified as explicit subjects, given that this class is the most heterogeneous one.

[5]Accuracy is understood as the percentage of the correctly classified instances.

Algorithm	Explicit Subject	Zero Subject	Impersonals
Palavras	71.4%	77.3%	
Our method	94.43%	55.87%	69.09%

Table 4: Summary of accuracy comparison with the baseline.

Class	Explicit Subject	Zero Subject	Impersonal
Explicit Subj.	3957	409	10
Zero Subj.	493	633	7
Impersonals	24	10	76

Table 5: Confusion matrix.

6 Conclusions and Future Work

This is only a first approach to the automatic classification of zero pronouns and impersonal constructions in Brazilian Portuguese. However, to the best of our knowledge, it is the first time to tackle this task for this language using a ML-based approach. The results are modest but promising and provide insights for future work.

A first goal for future work is to improve our method by extending the set of features and performing feature analysis to select the best performing ones. For the creation of new features we will conduct an errors analysis of the wrongly classified instances taking into account their linguistic context in the corpus. Another future research is the extrinsic evaluation of our system by integrating our system in NLP tasks.

References

Artstein, R. and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Bergsma, S., D. Lin, and R. Goebel. 2008. Distributional identification of non-referential pronouns. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT-08)*, pages 10–18.

Bergsma, S. and D. Yarowsky. 2011. Nada: A robust system for non-referential pronoun detection. *Anaphora Processing and Applications*, pages 12–23.

Bick, Eckhard. 2000. *The Parsing System Palavras - Automatic Grammatical Anal-*

ysis of Portuguese in a Constraint Grammar Famework. Aarhus University Press, Aarhus.

Bick, Eckhard. 2010. A dependency-based approach to anaphora annotation. In *Extended Activities Proceedings, 9th International Conference on Computational Processing of the Portuguese Language (PROPOR)*.

Boyd, A., W. Gegg-Harrison, and D. Byron. 2005. Identifying non-referential *it*: a machine learning approach incorporating linguistically motivated patterns. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing. 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, pages 40–47.

Breiman, L. 1984. *Classification and regression trees*. Chapman & Hall/CRC.

Chaves, A. and L. Rino. 2008. The mitkov algorithm for anaphora resolution in portuguese. *Computational Processing of the Portuguese Language*, pages 51–60.

Chomsky, N. 1981. *Lectures on government and binding*. Mouton de Gruyter, Berlin, New York.

Clara, Daniela. 2008. *A aquisição da elipse nominal em português europeu - produção e compreensão*. Ph.D. thesis, Universidade Nova de Lisboa, Lisboa.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Cunha, Celso and Lindley Cintra. 1984. *Nova Gramática do Português Contemporâneo*. Sá da Costa, Lisboa.

Danlos, L. 2005. Automatic recognition of French expletive pronoun occurrences. In Robert Dale, Kam-Fai Wong, Jiang Su, and Oi Yee Kwong, editors, *Natural language processing. Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-05)*, pages 73–78, Berlin, Heidelberg, New York. Springer. Lecture Notes in Computer Science, Vol. 3651.

Evans, R. 2001. Applying machine learning: toward an automatic classification of *it*. *Literary and Linguistic Computing*, 16(1):45–57.

Ferrández, A. and J. Peral. 2000. A computational approach to zero-pronouns in Spanish. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, pages 166–172.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Gayo, I. and L. Rello. 2011. El fenómeno pro-drop en portugués brasileño y español peninsular. In *III Congreso Internacional de Lingüística de Corpus (CILC 2011)*.

Karlsson, F. 1990. Constraint grammar as a framework for parsing running text. In *Proceedings of the 13th conference on Computational linguistics-Volume 3*, pages 168–173. Association for Computational Linguistics.

Karlsson, F., A. Voutilainen, and A. Anttila. 1995. *Constraint Grammar: a language-independent system for parsing unrestricted text*. Walter de Gruyter.

Kato, M. A. and E.V. Negrão. 2000. *Brazilian Portuguese and the Null Subject Parameter*. Vervuert-Iberoamericana, Frankfurt.

Mateus, Maria H., Ana M. Brito, Inês Duarte, Isabel H. Faria, Sónia Frota, Gabriela Matos, Fátima Oliveira, Marina Vigário, and Alina Villalva. 2003. *Gramática da Língua Portuguesa*. Editorial Caminho, Lisboa, 5 edition.

Mihaila, C., I. Ilisei, and D. Inkpen. 2011. Zero pronominal anaphora resolution for the romanian language. *Research Journal on Computer Science and Computer Engineering with Applications POLIBITS*, 42.

Mitkov, R. 2002. *Anaphora resolution*. Longman, London.

Mitkov, R. 2010. Discourse processing. In Alexander Clark, Chris Fox, and Shalom Lappin, editors, *The handbook of computational linguistics and natural language processing*. Wiley Blackwell, Oxford, pages 599–629.

Ng, V. and C. Cardie. 2002. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th International*

Conference on Computational Linguistics (COLING-02), pages 1–7.

Peral, J. and A. Ferrández. 2000. Generation of Spanish zero-pronouns into English. In D. N. Christodoulakis, editor, *Natural Language Processing - NLP 2000. Proceedings of the 2nd International Conference on Natural Language Processing (NLP-2000)*. Springer, Berlin, Heidelberg, New York, pages 252–260. Lecture Notes in Computer Science, Vol. 1835.

Rello, L., R. Baeza-Yates, and R. Mitkov. 2012. Elliphant: Improved automatic detection of zero subjects and impersonal constructions in spanish. In *Proceedings of the European chapter of the Association for Computational Linguistics (EACL 2012)*. Association for Computational Linguistics.

Rello, L. and I. Gayo. 2011. Classification criteria for the automatic identification of subject ellipsis and impersonal constructions in Portuguese. In *Proceedings of the 25th Sociedad Española de Lingüística (SEL 2011)*.

Rello, L. and I. Gayo. 2012. A comparable Portuguese-Spanish corpus with ellipsis annotations. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC 2012)*.

Rello, L. and I. Illisei. 2009. A rule-based approach to the identification of Spanish zero pronouns. In *Student Research Workshop. International Conference on Recent Advances in Natural Language Processing (RANLP-09)*, pages 209–214.

Rocha, M. 1999. Coreference resolution in dialogues in english and portuguese. In *Proceedings of the Workshop on Coreference and its Applications*, pages 53–60. Association for Computational Linguistics.

Scott, W.A. 1955. Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly*.

Witten, I. H. and E. Frank. 2005. *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann, London, 2 edition.

Yeh, C. and Y. Chen. 2003. Using zero anaphora resolution to improve text categorization. In *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation (PACLIC-03)*, pages 423–430.

Yoshimoto, K. 1988. Identifying zero pronouns in Japanese dialogue. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING-88)*, pages 779–784.

Zhao, S. and H.T. Ng. 2007. Identification and resolution of Chinese zero pronouns: a machine learning approach. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CNLL-07)*, pages 541–550.

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 171-178 recibido 05-05-12 revisado 28-05-12 aceptado 04-06-12

Optimizing Planar and 2-Planar Parsers with MaltOptimizer

Optimizando los Parsers Planar y 2-Planar con MaltOptimizer

Miguel Ballesteros†, Carlos Gómez-Rodríguez‡, Joakim Nivre§
†Universidad Complutense de Madrid, Spain
‡Universidade da Coruña, Spain
§Uppsala University, Sweden
miballes@fdi.ucm.es, carlos.gomez@udc.es, joakim.nivre@lingfil.uu.se

Resumen: MaltOptimizer es una herramienta capaz de proporcionar una optimización para modelos generados mediante MaltParser. Los analizadores de dependencias actuales requieren una completa configuración para obtener resultados a la altura del estado del arte, y para ello es necesario un conocimiento especializado. Los analizadores Planar y 2-Planar son dos algoritmos diferentes y de reciente incorporación en MaltParser. En el presente artículo presentamos cómo estos dos analizadores pueden incluirse en MaltOptimizer comparándolos con el resto de familias de algoritmos incluidas en MaltParser, y cómo se puede definir una búsqueda y selección de atributos (o "features") usando el propio sistema para estos dos parsers. Los experimentos muestran que usando estos métodos podemos mejorar la precisión obtenida hasta un porcentaje absoluto del 8 por ciento (labeled attachment score) si lo comparamos con una configuración básica de estos 2 parsers.
Palabras clave: Análisis sintáctico de dependencias, MaltOptimizer, MaltParser, Planar y 2-Planar

Abstract: MaltOptimizer is a tool that is capable of finding an optimal configuration for MaltParser models, taking into account that nowadays dependency parsers require careful tuning in order to obtain state-of-the-art results, and this tuning is normally based on specialized knowledge. The Planar and 2-Planar parsers are two different parsing algorithms included in MaltParser. In the present paper, we show how these two parsers can be included in MaltOptimizer processes comparing them with the rest of MaltParser algorithm families, and how we can define a deep feature search and selection by using MaltOptimizer for these two algorithms. The experiments show that by using MaltOptimizer we can improve parsing accuracy for Planar and 2-Planar parsers by up to 8 percent absolute (labeled attachment score) compared to default settings.
Keywords: Dependency parsing, MaltOptimizer, MaltParser, Planar and 2-Planar

1. Introduction

Data-driven applications are very useful because as soon as we have new data, they can be applied to different domains just by training the new models. However, this training normally requires careful tuning in order to obtain a reliable outcome. MaltOptimizer[1] (Ballesteros and Nivre, 2012b; Ballesteros and Nivre, 2012a) automates[2] the search of an optimal configuration for models obtained with MaltParser (Nivre, Hall, and Nilsson, 2006), which is a typical example of a widely used data-driven system. This requires finding preliminary parameters, such as the handling of covered roots or multiple root labels, selecting the parsing algorithm that best suits the data, and finally, a backward and forward feature selection with the intention of making use of all the possible information included in the data format. MaltOptimizer is an expert system, since it was built by using previous experience in the optimization of MaltParser models during the last years.

MaltParser, which is a transition-based dependency parser with state-of-the-art performance for many languages, was one of the top parsers in the CoNLL Shared Tasks

[1] http://nil.fdi.ucm.es/maltoptimizer

[2] MaltOptimizer is fully automatic and it can be run in batch mode, but it also allows an interaction with the user between phases.

ISSN 1135-5948

on dependency parsing (Buchholz and Marsi, 2006; Nivre et al., 2007). It implements different families of transition-based parsing algorithms: (i) Nivre's algorithms (Nivre, 2003; Nivre, 2008), (ii) Covington's algorithms (Covington, 2001; Nivre, 2008), and (iii) Stack algorithms (Nivre, 2009; Nivre, Kuhlmann, and Hall, 2009). However, there is another family that was not handled by the initial version of MaltOptimizer: (iv) the Multiplanar parsers (Gómez-Rodríguez and Nivre, 2010), which include two algorithms: the Planar parser and the 2-Planar parser. These are linear-time algorithms that cover the sets of *multiplanar* dependency structures described by Yli-Jyrä (2003): while the Planar parser is limited to dependency graphs with no crossing arcs, which are a rather tight superset of projective dependency graphs, the 2-Planar algorithm supports the vast majority of phenomena present in natural language treebanks, and does so in linear time and with competitive accuracy (Gómez-Rodríguez and Nivre, 2010).[3] For these reasons, the Planar and 2-Planar algorithms have been used in several applications and studies in recent literature (Ott and Ziai, 2010; Krivanek and Meurers, 2011; Beuck, Köhn, and Menzel, 2011; Gómez-Rodríguez and Fernández-González, 2012a; Gómez-Rodríguez and Fernández-González, 2012b).

This is why we believe it necessary to include the Multiplanar parsers in MaltOptimizer, and in this paper we present how we addressed this problem. We present how the Multiplanar parsers can be included into the system decision trees and how we refined the feature selection for these parsers.

In the rest of the paper, we introduce transition-based dependency parsing and the Planar and 2-Planar parsers (Section 2). We describe MaltOptimizer in a deeper way by also citing some similar systems and the modifications that we included in order to host these parsers (Section 3). We report experiments, showing how MaltOptimizer finds suitable parameters and feature models when the Multiplanar parsers are selected (Section 4). Finally, we show conclusions and plans for future work (Section 5).

2. The Planar and 2-Planar Dependency Parsers

Given a sentence $w_1 \dots w_n$, the goal of a dependency parser is to assign it a *dependency graph*, which is a directed graph $G = (V, A)$ where $V = \{1, \dots, n\}$ and $A \subseteq V \times V$.[4] We will assume that such dependency analyses are required to satisfy the *acyclicity constraint* (i.e. that the graph cannot have cycles) and the *single-head constraint* (that a node cannot have more than one incoming arc).

Transition-based dependency parsers assign dependency analyses to natural language sentences by using non-deterministic state machines, called *transition systems*, whose actions (transitions) manipulate input words and build dependency relations between them. The choice of the particular sequence of actions to parse each given sentence is performed by scoring transitions using a model learned from training data, and using these scores to select a suitable transition sequence. In the particular case of MaltParser, SVM classifiers are used to learn the model, and the selection of a transition sequence is done by greedy deterministic search, which proceeds by choosing the highest-scoring transition at each parser state.

The Planar and 2-Planar parsers, introduced by Gómez-Rodríguez and Nivre (2010), are among the parsing algorithms implemented in MaltParser. These algorithms use the transition systems shown in Figure 1. We give here a brief description of how both parsers work, a more thorough explanation can be found in Gómez-Rodríguez and Nivre (2010).

The Planar parser, like other algorithms implemented in MaltParser, uses two data structures to manipulate input words: a *buffer*, which holds the words that have not yet been read, and a *stack*, containing words that have already been read but that we still may wish to connect to other words via dependency arcs. The SHIFT transition is used

[3] In theory, the Multiplanar family of parsers is an infinite hierarchy of parsers with increasing coverage (an m-Planar parser can be defined for any natural number m). However, only the parsers with $m = 1$ and $m = 2$ have been implemented in MaltParser because the 2-Planar parser already has a very large coverage of non-projective phenomena (Gómez-Rodríguez and Nivre, 2010), so the practical interest of m-Planar parsers with $m > 2$ is dubious.

[4] In practice, the arcs in the set A are labeled, but we ignore arc labels in this explanation for simplicity of presentation.

Planar transition system:

Initial/terminal configurations: $c_s(w_1 \dots w_n) = \langle [], [1 \dots n], \emptyset \rangle$, $C_f = \{\langle \Sigma, [], A \rangle \in C\}$

Transitions:		
	SHIFT	$\langle \Sigma, i\|B, A\rangle \Rightarrow \langle \Sigma\|i, B, A\rangle$
	REDUCE	$\langle \Sigma\|i, B, A\rangle \Rightarrow \langle \Sigma, B, A\rangle$
	LEFT-ARC	$\langle \Sigma\|i, j\|B, A\rangle \Rightarrow \langle \Sigma\|i, j\|B, A \cup \{(j,i)\}\rangle$ only if $\nexists k \mid (k,i) \in A$ (single-head) and $A \cup \{(j,i)\}$ is acyclic.
	RIGHT-ARC	$\langle \Sigma\|i, j\|B, A\rangle \Rightarrow \langle \Sigma\|i, j\|B, A \cup \{(i,j)\}\rangle$ only if $\nexists k \mid (k,j) \in A$ (single-head) and $A \cup \{(i,j)\}$ is acyclic.

2-Planar transition system:

Initial/terminal configurations: $c_s(w_1 \dots w_n) = \langle [], [], [1 \dots n], \emptyset \rangle$, $C_f = \{\langle \Sigma_0, \Sigma_1, [], A \rangle \in C\}$

Transitions:		
	SHIFT	$\langle \Sigma_0, \Sigma_1, i\|B, A\rangle \Rightarrow \langle \Sigma_0\|i, \Sigma_1\|i, B, A\rangle$
	REDUCE	$\langle \Sigma_0\|i, \Sigma_1, B, A\rangle \Rightarrow \langle \Sigma_0, \Sigma_1, B, A\rangle$
	LEFT-ARC	$\langle \Sigma_0\|i, \Sigma_1, j\|B, A\rangle \Rightarrow \langle \Sigma_0\|i, \Sigma_1, j\|B, A \cup \{(j,i)\}\rangle$ only if $\nexists k \mid (k,i) \in A$ (single-head) and $A \cup \{(j,i)\}$ is acyclic.
	RIGHT-ARC	$\langle \Sigma_0\|i, \Sigma_1, j\|B, A\rangle \Rightarrow \langle \Sigma_0\|i, \Sigma_1, j\|B, A \cup \{(i,j)\}\rangle$ only if $\nexists k \mid (k,j) \in A$ (single-head) and $A \cup \{(i,j)\}$ is acyclic.
	SWITCH	$\langle \Sigma_0, \Sigma_1, B, A\rangle \Rightarrow \langle \Sigma_1, \Sigma_0, B, A\rangle$

Figure 1: The planar and 2-planar transition systems. Note that we use the notation $\Sigma|i$ for a stack with i at the top and tail Σ, and $i|B$ for a buffer with i as its first word and tail B.

to read the first word from the buffer, moving it to the stack. The LEFT-ARC and RIGHT-ARC transitions create a leftward (rightward) dependency arc involving the topmost stack node and the first node remaining in the buffer, and can only be executed if this arc does not create a cycle or violate the single-head constraint when combined with the already created arcs. Finally, the REDUCE transition can be employed to remove a word from the stack when we do not need to create more arcs using it. This set of actions allows the Planar parser to build any dependency graph that is planar, i.e., not containing crossing arcs. Note that planarity is a very mild relaxation of the well-known notion of projectivity, meaning that the practical coverage of the planar parser is only very slightly larger than that of projective dependency parsers.

To expand this coverage further and obtain a parser that would be able to analyze in linear time the vast majority of dependency structures present in natural language treebanks, the Planar parser was extended by adding an extra stack, obtaining the 2-Planar parser. The 2-Planar parser is capable of building any dependency graph that can be divided into two subgraphs (planes) that are planar. To do so, it uses two stacks, one per plane, such that one of them is marked as the *active* stack at each given configuration, the other being the *inactive* stack. The SHIFT transition works analogously to that in the Planar parser but it moves the first word in the buffer to both stacks. The REDUCE, LEFT-ARC and RIGHT-ARC transitions also work like their Planar counterparts, but they involve only the active stack. Finally, an additional SWITCH transition makes the active stack inactive and vice versa.

The paper by Gómez-Rodríguez and Nivre (2010) shows that the 2-Planar parser can produce results that are on par with well-known state of the art dependency parsers, like the arc-eager pseudo-projective parser by Nivre et al. (2006).

3. *MaltOptimizer*

MaltOptimizer (Ballesteros and Nivre, 2012b; Ballesteros and Nivre, 2012a) implements a search of different parameters for MaltParser based mainly on the heuristics described by Nivre and Hall (2010) and previous experience acquired during the last years. MaltOptimizer takes a single input, which is a training set in CoNLL data format,[5] and returns suggestions of an optimal configuration

[5] http://ilk.uvt.nl/conll/#dataformat

for MaltParser models, providing a complete option file and a feature specification file.

MaltOptimizer also estimates the expected results by providing labeled attachment score results (LAS) (Buchholz and Marsi, 2006).[6] It only explores linear multiclass SVMs in LIBLINEAR (Fan et al., 2008), excluding LIBSVM (Chang and Lin, 2001) for efficiency reasons. It has been observed that the expected outcomes are similar between both libraries but LIBLINEAR takes much less running time. This fact makes the experiments shown in this paper even more interesting, because most of the feature models that have been obtained manually for Multiplanar parsers are based on LIBSVM (Gómez-Rodríguez and Nivre, 2010; Gómez-Rodríguez and Fernández-González, 2012a; Gómez-Rodríguez and Fernández-González, 2012b).

It is worth noting that both LIBSVM and LIBLINEAR provide outcomes in the same range of accuracy, or as stated in (Prudhvi Kosaraju and Kukkadapu, 2010) or (Gómez-Rodríguez and Fernández-González, 2012b), LIBLINEAR can even provide better results than LIBSVM.

There are some other systems with the same intention as MaltOptimizer, due to the importance of feature selection and parameter optimization in machine-learning based systems. However, in the NLP community, the set of such systems is very limited. We can find feature selection and parameter optimization systems, such as Kool, Zavrel and Daelemans (2000) and Daelemans et al. (2003). More recently, Nilsson and Nugues (2010) explored automatic feature selection specifically for MaltParser, but MaltOptimizer is the first system that implements a complete customized optimization process for this system and for a big set of algorithms.

MaltOptimizer is divided in three different phases: (i) data analysis, (ii) parsing algorithm selection and (iii) feature selection. In the following subsections we describe each of them and we show the modifications that we had to make in order to host the Multiplanar parsers in the MaltOptimizer processes.

[6] In the default settings, it provides LAS including punctuation symbols, but it can be configured excluding punctuation symbols and excluding the labeling returning unlabeled attachment scores (UAS).

3.1. Phase 1: Data Analysis

In the first phase, MaltOptimizer gathers some information that leads the optimization for the following phases. Some of the properties are crucial in order to select the best parsing algorithm, which makes them very important due to the aim of the present paper.

1. Size of the training set: number of words/sentences.
2. Existence of "covered roots" (arcs spanning tokens with HEAD = 0).
3. Frequency of labels used for tokens with HEAD = 0.
4. Percentage of non-projective arcs/trees.
5. Existence of non-empty feature values in the LEMMA and FEATS columns.
6. Identity (or not) of feature values in the CPOSTAG and POSTAG columns.

In order to host Multiplanar parsers, we did not change the behavior of the first phase at all. The first three properties are used to set basic parameters, such as the way of handling covered roots by the model; 4 is used in the choice of parsing algorithm (phase 2), and for Multiplanar parsers it is basically the same as for other algorithms, the modifications are shown in the Phase 2 description; 5 and 6 are relevant for feature selection experiments (phase 3).

3.2. Phase 2: Parsing Algorithm Selection

The original version of MaltOptimizer handled three different groups of transition-based parsing algorithms: (i) Nivre's algorithms (Nivre, 2003; Nivre, 2008), (ii) Covington's algorithms (Covington, 2001; Nivre, 2008), and (iii) Stack algorithms (Nivre, 2009; Nivre, Kuhlmann, and Hall, 2009). However, this initial release was not able to handle Multiplanar parsers (Gómez-Rodríguez and Nivre, 2010), which is basically the goal of this paper. The Multiplanar group of algorithms contains algorithms that can handle non-projective dependency trees (2-Planar or 2-Planar arc-eager parser), and a roughly projective[7] version (Planar or Planar arc-eager) that can be combined with

[7] As mentioned above, the Planar parser has coverage over the set of planar dependency graphs, which is a rather tight superset of the set of projective dependency graphs.

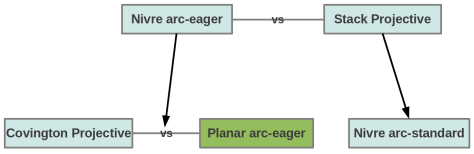


Figure 2: New decision tree for best projective algorithm in which Planar arc-eager is considered.

pseudo-projective parsing to recover non-projective dependencies in post-processing by using the pseudo-projective parsing approach (Nivre and Nilsson, 2005).

MaltOptimizer has two different decision trees, the first one is only used if the number of non-projective dependencies in the training set is small (or zero), and the second one is used if the number of non-projective trees is not zero. Therefore, for most of the languages both decision trees were explored in previous experiments (Ballesteros and Nivre, 2012b), with some exceptions, such as Chinese.[8]

We needed to locate the Multiplanar parsers in the decision trees in order to be able to select them as best parsers. Planar arc-eager has the same parsing order and a similar way of handling left and right attachments as Nivre arc-eager, this is why we locate it in the same branch in which the main parent is Nivre arc-eager, making a comparison with Covington projective. The new projective decision tree is shown in Figure 2.

When the training set contains a substantial amount of non-projective dependencies, the older version of MaltOptimizer tests the non-projective versions of Covington's algorithm and the Stack algorithm (including a lazy and an eager variant), and also the projective algorithms in combination with pseudo-projective parsing. We have 2-Planar arc-eager and Planar arc-eager, which is a (roughly) projective parsing algorithm, run in combination with pseudo-projective parsing. Therefore, according to parsing order the Multiplanar parsers are classified with Covington and arc-eager. The new non-projective decision tree is shown in Figure 3.

At the end of Phase 2, if 2-Planar arc-eager has been selected as best option, then MaltOptimizer tries with and without its

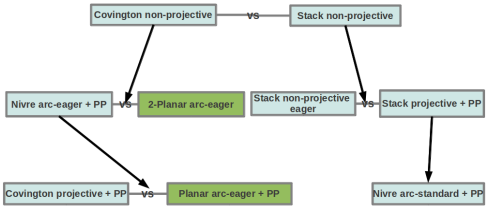


Figure 3: New decision tree for best non-projective algorithm (+PP for pseudo-projective parsing) in which Planar arc-eager and 2-Planar arc-eager are considered.

`-2pr` option, which can improve performance for some datasets by applying a REDUCE transition automatically after each SWITCH transition. There is also a `-prh` option, governing whether root arcs from the artificial root node 0 are constructed explicitly by the algorithm or left to be added automatically after parsing, and MaltOptimizer also explores it in order to find the best possible configuration.

3.3. Phase 3: Feature Selection

Once MaltOptimizer has gathered the preliminary parameters and the best parsing algorithm for the input data, it starts with the more challenging problem: the feature selection. MaltOptimizer tunes the feature model given all the parameters chosen so far, such as the parsing algorithm. It starts with default feature models for each parsing algorithm and then it tries with a greedy feature selection. In order to host Multiplanar parsers, we needed to modify the behavior in some cases. However, it is worth noting that the new processes are similar to the ones already implemented for Nivre arc-eager.

The features in MaltOptimizer can be explained dividing them into different feature windows: (i) part-of-speech, (ii) morphology, (iii) partially built dependency structure (dependency relation features), (iv) coarse grained part-of-speech, (v) extra morphological features such as gender or number (FEATS column) and (vi) lemma features.[9] For each window, it tries with backward selection experiments to ensure that all features in the default model for the given parsing algorithm are actually useful. After that, the system proceeds with forward selection experiments, trying potentially useful features

[8] To check statistics about non-projectivity in the training corpora, please visit http://ilk.uvt.nl/conll/paper_submission.html#table

[9] For an explanation of the different feature columns see (Buchholz and Marsi, 2006) or see http://ilk.uvt.nl/conll/#dataformat

one by one.

The major steps (and modifications to include Multiplanar parsers) of the feature selection experiments are the following:

1. Tune the window of POSTAG n-grams over the parser state.
   - It has been observed empirically (Gómez-Rodríguez and Nivre, 2010) that the planar algorithm tends to benefit from adding more features than in arc-eager, so for Multiplanar parsers we extend the search a bit over the buffer and the parser state.
   - For 2-Planar arc-eager we needed to distinguish between the active stack and the inactive stack, adding possible features involving the inactive stack when they are available (see Section 2). This fact was a challenge, because the inactive stack data structure is something new that is not available in any other parsing algorithm.
2. Tune the window of FORM features over the parser state. We added the same modifications as the ones added for POSTAG feature selection.
3. Tune DEPREL and POSTAG features over the partially built dependency tree.
4. Add POSTAG and FORM features over the input string.
5. Add CPOSTAG, FEATS, and LEMMA features if available. For 2-Planar arc-eager, MaltOptimizer tries with new features for the inactive stack when available.
6. Add conjunctions of POSTAG and FORM features.

Additionally, while in Nivre's arc-eager parser the first word in the buffer cannot be linked to any other, in Multiplanar parsers this can (and does) happen, so it makes sense to have features using the head word and dependency label associated with the first position of the buffer (Input[0]). This is a general modification that we included all over the process when Planar and 2-Planar are selected as best parsers.

4. Experiments

In this Section we present two different sets of experiments with corpora from the CoNLL-X Shared Task on multilingual dependency parsing (Buchholz and Marsi, 2006). We force the optimizer (by using the possible interaction between phases) to use Planar arc-eager and 2-Planar arc-eager as selected parsing algorithms,[10] in order to run a full feature selection for these two parsers and to observe how far we can go with the updated feature selection.

Table 1 shows the results for all the MaltOptimizer phases. Default and Phase 1 columns present the outcomes of the first phase in which Nivre arc-eager was selected as default parsing algorithm. Phase 2 columns, is divided in 2, in which we show the results of Planar and 2-Planar.[11] Finally, in Phase 3, we show the results of feature selection again for Planar and 2-Planar, interacting with the possibility that MaltOptimizer provides, stopping the process between phases. For phase 3, when the training corpora has non zero non-projective arcs/trees we run Planar arc-eager with pseudo-projective parsing, otherwise, we run it with default settings.

			Phase 2		Phase 3	
Language	**Default**	**Phase 1**	**Planar**	**2-Planar**	**Planar**	**2-Planar**
Arabic	63.02	63.03	62.81	63.42	65.53	64.94
Bulgarian	83.19	83.19	82.89	84.09	83.55	84.09
Chinese	84.14	84.14	81.66	82.79	83.63	83.54
Czech*	69.85	70.24	70.34	70.45	75.60	74.76
Danish	81.01	81.01	80.86	81.18	82.08	82.75
Dutch	74.77	74.77	76.55	76.81	76.55	81.41
German	82.36	82.36	81.34	82.28	84.64	84.84
Japanese	89.70	89.70	86.62	88.19	87.95	89.79
Portuguese	84.11	84.31	84.06	84.19	84.06	86.10
Slovene	66.08	66.52	65.94	66.43	69.72	70.13
Spanish	76.45	76.45	75.69	76.52	78.29	79.15
Swedish	83.34	83.34	82.38	82.83	83.67	83.78
Turkish	57.79	57.79	55.94	56.08	64.44	64.00

Table 1: Labeled attachment score per phase and with comparison to default settings for the training sets from the CoNLL-X shared task (Buchholz and Marsi, 2006). *The results for Czech were obtained using half of the training corpus due to a memory heap issue.

We can note that the performance improved substantially over all the selected corpora. However, for some data sets we got a much more substantial improvement, such as Turkish or Slovene, than in the other corpora.

[10] The Multiplanar parsers are not selected as best parsers for the corpora in which we carried out experiments with MaltOptimizer, due to the action of Stack non-projective and Stack projective parsing algorithms, which provide normally the highest results.

[11] For 2-Planar, we force MaltOptimizer to select between the 2-Planar options (`-2pr` and `-prh`) making it believe that 2-Planar is the best parsing algorithm for the data.

It is also worth noting that 2-Planar arc-eager normally provides better results, both in default settings (Phase 2) and in the optimized version (Phase 3), than Planar arc-eager. However, in some cases this is not the case, and the feature selection for Planar arc-eager reaches a higher attachment score.

5. Conclusions and Future Work

In this paper we have demonstrated that MaltOptimizer, which is an open-source system, can be updated with new parsing algorithms that are included (or will be) in MaltParser. This fact demonstrates that it is a very useful tool in order to get a reliable outcome when a user wants to use MaltParser for a new data set.

We have also demonstrated that the improvement is substantial, and we therefore suggest using MaltOptimizer when Planar and 2-Planar parsers are going to be used in comparison between parsers, so as to obtain reliable results.

As future work, we intend to add new feature selection algorithms, not only for Planar and 2-Planar parsers. Therefore, we are thinking on new procedures that select the best feature set possible and make them able to beat the results shown in previous publications of MaltParser feature selection and the present one. We could even consider a more complex version of MaltOptimizer, in which we run Phase 3 before Phase 2 for all the algorithms, or a subset of them, and then we could guarantee better results. However, for big data sets, it will take a lot of time to run several feature selection experiments.

Acknowledgments

MB has been funded by the Spanish Ministry of Education and Science (TIN2009-14659-C03-01 Project).
CGR has been partially funded by the Spanish Ministry of Economy and Competitiveness and FEDER (project TIN2010-18552-C03-02) and Xunta de Galicia (Rede Galega de Recursos Lingüísticos para unha Sociedade do Coñecemento).

References

Ballesteros, Miguel and Joakim Nivre. 2012a. MaltOptimizer: A System for MaltParser Optimization. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*.

Ballesteros, Miguel and Joakim Nivre. 2012b. MaltOptimizer: An Optimization Tool for MaltParser. In *Proceedings of the System Demonstration Session of the Thirteenth Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Beuck, Niels, Arne Köhn, and Wolfgang Menzel. 2011. Incremental parsing and the evaluation of partial dependency analyses. In *Proceedings of the Int. Conference on Dependency Linguistics (Depling 2011)*.

Buchholz, Sabine and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL)*, 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.

Covington, Michael A. 2001. A fundamental algorithm for dependency parsing. In *Proceedings of the 39th Annual ACM Southeast Conference*, pages 95–102.

Daelemans, Walter, Véronique Hoste, Fien De Meulder, and Bart Naudts. 2003. Combined optimization of feature selection and algorithm parameters in machine learning of language. In Nada Lavrac, Dragan Gamberger, Hendrik Blockeel, and Ljupco Todorovski, editors, *Machine Learning: ECML 2003*, volume 2837 of *Lecture Notes in Computer Science*. Springer.

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.

Gómez-Rodríguez, Carlos and Daniel Fernández-González. 2012a. Dependencias no dirigidas para el análisis basado en transiciones. *Procesamiento del Lenguaje Natural*, 48:43–50.

Gómez-Rodríguez, Carlos and Daniel Fernández-González. 2012b. Dependency parsing with undirected graphs.

In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 66–76, Avignon, France, April. Association for Computational Linguistics.

Gómez-Rodríguez, Carlos and Joakim Nivre. 2010. A transition-based parser for 2-planar dependency structures. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1492–1501.

Kool, Anne, Jakub Zavrel, and Walter Daelemans. 2000. Simultaneous feature selection and parameter optimization for memory-based natural language processing. In A. Feelders, editor, *BENELEARN 2000. Proceedings of the Tenth Belgian-Dutch Conference on Machine Learning*. Tilburg University, Tilburg, pages 93–100.

Krivanek, Julia and Detmar Meurers. 2011. Comparing rule-based and data-driven dependency parsing of learner language. In *Proceedings of the Int. Conference on Dependency Linguistics (Depling 2011)*.

Nilsson, Peter and Pierre Nugues. 2010. Automatic discovery of feature sets for dependency parsing. In *COLING*, pages 824–832.

Nivre, Joakim. 2003. An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT)*, pages 149–160.

Nivre, Joakim. 2008. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34:513–553.

Nivre, Joakim. 2009. Non-projective dependency parsing in expected linear time. 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 (ACL-IJCNLP)*, pages 351–359.

Nivre, Joakim and Johan Hall. 2010. A quick guide to MaltParser optimization. Technical report, maltparser.org.

Nivre, Joakim, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task of EMNLP-CoNLL 2007*, pages 915–932.

Nivre, Joakim, Johan Hall, and Jens Nilsson. 2006. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 2216–2219.

Nivre, Joakim, Johan Hall, Jens Nilsson, G. Eryiğit, and S. Marinov. 2006. Labeled pseudo–projective dependency parsing with support vector machines. In *CoNLL–X*.

Nivre, Joakim, Marco Kuhlmann, and Johan Hall. 2009. An improved oracle for dependency parsing with online reordering. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, pages 73–76.

Nivre, Joakim and Jens Nilsson. 2005. Pseudo-projective dependency parsing. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 99–106.

Ott, Niels and Ramon Ziai. 2010. Evaluating dependency parsing performance on German learner language. In *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories (TLT9)*, NEALT Proceeding Series.

Prudhvi Kosaraju, Sruthilaya Reddy Kesidi, Vinay Bhargav Reddy Ainavolu and Puneeth Kukkadapu. 2010. Experiments on indian language dependency parsing. In *ICON-2010 Tools Contest on Indian Language Dependency Parsing. Kharagpur, India.*

Yli-Jyrä, Anssi Mikael. 2003. Multiplanarity – a model for dependency structures in treebanks. In Joakim Nivre and Erhard Hinrichs, editors, *TLT 2003. Proceedings of the Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modelling in Physics, Engineering and Cognitive Sciences*, pages 189–200, Växjö, Sweden, 14-15 November. Växjö University Press.

Proyectos

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 181-184 recibido 12-04-12 revisado 31-05-12 aceptado 05-06-12

IARG-AnCora: Anotación de los corpus AnCora con argumentos implícitos

IARG-AnCora: Annotating AnCora corpus with implicit arguments

Mariona Taulé, M. Antònia Martí, Aina Peris
CLiC-Universidad de Barcelona
Gran Via 585, 08007 Barcelona
{mtaule,amarti,aperis}@ub.edu

Horacio Rodríguez
TALP-Universidad Politécnica de Cataluña
Jordi Girona Salgado 1-3, Barcelona
horacio@lsi.upc.edu

Lidia Moreno
ELiRF-Universidad Politécnica de Valencia
Camino de Vera s/n, 46020 Valencia
lmoreno@dsic.upv.es

Paloma Moreda
GPLSI-Universidad de Alicante
Campus de San Vicente del Raspeig,
03080, Alicante
moreda@dlsi.ua.es

Resumen: IARG-AnCora tiene como objetivo la anotación con papeles temáticos de los argumentos implícitos de las nominalizaciones deverbales en el corpus AnCora. Estos corpus servirán de base para los sistemas de etiquetado automático de roles semánticos basados en técnicas de aprendizaje automático. Los analizadores semánticos son componentes básicos en las aplicaciones actuales de las tecnologías del lenguaje, en las que se quiere potenciar una comprensión más profunda del texto para realizar inferencias de más alto nivel y obtener así mejoras cualitativas en los resultados.
Palabras clave: Estructura argumental, nominalización deverbal, argumentos implícitos, anotación de corpus, recursos lingüísticos

Abstract: Iarg-AnCora aims to annotate the implicit arguments of deverbal nominalizations in AnCora corpus. This corpus will be the basis for systems of automatic semantic role labeling based on machine learning techniques. Semantic analyzers are essential components in the current applications of language technologies, in which it is important to obtain a deeper understanding of the text to make inferences on the highest level in order to obtain qualitative improvements in the results.
Keywords: Argument structure, deverbal nominalization, implicit arguments, corpus annotation, linguistic resources

1 Motivación y Objetivos

Tradicionalmente el análisis de la estructura argumental se ha centrado principalmente en los predicados verbales, aunque recientemente se ha extendido también a los predicados nominales y adjetivales. En la gran mayoría de estas propuestas la identificación de los argumentos se restringe a los que aparecen en la oración, en el caso de los verbos, y al SN, en el caso de los nombres, es decir a los argumentos explícitos. Lo mismo es aplicable a los sistemas de etiquetado automático de roles semánticos (*Semantic Role Labeling*), que utilizan estos recursos para el aprendizaje del modelo de etiquetado, la mayoría de los cuales se aplican a verbos y sólo reconocen y clasifican los argumentos explícitos (Márquez et al., 2008), (Palmer, Gildea y Xue, 2010). Explicitar esta información permite una cobertura mucho más amplia del contenido semántico de los documentos (Gerber, Chai, y Meyers, 2009).

El proyecto *IARG-AnCora: Anotación de los corpus AnCora con argumentos implícitos*[1]

[1] Acción complementaria (FFI2011-13737-E), asociada al proyecto TextMess 2.0 (TIN2009-13391-C04-03/04).

tiene como objetivo principal enriquecer los corpus AnCora[2] del español y del catalán con la anotación de los argumentos implícitos de los predicados nominales derivados de verbos. Hasta el momento sólo se habían anotado los argumentos explícitos de estos nombres.

AnCora está formado por un corpus del catalán y otro del español de 500.000 palabras cada uno anotado con información morfológica (lema y categoría), sintáctica (constituyentes y funciones), semántica (estructura argumental, roles semánticos, entidades nombradas y sentidos nominales de WordNet) y pragmática (correferencia).

1.1 Argumento implícito

Se entiende por argumento implícito aquel argumento que no se realiza en el sintagma nominal (SN) el núcleo del cual es el nombre deverbal, pero que se encuentra en el contexto oracional (1) o textual (2) de la nominalización.

1) *Las escuelas de samba de Sao Paulo*$_{\textbf{iarg1-pat}}$ han conseguido **[**el **apoyo** *de la empresa privada*$_{\textbf{arg0-agt}}$ para mejorar las fiestas de carnaval**]**.

2) **"***El carnaval de Sao Paulo es feo***"**$_{\textbf{iarg1-pat}}$**,** dijo hoy *el alcalde de Río de Janeiro*$_{\textbf{iarg0-agt}}$ en una conversación informal con periodistas cariocas, y encendió la polémica. **[**Esa **opinión]** fue respaldada por el gobernador de Río de Janeiro, quien incluso fue más allá en su crítica al comentar que el carnaval que se organiza en Sao Paulo es "más aburrido que un desfile militar".

En el ejemplo (1), el nombre deverbal 'apoyo' tiene el argumento agente (arg0-agt) explicitado en el mismo SN, mientras que el argumento paciente 'las escuelas de samba de Sao Paulo' está implícito (iarg1-pat), porque se realiza en la misma oración pero fuera del SN. En el ejemplo (2), en cambio, el nombre deverbal 'opinión' aparece en el SN sin ningún argumento explícito. Sin embargo, tanto el argumento agente, 'el alcalde de Río de Janeiro', como el argumento paciente, 'el carnaval de Sao Paulo es feo', se consideran argumentos implícitos (iarg-agt y iarg-pat, respectivamente) porque se realizan en la oración previa. En la anotación actual del corpus AnCora, sólo los argumentos dentro del SN están anotados, por lo tanto, 'opinión' no tiene ningún argumento asociado y 'apoyo' sólo el argumento agente.

La tarea a desarrollar en este proyecto consiste básicamente en identificar los argumentos implícitos y asignarles una posición argumental –iarg0, iarg1, etc.– con el correspondiente papel temático (agente, paciente, causa, etc.).

Estos argumentos pueden recuperarse si se tiene en cuenta un contexto discursivo más amplio (Ruppenhofer et al., 2009). Tenerlos identificados es, por lo tanto, importante para poder proporcionar una interpretación semántica completa de las oraciones y textos.

2 Corpus con argumentos implícitos anotados

Sólo existen dos corpus que contengan los nombres deverbales anotados con argumentos implícitos y ambos son del inglés:

1) El corpus de entrenamiento y evaluación creado para llevar a cabo la tarea 10 de SemEval-2010, *Linking events and their participants in discourse*[3]. Se trata de un corpus formado por textos literarios de ficción y etiquetado siguiendo el esquema de anotación de FrameNet (Erk y Padó, 2004).
2) Un subconjunto de la sección de entrenamiento, desarrollo y evaluación del corpus periodístico Penn TreeBank (Marcus et al., 1993). El esquema de anotación sigue las propuestas de PropBank (Palmer et al., 2005) y NomBank (Meyers, Reeves, y Macleod, 2004) y (Meyers, 2007).

En el primer caso se han anotado un total de 3.073 ocurrencias que cubren distintos predicados nominales, pero cada uno de ellos asociado con un número pequeño de ocurrencias anotadas. En el segundo caso, sólo se han seleccionado los 10 nombres más frecuentes con sentidos no ambiguos, y se han anotado todas las ocurrencias del subconjunto del corpus en las que aparecen, un total de

[2]Los corpus AnCora están disponibles en la página web: http://clic.ub.edu/corpus/ancora.

[3]http://www.coli.uni-saarland.de/projects/semeval2010_FG/

1.253 (Gerber y Chai, 2010). En ambos corpus sólo se han anotado los argumentos implícitos 'nucleares' de nombres derivados de verbos, en ningún caso los argumentos adjuntos (o periféricos siguiendo la terminología de FrameNet[4]).

IARG-AnCora será el primer corpus anotado semánticamente con argumentos implícitos para el español y catalán. A diferencia de los corpus ingleses, la cobertura de IARG-AnCora será más amplia en dos sentidos: por un lado, se anotarán todos los argumentos implícitos de todas las ocurrencias nominales deverbales de los corpus AnCora (del orden de 23.000 aproximadamente para cada lengua); por el otro lado, se tendrán en cuenta tanto los argumentos implícitos 'nucleares' (iarg0, iarg1, iarg2, iarg3 y iarg4) como los argumentos adjuntos (iargM), entre los que se priorizarán los argumentos locativos (iargM-loc), temporales (iargM-tmp) y finales (iargM-fin).

Estas dos características los diferencian de los corpus anotados con argumentos implícitos.

3 Metodología

Se utilizará el mismo esquema de anotación que se ha adoptado en la anotación de los argumentos explícitos de las nominalizaciones deverbales (Peris y Taulé, 2011), que, a su vez, es el mismo que se ha utilizado para la anotación de la estructura argumental de los verbos (Taulé et al., 2008). El esquema de anotación sigue la propuesta de PropBank y NomBank enriquecida con papeles temáticos. De esta manera, se asegura la consistencia de la anotación de los argumentos entre diferentes predicados ─nombres y verbos─ así como la compatibilidad de los recursos del español y catalán con los del inglés.

En el caso de los argumentos implícitos, la etiqueta que se utilizará es $iarg_n$ para diferenciarlos de los argumentos explícitos (arg_n) (Gerber y Chai, 2010). La lista de papeles temáticos incluye 20 etiquetas distintas[5] ampliamente reconocidas en lingüística. La combinación de las 6 etiquetas argumentales (iarg0, iarg1, iarg2, iarg3, iarg4, iargM) con los distintos papeles temáticos da como resultado un total de 36 etiquetas semánticas posibles (iarg0-cau, iarg1-agt, iarg0-agt, iarg2-loc, etc.) que servirán para describir la relación semántica que se establece entre los argumentos y sus predicados.

La anotación del corpus con argumentos implícitos se realizará en dos etapas, la primera se llevará a cabo de manera automática y la segunda manualmente.

a) En la primera etapa se desarrollará un modelo de etiquetado de roles semánticos basado en técnicas de aprendizaje automático, cuyo objetivo será la identificación y clasificación de los argumentos implícitos y con el cual se etiquetará automáticamente todo el corpus (tanto la parte del español como del catalán). Este modelo se inferirá a partir de un corpus de entrenamiento anotado previamente de manera manual consistente en una muestra seleccionada de 500 ocurrencias nominales del corpus del español.
b) En una segunda etapa se procederá a la revisión manual de la anotación obtenida en el proceso automático anterior con el fin de garantizar la calidad final del recurso. Esta anotación manual servirá también para evaluar la precisión y cobertura del sistema automático desarrollado. Dado que dicho sistema se utilizará para la anotación automática del catalán, será posible analizar también el grado de portabilidad del sistema a otra lengua.

Tanto en el proceso automático como en el manual, se utilizarán los léxicos AnCora-Verb (Aparicio et al., 2008) y AnCora-Nom (Peris y Taulé, 2011) como fuentes léxicas a partir de las cuales obtener información de los argumentos implícitos posibles de cada predicado. Los argumentos potenciales a localizar en el contexto discursivo local, y posteriormente a etiquetar, serán aquellos que aparecen declarados en la entrada léxica nominal o verbal y que no están realizados dentro del SN el núcleo del cual es la nominalización deverbal.

El proyecto dará lugar, por un lado, a una versión enriquecida de los corpus AnCora con información sobre los argumentos implícitos de

[4] http://framenet.icsi.berkeley.edu/

[5] Los papeles temáticos son: 'agt' (agente), 'cau' (causa), 'exp' (experimentador), 'scr' (origen), 'pat' (paciente), 'tem' (tema), 'cot' (cotema), 'atr' (atributo), 'ben' (beneficiario), 'ext' (extensión), 'ins' (instrumento), 'loc' (locativo), 'tmp' (tiempo), 'mnr' (manera), 'ori' (origen), 'des' (destino), 'fin' (finalidad), 'ein' (estado inicial), 'efi' (estado final) y 'adv' (adverbial).

los nombres deverbales y, por otro, se dispondrá de un primer modelo para desarrollar un sistema de etiquetado de roles semánticos basado en rasgos para predicados nominales que contemple todos los argumentos, explícitos e implícitos.

4 Conclusiones

Disponer de corpus de amplia cobertura que incluyan la anotación tanto de los argumentos explícitos como de los implícitos y para predicados verbales y nominales, convierte dichos recursos en una fuente de conocimiento de gran valor. Los corpus AnCora del español y del catalán enriquecidos con dicha información, serán los primeros de esta extensión con este tipo de información. Estos corpus se podrán utilizar tanto para el estudio y análisis de la estructura argumental de nombres y verbos en general, como para derivar sistemas automáticos de etiquetado de roles semánticos que tengan en cuenta tanto los argumentos explícitos como implícitos de los nombres, para el estudio de las cadenas correferenciales, el referente de los SN, etc. La derivación de estos sistemas no se puede llevar a cabo si no se dispone de este tipo de recursos.

Bibliografía

Aparicio, J., M. Taulé y M.A. Martí, (2008). 'AnCora-Verb: A Lexical Resource for the Semantic Annotation of Corpora'. *Proceedings of 6th International Conference on of Language, Resources and Evaluation*. Marrakech, Morocco.

Erk K. y S. Padó, (2004). 'A powerful and versatile XML Format for representing role-semantic annotation'. *Proceedings of 4th International Conference on Language Resources and Evaluation*. Lisbon, Portugal.

Gerber, M., J. Chai, A. Meyers, (2009). 'The role of implicit argumentation in nominal SRL', *Proceedings of Human Language Technologies: NACL-2009*, pp. 146–154, Boulder, Colorado.

Gerber, M. y J.Y. Chai, (2010). 'Beyond NomBank: A Study of Implicit Argumentation for Nominal Predicates. *Proceedings of the ACL conference 2010*, pp. 1583–1592, Uppsala, Sweden, ACL.

Marcus, M., B. Santorini, y M. Marcinkiewicz, (1993). 'Building a large annotated corpus of English: the Penn treebank'. *Computational Linguistics*, 19:313-330.

Márquez, L., X. Carreras, C. Kenneth, Litkowski, y S. Stevenson, (2008). 'Semantic role labeling: an introduction to the special issue'. *Computational Linguistics*, 34(2):145–159.

Meyers, A., R. Reeves, y C. Macleod, (2004). 'NP-external arguments, a study of argument sharing in English. *Proceedings of the Workshop on Multiword Expressions: Integrating Processing (MWE'04)*, pp. 96–103, Stroudsburg, PA, USA. ACL.

Meyers, A. (2007). 'Anotation Guidelines for NomBank-Noun Argument Structure for PropBank'. Technical report, University of New York.

Palmer, M., Kingsbury, P. and Gildea, D. (2005): The Proposition Bank: An Annotated Corpus of Semantic Roles, *Computational Linguistics*, 21 (1). USA: MIT Press.

Palmer, M., D. Gildea, y N. Xue, (2010). Semantic Role Labeling. Synthesis on Human Languages Technologies. Morgan and Claypool Publishers.

Peris, A y M. Taulé, (2011). 'Annotating the Argument Structure of Deverbal Nominalizations in Spanish'. *Language Resources and Evaluation*, Springer. DOI 10.1007/s10579-011-9172-x.

Peris, A y M. Taulé, (2011). 'AnCora-Nom: A Spanish lexicon of deverbal nominalizations', *Procesamiento del Lenguaje Natural*, nº46, pp. 11-18.

Ruppenhofer, J., Sporleder, C., Morante, R., Baker, C, Palmer, M. (2009). Semeval-2010 task 10: Linking events and their participants in discourse. Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Direvtions (SEW 2009), p. 106-11, ACL, Boulder, Colorado.

Taulé, M., M.A., Martí, y M. Recasens (2008). 'Ancora: Multilevel Annotated Corpora for Catalan and Spanish'. *Proceedings of 6th International Conference on Language Resources and Evaluation.* Marrakesh Morocco.

Procesamiento de Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 185-188 recibido 14-04-12 revisado 30-05-12 aceptado 05-06-12

METANET4U: Enhancing the European Linguistic Infrastructure

METANET4U: Aumentar la Infraestructura Lingüística Europea

Núria Bel
Universitat Pompeu Fabra
Roc Boronat, 138
08028 Barcelona
nuria.bel@upf.edu

Asunción Moreno
Universitat Politècnica de Catalunya
Jordi Girona 1-3 Edifici D5
08034 Barcelona
asuncion.moreno@upc.edu

Resumen: El proyecto METANET4U está contribuyendo a la creación de una plataforma digital pan-europea que sustentará la distribución y el intercambio de recursos y servicios lingüísticos con el objetivo último de apoyar el desarrollo de aplicaciones basadas en tecnologías lingüísticas.
Palabras clave: recursos lingüísticos, infraestructuras

Abstract: METANET4U Project is participating in the creation of a digital pan-European platform that will support the distribution and the interchange of linguistic resources and services. The ultimate goal is to support the development of applications based on Language Technologies.
Keywords: Language Resources, infraestructures

1 Introducción

El proyecto METANET4U está contribuyendo a la creación de una plataforma digital pan-europea que sustentará la distribución y el intercambio de recursos y servicios lingüísticos con el objetivo último de apoyar el desarrollo de aplicaciones basadas en tecnologías lingüísticas. METANET4U es un proyecto co-financiado al 50% por el programa CIP-PSP (ICT *Policy Support Programme Competitiveness and Innovation framework Programme*) y por el consorcio participante. En él participan el *Institut Universitari de Lingüística Aplicada*[1], IULA, de la Universitat Pompeu Fabra y el *Centre de Tecnologies i Aplicacions del Llenguatge i la Parla*[2], TALP de la Universitat Politècnica de Catalunya. Coordinado por la Universidad de Lisboa, cuenta además con la participación de los siguientes centros: la Universidad de Manchester, la Universidad Alexandru Ioan Cuza, Insitutul de Cercetarri Pentru Inteligenta Artificiala y la Universidad de Malta.

El proyecto, que empezó en febrero de 2011, tiene prevista una duración de 24 meses y forma parte de la red de excelencia META-NET: *Multilingual Europe Technology Alliance*[3].

En este artículo presentamos en la sección 2 el contexto de iniciativas europeas que sirve de marco para entender las acciones de la red META-NET y de sus proyectos asociados. En la sección 3, se presenta la iniciativa META-SHARE para compartir recursos lingüísticos. En la sección 4, se presentan las contribuciones concretas realizadas en el proyecto METANET4U por las dos universidades españolas participantes y en la sección 5 se muestran las conclusiones.

2 METANET4U en contexto

El Séptimo Programa Marco de la Unión Europea (2007-2013) está teniendo una gran repercusión para el ámbito de las tecnologías lingüísticas en Europa. El Programa Marco ha

[1] http://www.iula.upf.edu
[2] http://www.talp.cat
[3] http://www.meta-net.eu

ISSN 1135-5948

dado financiación a un número importante de proyectos (hasta el momento más de 25 de ellos con participación española) en el ámbito de las tecnologías lingüísticas gracias a la creciente preocupación por la diversidad lingüística y las aplicaciones que estas tecnologías pueden tener en la llamada Europa digital. Además, varias acciones financiadas por la Comisión Europea han fomentado la creación de un espacio para la concertación y la cooperación entre todos los agentes y miembros de la comunidad de desarrolladores y de usuarios potenciales de estas tecnologías, con el objetivo de definir las prioridades para el óptimo desarrollo de este ámbito en el futuro. Este espacio se ha materializando en diferentes iniciativas, quizá las más conocidas han sido CLARIN[4], FLaReNet[5] y META-NET. Estas tres redes comparten el objetivo de fomentar el uso y la aplicación de tecnologías lingüísticas en diferentes aplicaciones, usos o audiencia. En particular han coincidido en proponer la visión de que es necesario crear una infraestructura que facilite la investigación y el desarrollo en este ámbito.

CLARIN (*Common Language Resources and Technology Infrastructure*, 2009-2011) ha sido la fase piloto de lo que es ya una infraestructura europea de investigación que pretende dar apoyo a los investigadores en humanidades para el acceso y explotación de textos gracias, entre otros factores, al uso de tecnologías lingüísticas. Ha desarrollado una red de instituciones colaboradoras de más de 200 nodos, y a partir de la participación de 9 Estados miembros de la Unión han constituido el CLARIN *European Research Infrastructure Consortium* (CLARIN-ERIC).

FLaReNet (Fostering Language Resources Network, 2009-2011) ha constituido una red de 99 miembros institucionales y 365 socios representando a 33 países diferentes, con la misión de consensuar y difundir las prioridades y objetivos estratégicos de la toda la comunidad relacionada con los la tecnología y recursos lingüísticos. Los resultados de los estudios y discusiones (Calzolari et al., 2012) se han hecho públicos y ya han sido enviados a representantes de instituciones europeas y nacionales para dar las claves de la futura coordinación del área.

En las primeras recomendaciones de FLaReNet ya se hacía énfasis en la necesidad de que la comunidad actúe de forma coordinada y como una unidad para conseguir mantener el apoyo a un área tan sensible para Europa: defender el multilingüismo que la caracteriza y desarrollar una industria que lo haga sostenible. Esta recomendación ha tenido ya un resultado en la constitución de la Red de Excelencia META-NET: *Multilingual Technology Alliance Network*, iniciada en 2010 gracias al proyecto T4ME financiado también por el 7PM. Esta red de excelencia ha puesto de manifiesto, ya en sus primeros estudios, la urgente necesidad de disponer de información, documentación y acceso a recursos y tecnologías de todas las lenguas de Europa y ha puesto en marcha META-SHARE, una plataforma para la creación de un catálogo y repositorio digital de recursos lingüísticos.

Para contribuir a poner en marcha la versión operativa de META-SHARE dotándola de contenidos que puedan acelerar su utilización, la Comisión coordinó la colaboración de tres proyectos del programa europeo *Information and Communication Technology Policy Support Programme* (ICT-PSP) y *Competitiveness and Innovation Framework Programme* (CIP): METANET4U[6], CESAR[7] y METANORD[8]. Estos proyectos tienen como misión: (i) extender los objetivos de META-NET difundiendo sus objetivos en el ámbito de los diferentes Estados europeos y (ii) compilar y aportar una masa crítica de recursos lingüísticos para la plataforma META-SHARE.

3 Acceso y disponibilidad de recursos lingüísticos en META-SHARE

En la actualidad hay un gran número de recursos lingüísticos que potencialmente pueden ser la base para diferentes tecnologías y aplicaciones: texto, datos lingüísticos, ficheros multimedia, herramientas, etc. Su descripción, archivo y mantenimiento a largo plazo tienen, no obstante, realizaciones diversas y heterogéneas.

Los repositorios digitales proporcionan ahora la infraestructura necesaria para hacer que la búsqueda y acceso a ellos sea no solamente posible sino también fácil. Estos repositorios

[4] http://www.clarin.eu
[5] http://www.flarenet.eu
[6] http://metanet4u.eu
[7] http://www.meta-net.eu/projects/cesar
[8] http://www.meta-nord.eu/

son una evolución del paradigma de las bibliotecas digitales en la que no solo se catalogan recursos, también se da acceso a los mismos recursos permitiendo la descarga, y se proporcionan nuevas capacidades de búsqueda basadas en descripciones formalizadas o metadatos.

META-SHARE tiene como objetivo ofrecer un repositorio digital distribuido de recursos y servicios lingüísticos que aporte estas nuevas funcionalidades y que sea un paso adelante en relación a otras iniciativas ya existentes de recopilación de información y distribución de recursos (el catálogo de ELRA, *European Language Resources Association*, o el modelo de distribución de ELDA, Evaluation and Language Distribution Agency, por ejemplo). META-SHARE quiere aportar también un marco de interoperabilidad entre recursos y tecnologías y ofrecer una infraestructura abierta que incluya recursos libres o bajo licencia, gratuitos o de pago. Además el proyecto está prestando atención a las cuestiones legales que a menudo comportan problemas para la distribución de estos recursos.

4 METANET4U

El objetivo del proyecto METANET4U es contribuir a la creación de esta plataforma digital pan-europea que sustentará la distribución y el intercambio de recursos y servicios lingüísticos con el objetivo último de apoyar el desarrollo de aplicaciones basadas en tecnologías lingüísticas. Este objetivo central se articula a partir de las siguientes iniciativas:

1) Recoger, organizar y difundir información sobre el estado de las actividades relacionadas con las tecnologías lingüísticas en las lenguas representadas por los participantes en el proyecto. Para ello ha secundado la iniciativa de META-NET de redactar una serie de libros blancos de las lenguas en Europa que identifiquen los beneficios que se pueden esperar de estas tecnologías y el nivel de disponibilidad para las diferentes lenguas europeas. METANET4U ha redactado 3 de estos informes: catalán, gallego y vasco. Antes de final del proyecto se editarán como libros independientes en la editorial Springer.

2) Recopilar recursos lingüísticos ya existentes y disponibles de las lenguas representadas en el proyecto para documentarlos siguiendo los esquemas de metadatos de la plataforma y, en su caso, convertirlos a los formatos que garanticen la interoperabilidad entre recursos y tecnologías. En este primer año, METANET4U ha hecho una primera entrega de 87 recursos que incluyen 27 corpus (monolingües, bilingües, alineados, etc.), 41 léxicos y diccionarios, 15 bases de datos de registros de habla y modelos de lenguaje, y 4 gramáticas y reglas de inferencia. En la mayoría de los casos las versiones de los recursos que se encuentran en los servidores META-SHARE son versiones revisadas, documentadas con metadatos para permitir la búsqueda inteligente, y, en su caso, convertidas a formatos estándar y a UTF8. Al final del proyecto, METANET4U habrá dado acceso a cerca de 300 recursos lingüísticos, entre datos y herramientas. Para conseguirlo, el proyecto tiene la misión de ponerse en contacto con desarrolladores de recursos externos al proyecto brindándoles la posibilidad de dar acceso a sus recursos desde META-SHARE además de aprovechar la posibilidad de que el proyecto lleve a cabo la adaptación a estándares y documentación específica necesarias. Así, en cuanto a recursos españoles, se han hecho versiones LMF de diccionarios tanto bilingües de acceso abierto como los de Apertium[9], como monolingües como el léxico PAROLE-SIMPLE catalán desarrollado por el Institut d'Estudis Catalans[10]; glosarios terminológicos multilingües, como el Banco de datos terminológico[11] da Universidade de Vigo, corpus anotados con información de dependencias, como los corpus AnCora[12] del grupo CLIC o el corpus Grial[13], del grupo GRIAL, ambos de la Universidad de Barcelona. Se ha abierto también una colaboración con la Universidade de Vigo y con la Universidad del País Vasco que, juntamente con el TALP han dado acceso a recursos lingüísticos relacionados con las tecnologías del habla, concretamente, se ha dado acceso a grandes bases de datos orales para entrenar sistemas de reconocimiento del habla en español, catalán, gallego y vasco, grabaciones preparadas para ser utilizadas en sistemas automáticos de conversión texto voz en español, catalán y vasco, tanto monolingües como bilingües, lexicones específicamente diseñados para ser incorporados en sistemas de

[9] http://www.apertium.org
[10] http://www.iec.cat
[11] http://sli.uvigo.es/TUVI/
[12] http://clic.ub.edu/corpus/AnCora
[13] http://grial.uab.es/sensem/download/main,es

reconocimiento automático del habla y en sistemas de conversión de texto a voz en las cuatro lenguas mencionadas, así como textos paralelos alineados y etiquetados para ser utilizados en el entrenamiento de sistemas de traducción automática entre el español y el catalán y el español y el gallego.

También se han incorporado a la plataforma META-SHARE programas para realizar automáticamente transcripciones fonéticas en catalán, español y gallego a partir de textos.

3) Difundir la existencia de estos recursos así como la disponibilidad de tecnologías que pueden hacer del espacio digital europeo un espacio multilingüe sin barreras. El público objetivo son investigadores, desarrolladores de aplicaciones y responsables políticos, sectores críticos para garantizar la diversidad lingüística en Europa.

METANET4U ha supuesto también la puesta en marcha de varios proyectos para mejorar y enriquecer recursos propios de los participantes. En el caso del IULA, se han estandarizado recursos que gracias a los grupos de investigación del IULA ya estaban a disposición de todos los usuarios para consulta *on line* en su página web. En particular, se está enriqueciendo el Corpus Tècnic de l'IULA (Cabré et al. 2006) hasta ahora solamente anotado con información morfosintáctica con información sintáctica y de dependencias. El IULA Treebank (Marimon et al. 2012) contribuirá así a la disponibilidad de más datos analizados sintácticamente del castellano que hasta ahora contaba con el corpus ANCORA (Taulé et al., 2008), y el UAM Spanish Treebank (Moreno and López, 1999).

Por su parte, el TALP, ha iniciado un proyecto para ampliar y mejorar recursos lingüísticos multimodales como son los seminarios interactivos CHIL, inicialmente consistentes en seminarios de una hora de duración grabados en inglés y transcritos ortográficamente. Se ha procedido a grabar seminarios en catalán y castellano y a extender la transcripción también a la parte de video incorporando información de movimientos, gestos, posiciones, y estados anímicos tanto del ponente como de los asistentes al seminario. Por otra parte, se ha realizado una mejora de los datos disponibles para realizar conversión texto a voz. Concretamente se ha realizado una compatibilización de los sistemas y anotaciones existentes en bases de datos en catalán y español (Bonafonte et al. 2008) para que puedan ser fácilmente integrados en el sistema de código abierto Festival.

5 Conclusiones

En este artículo se han presentado los proyectos que la Comisión Europea está cofinanciando para fomentar la participación del mayor número posible de grupos de investigación en la definición de estrategias para desarrollar tecnologías y recursos lingüísticos en Europa.

Se ha presentado también la iniciativa META-SHARE para crear una plataforma que facilite la disponibilidad pública de recursos y servicios lingüísticos así como el proyecto METANET4U gracias al cual se han puesto a disposición de los usuarios recursos relacionados con las lenguas oficiales habladas en España.

Bibliografía

Calzolari, N.; Quochi, V. y Soria, C. 2012, The Strategic Language Resource Agenda. En http://www.flarenet.eu/sites/default/files/FLaReNet_Strategic_Language_Resource_Agenda.pdf.

Marimon, M., Fisas, B., Bel, N., Arias, B., Vázquez, S., Vivaldi, J., Torner, S., Villegas, M. y Lorente, M. 2012. The IULA Treebank. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC2012)*, Estambul.

Moreno, A. and S. López. 1999. Developing a Spanish Tree Bank. In *Proc. Journées ATALA, Corpus annotés pour la syntaxe*. Paris, 18-19 June 1999.

Taulé, M.; M.A. Martí and M. Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. *Proceedings of 6th International Conference on Language Resources and Evaluation (LREC2008)*. Marrakesh.

Bonafonte, A.; J. Adell, I. Esquerra, S. Gallego, A. Moreno, J. Pérez . 2008. Corpus and Voices for Catalan Speech Synthesis. *Proc. of the Sixth Int. Conf. on Language Resources and Evaluation LREC 2008* Marrakech Marruecos

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 189-192 recibido 13-04-12 revisado 01-06-12 aceptado 06-06-12

Mejorando el Acceso, el Análisis y la Visibilidad de la Información y los Contenidos Multilingües y Multimedia en Red para la Comunidad de Madrid

Improving the Access, Analysis and Visibility of the Information and the Multilingual and Multimedia Net Content for the Madrid Community

F.Verdejo, R.Martínez , J.Cigarran, V. Fresno, A. García Serrano
(NLP\&IR-UNED)
P.Martínez
(LABDA-UC3M)
J M Pardo, Á Martínez
(THALES-UPM)

P. Castells, A. Moreno, D. Torre, I. Cantador, D. Vallet
(HLT\&IR-UAM)
A. Duarte
(GAVAB-URJC)
M. de Buenaga
(GSI-UEM)

Resumen: Presentación de las actividades del segundo programa de la red de investigación MAVIR de la Comunidad de Madrid.
Palabras clave: acceso a la información multimedia y multilingüe, análisis textual para la web social, evaluación centrada en el usuario.
Abstract: Presentation of the second program activities of the research net MAVIR of Madrid Region.
Keywords: multimedia and multilingual information access, textual analysis for social web, user centered evaluation.

1 Descrición general

MAVIR es una red de investigación cofinanciada por la Comunidad de Madrid y el Fondo Social Europeo en dos programas de I+D en TIC. En el segundo programa, actualmente en vigor y denominado $MA_2VI^{C+}R$, el núcleo del consorcio está formado por siete grupos de investigación de universidades y centros de la Comunidad de Madrid, en concreto: el Laboratorio de Cibermetría del CSIC, el grupo de Tecnologías del Lenguaje Humano y Recuperación de Información (HLT\&IR-UAM) de la UAM, el Laboratorio de Bases de Datos Avanzadas (LABDA-UC3M) de la UC3M, el Grupo de Sistemas Inteligentes (GSI-UEM) de la UEM y la UCM, el grupo de Tecnologías de Audio, Habla y Lenguaje Natural en Sistemas Inteligentes (THALES-UPM) de la UPM, el Grupo de Algorítmica aplicado a la Visión Artificial y la Biometría (GAVAB-URJC) de la URJC y finalmente el Grupo de Procesamiento del Lenguaje Natural y Recuperación de Información (NLP\&IR-UNED) de la UNED, que es el grupo coordinador del consorcio. Además hay 35 empresas, organismos culturales, hospitales y grupos de investigación de fuera de la Comunidad de Madrid. Todos ellos colaboran en las actividades del consorcio en calidad de "entidades asociadas".

En este segundo programa se han incorporado investigadores que trabajan en tecnologías del habla (UPM y UAM), recuperación de información (UAM) y visión y algorítmica (URJC). Estas incorporaciones han supuesto un incremento de la multidisciplinariedad del consorcio que nos ha permitido abordar nuevos retos relacionados con el

ISSN 1135-5948

tratamiento de contenidos multimedia y los sistemas de recomendación.

Las técnicas, recursos y aplicaciones desarrolladas por los grupos participantes en la línea de procesado de la información se encuentran detalladas en la web de MAVIR[2].

2 *Caso de estudio*

Hemos planteado un caso de estudio con el fin de articular y crear sinergia entre los diferentes grupos del consorcio, nos hemos centrado en la línea Acceso a la Información Multilingüe y Multimedia, y en particular en el objetivo "Interacción con el usuario: aplicación al dominio audiovisual".

La metodología seguida fue la siguiente:

- Elaboración de escenarios y casos de uso donde las diferentes tecnologías se complementan ofreciendo un valor añadido novedoso respecto al estado del arte
- Definición de la funcionalidad requerida
- Diseño arquitéctonico determinando módulos y componentes que implementan las tecnologías identificadas .
- Elección de un escenario para la elaboración de un demostrador en un dominio, considerando la disponibilidad de fuentes de información multimedia.

En la figura 8 se pueden ver todos los módulos involucrados en el escenario. Podemos distinguir cuatro niveles, un primer nivel de procesado de la informacion multimedia para la extracción de características y transcripción de la misma. Un segundo nivel en donde se lleva a cabo la anotacion e indexacion de la informacion transcrita, un tercer nivel que incluye la recuperación y personalización de la informacion, y un último en donde se realiza la presentación de información y la interacción con el usuario. En cada nivel aparecen los diferentes componentes con las siglas que identifican al miembro del consorcio responsable del mismo.

2.1 Escenario ilustrativo

El sistema de búsqueda MAVIR dispone de una colección de 31 vídeos sobre el patrimonio nacional. Estos vídeos son documentos audiovisuales de unos 2 ó 3 minutos de duración en los que un narrador describe en castellano aspectos históricos, artísticos, arquitectónicos y socio-culturales de un conjunto de ciudades españolas, patrimonio nacional. MAVIR dispone de una serie de módulos software que automáticamente procesan la imagen y el audio de los vídeos para extraer de ellos información muy variada: índices visuales, transcripciones de audio, y metadatos (ver figuras 1, 2 y 3).

Figura 1: Tratamiento de video

Figura 2: Transcripción de habla

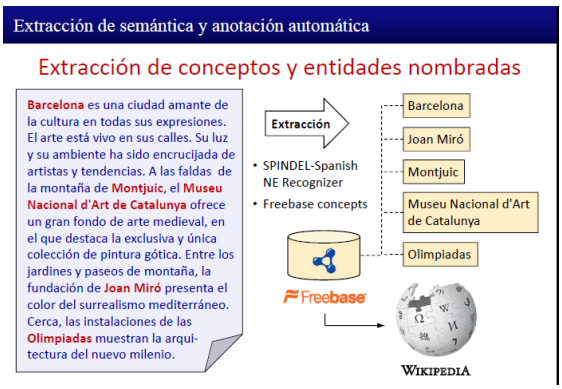


Figura 3: Extracción de conceptos y entidades nombradas.

Un usuario planea una excursión de fin de semana e introduce en el sistema de búsqueda MAVIR la consulta "arte gótico", el sistema

responde con una serie de videos relevantes para dicha consulta, tal y como puede verse en la figura 4.

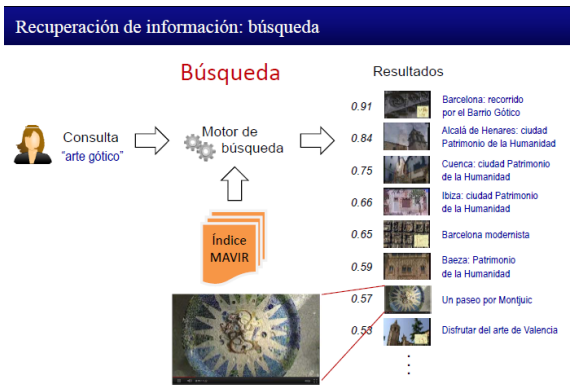


Figura 4: Resultados de búsqueda dada una consulta.

El usuario puede habilitar la búsqueda personalizada y el sistema de recomendación, de manera que el sistema personaliza los resultados de la búsqueda y muestra recomendaciones para el perfil del usuario (ver figuras 5 y 6).

Figura 5: Personalización

Para tener una impresión rápida sobre los contenidos recomendados, el usuario solicita una selección de "destacados" y el sistema selecciona vídeos y segmentos que reflejan valoraciones positivas (ver figura 7). Para una descripción completa puede consultarse http://ir.ii.uam.es/mavir/mavir-1.4-memoria.pdf.

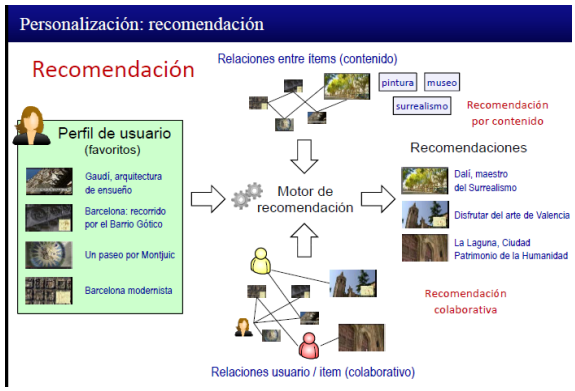


Figura 6: Recomendación

Figura 7: Análisis de opinión

3 *Conclusiones*

El desarrollo del caso de estudio ha supuesto para el consorcio un esfuerzo en cuanto a la puesta en común de las tecnologías desarrolladas por cada socio con el objetivo de identificar un escenario integrador. Dicho esfuerzo se ha traducido en una serie de beneficios:

- Posibilitar una visión integradora de las tecnologías y capacidades del consorcio y de las aplicaciones potenciales.
- Puesta en claro del potencial de convergencia y complementariedad que permitirá abordar futuros proyectos en común.
- Soporte para la colaboración e identificación de sinergias entre los grupos del consorcio.
- Identificación del valor añadido y potencial de transferencia tecnológica de las tecnologías desarrolladas.

Agradecimientos- $MA_2VI^{C+}R$ es un programa financiado por el plan regional de Ciencia y Tecnología de la Comunidad de Madrid

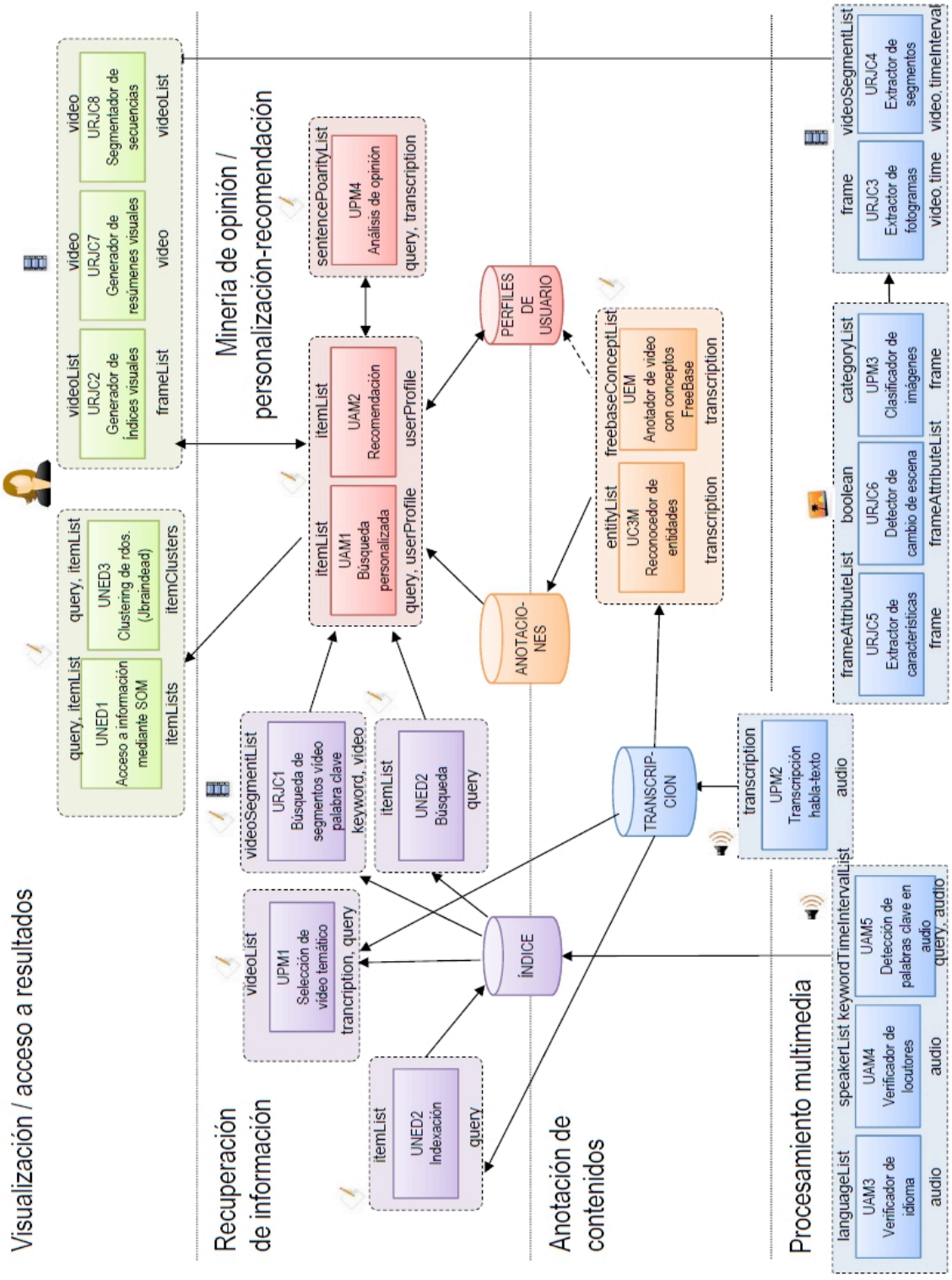


Figura 8: Esquema General

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 193-196 recibido 18-04-12 revisado 30-05-12 aceptado 05-06-12

Tratamiento de la dimensión espacial en el texto y su aplicación a la recuperación de información

Handling spatial dimension in text and its application to information retrieval

David Tomás, Fernando S. Peregrino, Fernando Llopis, Sonia Vázquez, Paloma Moreda, Estela Saquete, José M. Gómez
Depto. de Lenguajes y Sistemas Informáticos - Universidad de Alicante
{dtomas,fsperegrino,llopis,svazquez,moreda,stela,jmgomez}@dlsi.ua.es

Rubén Izquierdo
Induction of Linguistic Knowledge Research Group - Tilburg University
r.izquierdobevia@vu.nl

Óscar Ferrández
Department of Biomedical Informatics - University of Utah
oscar.ferrandez@utah.edu

Resumen: Proyecto emergente centrado en la desambiguación de topónimos y la detección del foco geográfico en el texto. La finalidad es mejorar el rendimiento de los sistemas de recuperación de información geográfica. Se describen los problemas abordados, la hipótesis de trabajo, las tareas a realizar y los objetivos parciales alcanzados.
Palabras clave: RI geográfica, desambiguación de topónimos, foco geográfico

Abstract: This project is focused on toponym disambiguation and geographical focus identification in text. The goal is to improve the performance of geographic information retrieval systems. This paper describes the problems faced, working hypothesis, tasks proposed and goals currently achieved.
Keywords: Geographic IR, toponym disambiguation, geographical focus

1. Datos del proyecto

Este proyecto está dirigido por David Tomás, miembro del Grupo de Procesamiento del Lenguaje y Sistemas de Información (GPLSI) de la Universidad de Alicante. Está financiado por la Universidad de Alicante (GRE10-33) y por la Generalitat Valenciana (GV/2012/110) dentro del programa de ayudas a proyectos emergentes.

Contacto

Email: dtomas@dlsi.ua.es
Teléfono: 965903400 ext. 2966
Dpto. de Lenguajes y Sistemas Informáticos,
Universidad de Alicante,
Carretera San Vicente del Raspeig s/n,
03690, Alicante, España.

2. Introducción

En los últimos años, debido a la implantación masiva de Internet en la empresa y en los hogares, se ha producido un incremento drástico de la información digital que se produce y distribuye. Los sistemas de *recuperación de información* (IR - *information retrieval*) surgen ante la necesidad de los usuarios de escudriñar este maremágnum de información digitalizada (Baeza-Yates y Ribeiro-Neto, 1999). Estos sistemas reciben una consulta por parte del usuario, devolviendo como resultado una lista de documentos relevantes a dicha petición. Esta lista se muestra ordenada siguiendo un criterio que intenta reflejar en qué medida cada documento contiene información que responde a las necesidades expresadas por el usuario. Los sistemas de IR más conocidos en la actualidad son aquellos que permiten localizar información en la Web. Google[1] y Bing[2] son dos claros exponentes de este tipo de sistemas.

[1] `http://www.google.com/`.
[2] `http://www.bing.com/`.

ISSN 1135-5948

Un aspecto que ha alcanzado especial relevancia en este tipo de sistemas es el tratamiento de la información geográfica. Estudios realizados sobre consultas efectuadas por usuarios de sistemas de IR en la Web (Gan et al., 2008), revelaron que las búsquedas de información delimitada geográficamente (p.ej. "hoteles en Alicante" o "altercados en París") suponen entre un 18 % y un 22 % del total de búsquedas realizadas. Esto supone una cantidad significativa de consultas que los sistemas actuales de IR basados en texto no son capaces de manejar de forma adecuada, ya que carecen del conocimiento suficiente para ubicar geográficamente (*georreferenciar*) los documentos consultados.

Los sistemas de *recuperación de información geográfica* (GIR - *geographic information retrieval*) son la respuesta dada por la comunidad científica a este problema. Estos sistemas suponen una especialización de los sistemas de IR, orientados a la indexación y recuperación de información relevante a una determinada región geográfica (Larson, 1996). Para su correcto funcionamiento, un sistema GIR debe ser capaz de realizar un análisis de la información espacial contenida en el documento, detectando las entidades geográficas que aparecen en él (o cercanas en el espacio) y determinando la relevancia de éstas con respecto al texto (es decir, si simplemente se nombran o si realmente el documento contiene información de interés sobre ellas). Para realizar este análisis de forma correcta, es necesario llevar a cabo dos tareas: la *desambiguación de topónimos* y la identificación del *foco geográfico*.

La *desambiguación de topónimos* es la tarea de asignar una representación formal (por ejemplo, unas coordenadas geográficas, una entrada en una base de datos o una localización dentro de una ontología geográfica) a las localizaciones espaciales (*topónimos*) identificadas en el texto (Rauch, Bukatin, y Baker, 2003). La ambigüedad en los topónimos puede ser de dos tipos: *geo/no-geo* y *geo/geo*. El primer tipo de ambigüedad se da cuando existe confusión entre un topónimo y un término que no lo es (por ejemplo, cuando en un texto "Washington" hace referencia a "Jorge Washington" y no a la ciudad). El segundo tipo de ambigüedad es el que se produce cuando dos localizaciones tienen el mismo nombre. En un estudio realizado por Smith y Crane (2001), se obtuvo que el 92 % de todos los nombres de lugar que ocurrían en su corpus de trabajo eran ambiguos. Otro estudio realizado por Roberts, Bejan, y Harabagiu (2010) reveló que el 83 % de los topónimos que aparecían en el texto presentaban ambigüedad, y que el 60 % de ellos tenía más de 5 posibles resoluciones. Sirvan como ejemplo las 42 ciudades con el nombre de Londres, los 18 Jerusalem y 63 Springfields de Estados Unidos, o los más de mil San Jose y Santa Ana que hay en el mundo.

Por otra parte, la identificación del *foco geográfico* de un documento consiste en determinar la principal o principales localizaciones a las que hace referencia un texto de entre todas las que se nombran en él (Amitay et al., 2004). Esto implica determinar el grado de relevancia que tienen para un documento dado las entidades geográficas presentes en él. Si bien la desambiguación de topónimos es una tarea ampliamente tratada dentro del campo de los GIR, no todos los sistemas de este tipo determinan el grado de relevancia de las entidades geográficas que en él aparecen.

Existen dos aproximaciones fundamentales a la desambiguación de topónimos y la detección del foco geográfico: la aproximación *basada en mapas* y la aproximación *basada en conocimiento* (Buscaldi y Rosso, 2008). La primera aproximación se basa en el uso de información geográfica *cuantitativa*, empleando propiedades espaciales y geométricas de las localizaciones encontradas en el texto, como puede ser el cálculo de distancias entre lugares o el cálculo del centroide de un área geográfica (Smith y Crane, 2001). La segunda aproximación se basa en la utilización de información geográfica *cualitativa*, empleando herramientas de *procesamiento del lenguaje natural* (PLN) y conocimiento externo mediante el uso de diccionarios geográficos (*gazetteers*) y ontologías (Garbin y Mani, 2005).

3. Objetivos del proyecto

El objetivo principal de este proyecto es el análisis de la información espacial en el texto, afrontando para ello el problema de la desambiguación de topónimos y la identificación del foco geográfico de los documentos. Ambos problemas serán abordados desde la aproximación basada en conocimiento, empleando para ello herramientas de PLN y recursos como *gazetteers* y ontologías. A diferencia de las aproximaciones actuales, centradas exclusivamente en el uso de informa-

ción geográfica, nuestro objetivo es mejorar la desambiguación de topónimos y la detección del foco geográfico mediante la incorporación de conocimiento general del mundo (como entidades, roles, fechas y eventos). Esta investigación básica se completará con su aplicación a un sistema GIR y con el desarrollo de una interfaz de visualización de los resultados siguiendo un paradigma de navegación basado en mapas (Rauch, Bukatin, y Baker, 2003).

El interés de este proyecto viene dado por la necesidad de mejorar el tratamiento y la recuperación automática de información geográfica en los documentos. Entender las referencias geográficas mencionadas en páginas Web, noticias de prensa o emails, puede beneficiar enormemente el rendimiento de los sistemas de IR. Los usuarios podrían añadir criterios geográficos a sus consultas de forma que los motores de búsqueda las procesaran de manera inteligente. La información recuperaría de esta manera su dimensión espacial.

Las aplicaciones de este tipo de tecnología son múltiples. Por ejemplo, para un usuario interesado en un producto comercial, la distribución geográfica de las páginas que hablan sobre dicho producto podría indicarle en qué lugares es popular y en cuáles no. Otra utilidad inmediata es la restricción de búsquedas de información a una cierta región (por ejemplo, procesando sólo páginas que hablen de Alicante). De igual manera, este tipo de información podría servir para buscar correlaciones entre localizaciones y determinados términos: podría detectarse qué lugares consideran los internautas que están más asociados con la moda, las fiestas, las vacaciones o la buena comida (Amitay et al., 2004). Un campo que podría beneficiarse enormemente de la información geográfica es el de la telefonía móvil, ya que podrían habilitarse una amplia variedad de servicios en esta plataforma basados en la localización del usuario (Baldauf y Simon, 2010).

4. Hipótesis de trabajo

La hipótesis seguida en este proyecto es que la información general del mundo asociada a las localizaciones geográficas puede mejorar la desambiguación de topónimos y la localización del foco geográfico en los documentos. La presencia en el texto de determinados eventos, nombres de personas, de organizaciones, fechas o incluso términos comunes, puede ser de gran utilidad para detectar de qué localidad concreta nos habla el texto (desambiguación de topónimos) y determinar su importancia con respecto al contenido del documento (detección del foco). Más aún, este tipo de información general podría servirnos para detectar el foco geográfico sin necesidad de que el nombre de la localización aparezca en el texto de forma explícita, infiriéndolo a partir de la aparición de determinados personajes, eventos, etc. relacionados con dicha localización.

Hasta donde alcanza nuestro conocimiento, el único sistema que ha empleado este tipo de información para la tarea de desambiguación de topónimos es el desarrollado por Roberts, Bejan, y Harabagiu (2010). En este trabajo incorporaban información de eventos, relacionando nombres de personas, organizaciones y otras localizaciones. En nuestro caso pretendemos ir más allá, incorporando también información relacionada con fechas y términos comunes que puedan ser representativos de un lugar (como pueden ser los nombres de determinadas comidas, expresiones artísticas, etc.). Además, pretendemos extender nuestra aproximación no sólo a la desambiguación de topónimos, sino también a la detección del foco geográfico.

5. Tareas a desarrollar

Para la consecución del proyecto será necesario completar el conjunto de tareas y subtareas que se mencionan a continuación.

Análisis del problema

En esta tarea se analizarán las distintas aproximaciones existentes a la detección de topónimos, su desambiguación y la identificación del foco geográfico. 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, dando como resultado un sistema capaz de detectar las entidades geográficas en un texto, desambiguarlas e identificar el foco geográfico de éste de forma automática. En este punto se evaluarán también los dos aspectos fundamentales de nuestra investigación: la desambiguación de topónimos y la detección del foco geográfico.

Construcción de un sistema GIR

Los sistemas de desambiguación de topónimos y de localización del foco geográfico se incorporarán a un sistema tradicional de IR, obteniendo un sistema GIR especializado en la localización de información georreferenciada.

Visualización de la información

En esta tarea se busca complementar el sistema GIR con una interfaz que permita la visualización y análisis de la información proporcionada por el sistema. Tener geolocalizada la información nos va a permitir ofrecer al usuario un nuevo paradigma de navegación, donde la interfaz visual es la propia superficie del planeta. Los resultados obtenidos para un determinado punto geográfico se mostrarán en un mapa, permitiendo al usuario una navegación espacial en busca de la información relacionada con el lugar que le interese (Rauch, Bukatin, y Baker, 2003). Por ejemplo, una consulta como "atentados de Al Qaeda" podría posicionar en el mapa, en las localidades correspondientes, toda la documentación que se considere relevante para esa consulta y ese lugar, dando al usuario la posibilidad de navegar por el mapa y acceder a la información en los lugares que resulten de su interés.

6. Situación actual del proyecto

Dentro de las tareas antes mencionadas, hasta el momento se ha completado el desarrollo de un sistema GIR (Peregino, Tomás, y Llopis, 2011) con el que se participó en la tarea GeoTime del NTCIR-9[3] y la creación de una interfaz de visualización de los resultados basada en OpenLayers.[4] Sobre este marco se incorporarán los avances que se vayan realizando en la desambiguación de topónimos y la localización del foco geográfico.

Bibliografía

Amitay, Einat, Nadav Har'El, Ron Sivan, y Aya Soffer. 2004. Web-a-where: geotagging web content. En *Proceedings of the 27th annual international ACM SIGIR conference*, SIGIR '04, páginas 273–280.

Baeza-Yates, Ricardo A. y Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Baldauf, Matthias y Rainer Simon. 2010. Getting context on the go: mobile urban exploration with ambient tag clouds. En *Proceedings of the 6th Workshop on Geographic Information Retrieval*, GIR '10, páginas 11:1–11:2.

Buscaldi, Davide y Paolo Rosso. 2008. Map-based vs. knowledge-based toponym disambiguation. En *Proceedings of the 2nd international workshop on Geographic information retrieval*, GIR '08, páginas 19–22.

Gan, Qingqing, Josh Attenberg, Alexander Markowetz, y Torsten Suel. 2008. Analysis of geographic queries in a search engine log. En *Proceedings of the first international workshop on Location and the web*, LOCWEB '08, páginas 49–56.

Garbin, Eric y Inderjeet Mani. 2005. Disambiguating toponyms in news. En *Proceedings of the conference on Human Language Technology*, HLT '05, páginas 363–370.

Larson, Ray R. 1996. Geographic information retrieval and spatial browsing. *GIS and Libraries: Patrons, Maps and Spatial Information*, páginas 81–124.

Peregino, Fernando S., David Tomás, y Fernando Llopis. 2011. University of alicante at ntcir-9 geotime. En *Proceedings of NTCIR-9 Workshop Meeting*, NTCIR-9, páginas 52–58.

Rauch, Erik, Michael Bukatin, y Kenneth Baker. 2003. A confidence-based framework for disambiguating geographic terms. En *Proceedings of the HLT-NAACL workshop on Analysis of geographic references*, GEOREF '03, páginas 50–54.

Roberts, Kirk, Cosmin Adrian Bejan, y Sanda M. Harabagiu. 2010. Toponym disambiguation using events. En *Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference*. AAAI Press.

Smith, David A. y Gregory Crane. 2001. Disambiguating geographic names in a historical digital library. En *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '01, páginas 127–136.

[3] `http://metadata.berkeley.edu/NTCIR-GeoTime/`.

[4] `http://openlayers.org/`.

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 197-200 recibido 07-05-12 revisado 22-05-12 aceptado 05-06-12

MILES (Modelos de Interacción centrados en Lenguaje, Espacio y Semántica computacional)*

MILES (Models of Interaction centred on Language, spacE and computational Semantics)

Pablo Gervás
U. Complutense de Madrid
Madrid
pgervas@sip.ucm.es

Angélica de Antonio
U. Politécnica de Madrid
Madrid
angelica@fi.upm.es

Gabriel Amores
U. de Sevilla
Sevilla
jgabriel@us.es

Resumen: La finalidad principal del proyecto es desarrollar una arquitectura para sistemas de interacción que conjugue un motor de diálogo, un generador de lenguaje natural, y una representación semántica basada en ontologías que abarque tanto el espacio (real o virtual) como el usuario que en él se ubica.
Palabras clave: Generación de lenguaje natural, ontologías, sistemas de diálogo

Abstract: The main goal of this project is to develop an architecture for interactive systems that combines a dialogue engine, a natural language generator, and a semantic representation based on ontologies covering both the real (or virtual) physical space and the user located within it.
Keywords: Natural language generation, ontologies, dialogue systems

1. Introducción

En los últimos años ha habido grandes avances en las áreas de investigación de sistemas de diálogo, generación de lenguaje natural, y representación de conocimiento mediante ontologías. No obstante, a pesar de que estas tres áreas tienen en común el tratar interacciones entre personas y máquinas con componentes importantes de lenguaje natural, ha habido muy poco trabajo centrado en la conjunción de las tres.

En este proyecto coordinado, en el que participan grupos de investigación de las universidades Complutense de Madrid (UCM), Politécnica de Madrid (UPM) y Universidad de Sevilla (USE), se aplican soluciones tecnológicas de interacción para resolver dificultades de percepción o localización que un usuario real pueda tener con un entorno espacial concreto. Para los propósitos del proyecto, se tienen en cuenta usuarios con distinto tipo de diversidad funcional (distintos grados de facilidad en la percepción o en la movilidad). Se están desarrollando tecnologías para trabajar con entornos reales y con simulaciones virtuales de los mismos. El sistema resultante ofrecerá una funcionalidad de diálogo mediante la cual el usuario puede interactuar con el sistema, que comprende sus peticiones y responderá coherentemente. Este tipo de sistema plantea la necesidad de resolver cuatro desafíos clave para hacer posibles las distintas funcionalidades contempladas. Primero, el sistema debe mantener un modelo del entorno. Segundo, debe mantener un modelo del propio usuario, sus capacidades y sus posibilidades de percepción. Tercero, debe ser capaz de generar mensajes en lenguaje natural para proporcionar al usuario descripciones e instrucciones acerca del entorno. Finalmente, debe ser capaz de mantener un diálogo multimodal con el usuario, conjugando todos estos ingredientes. Estos cuatro desafíos constituyen las claves de la investigación desarrollada.

El presente proyecto se basa en las siguientes hipótesis de partida. Para un sistema de diálogo es relevante el ser capaz de generar dinámicamente respuestas complejas en lenguaje natural integradas dentro de una arquitectura de presentación multimodal. Para un entorno virtual interactivo sería interesante establecer un diálogo fluido con el usuario. Finalmente, un sistema de generación de lenguaje natural debe ser capaz de enfrentarse a los retos específicos que puedan surgir de trabajar en contexto de diálogo y/o de aprovechar la riqueza de información disponible en entornos virtuales. Otra hipótesis importante para el proyecto es que las ontologías pueden constituir un nexo importante sobre

* Este proyecto de investigación ha sido financiado por el Ministerio de Innovación y Ciencia (TIN2009-14659-C03).

el que coordinar la interacción entre un sistema de diálogo, un generador de lenguaje natural, y un entorno virtual. La representación de conocimiento mediante ontologías ha sido objeto de investigación en tiempos recientes en distintos dominios de aplicación, pero su utilización en las áreas mencionadas se ha restringido hasta la fecha a prototipos exploratorios en cada área que validan la utilidad de la tecnología desde un punto de vista básico.

2. Trabajo previo

Como resultado de la investigación del grupo de investigación NIL (Natural Interaction based on Language) de la UCM en el contexto del proyecto GALANTE (TIN2006-14433-C02-01), financiado por el Plan Nacional de investigación, se desarrolló el framework TAP, diseñado para facilitar el desarrollo de aplicaciones de generación automática de lenguaje natural de forma que se pueda reutilizar el trabajo previamente desarrollado. Para ello se combinó trabajo previo sobre definición de arquitecturas genéricas para el desarrollo de aplicaciones de generación de lenguaje natural (Reiter y Dale, 2000; Cahill et al., 2001), con ideas de patrones de diseño (Gamma et al., 1995) y frameworks (Johnson y Foote, 1988). El framework TAP está capacitado para trabajar con representación de conocimiento en forma de ontologías. El grupo también ha participado con éxito en tareas de evaluación competitiva sobre generación de lenguaje natural relacionadas con la identificación de elementos observables a partir de descripciones textuales y con la generación de instrucciones para guiar a un usuario en un entorno virtual.

El grupo de investigación JULIETTA de la USE ha participado en numerosos proyectos de investigación y dispone de tecnología propia de sistemas de diálogo multimodal y procesamiento de lenguaje natural, desarrollada a lo largo de una trayectoria de más de 15 años de investigación. A lo largo de este tiempo ha desarrollado una serie de soluciones específicas para sistemas de diálogo que constituyen un punto de partida para la investigación que se está desarrollando en el proyecto MILES. Entre ellas se encuentran Episteme, un sistema de traducción automática que incluye en módulo completo de análisis léxico y gramatical junto con uno de generación basado en reglas, Delfos, un sistema de diálogo unimodal inspirado en el ISU approach que desarrolla un gestor de diálogo basado en expectativas, y MIMUS (de Amores Carredano et al., 2010), una evolución sobre Delfos, dotándolo de mecanismos específicos de multimodalidad, como estrategias de fusión y presentación multimodal. En particular, MIMUS pivota su funcionamiento sobre fuentes de conocimiento externo, específicamente sobre ontologías tipo OWL. MIMUS fue integrado con el generador TAP desarrollado por el grupo de la UCM en el contexto del proyecto coordinado DIVA-GALAN (TIN2006-14433), dentro del subproyecto GILDA.

El grupo de investigación de la UPM disponía por su parte de varios resultados de interés para este proyecto, obtenidos durante los últimos 10 años como consecuencia de sus trabajos de investigación en otros proyectos como MAEVIF (TIC00-1346) o ENVIRA (TIN2006-15202-C03), y de varias tesis doctorales desarrolladas en el ámbito del grupo:

- Un modelo preliminar de ontología semántica para entornos virtuales, especialmente centrada en la representación e inferencia de información espacial acerca del entorno 3D y los objetos en él situados, capaz de inferir relaciones espaciales y estructurales de diversos tipos entre objetos
- Un algoritmo de generación de trayectorias para la navegación a través de un entorno tridimensional constituido por diversos sub-entornos, con evitación de obstáculos, con la posibilidad de considerar áreas de tránsito de mayor o menor coste asociado, y con algunos mecanismos para hacer la trayectoria lo más similar posible a la forma de navegación humana en entornos reales.
- Un modelo de percepción visual computerizado que puede ser utilizado por un agente virtual inteligente para extraer información visual de su entorno, así como para modelar las capacidades perceptivas de los usuarios y anticipar sus posibilidades de percepción
- Una ontología para el modelado de estudiantes y un método de diagnóstico cognitivo especialmente adaptado para su aplicación en entornos virtuales

Sobre estas bases previas se consideró que

se daban las circunstancias necesarias para que los tres grupos acometieran de forma coordinada la tarea de demostrar empíricamente la validez de la hipótesis de partida.

3. Objetivo y plan de trabajo

El proyecto se inició en enero de 2010, y finalizará en diciembre de 2012. El proyecto coordinado ha contemplado tanto el desarrollo de una arquitectura de sistema de gestión de interacciones como su validación en una serie de casos de estudio particulares.

El subproyecto NOVA (Navegación basada en Ontologías mediante la Verbalización de mensajes de Ayuda) se ha centrado en la tarea de guiar a un usuario por un espacio físico que desconoce mediante instrucciones y descripciones verbales. Se ha prestado especial atención a usuarios con diversidad funcional (percepción o movilidad restringidas).

El subproyecto SEPIA (Modelos Semánticos del Entorno y la Persona para una Interacción Adaptativa en entornos virtuales) ha prestado especial atención al uso de ontologías para representación y razonamiento sobre conceptos espaciales, así como para el modelado del usuario, sus conocimientos previos relevantes, sus capacidades perceptivas, y los conocimientos que éste va adquiriendo a través de la interacción con el sistema. Se ha investigado acerca de las posibles interacciones entre dichas ontologías y el proceso de cálculo de trayectorias, su interpretación semántica para facilitar la generación de instrucciones, y el seguimiento y supervisión del recorrido del usuario a través del entorno.

El subproyecto DIMMO (DIalogos MultiModales basados en Ontologías) se ha centrado en desarrollar soluciones avanzadas de gestión de diálogo que tengan en cuenta la selección de movimientos del diálogo y la presentación multimodal guiada por ontologías.

El grupo Julietta de Sevilla ha aportado un motor de diálogo avanzado, validado ya en aplicaciones prácticas reales (Amores, Manchón, y Pérez, 2010). El grupo de la UCM ha aportadoa el motor de generación TAP, validado en diversas tareas de evaluación competitiva (Gervás, Hervás, y León, 2008; Hervás y Gervás, 2009; Gervás, 2011) y cuya integración con el motor de diálogo del grupo Julietta ya fue el objetivo del proyecto DIVAGALAN (TIN2006-14433). De cara a la coordinación entre estos dos grupos, el objetivo parcial en este proyecto ha sido comprobar que la integración realizada puede trasladarse a otros entornos de aplicación. El grupo de investigación de la UPM ha aportado su experiencia en el desarrollo de ontologías como modelos semánticos para guiar la interacción en entornos virtuales de simulación y aprendizaje. El motor de generación de la UCM ya se había integrado con las ontologías espaciales del grupo de la UPM en el marco del proyecto IVERNAO (CCG08-UCM/TIC-4300), financiado por la Comunidad de Madrid a lo largo del año 2009, que exploraba el papel de las ontologías de representación del espacio y el usuario a la hora de guiar a un usuario a través de un espacio físico que desconoce. La integración entre los módulos de UPM y USE se ha desarrollado mediante un interfaz basado en servicios web que hace accesible el diálogo verbal sobre el entorno virtual.

La arquitectura del sistema en desarrollo se basa en la propuesta realizada en la tesis doctoral (Méndez, 2008) de uno de los miembros del equipo UCM, Gonzalo Méndez, dirigida por la Investigadora principal del equipo UPM, Angélica de Antonio. En esta tesis se proponía una arquitectura que posibilitase la integración de distintos módulos de manera muy flexible, manteniendo la independencia entre ellos, lo que ha permitido que se puedan enchufar y desenchufar las aportaciones de cada grupo de manera transparente sin afectar a los módulos de los restantes grupos.

4. Conclusiones

El modelado del contexto de discurso interactivo que aporta un sistema de diálogo y la representación de relaciones espaciales complejas permitirán mejoras significativas en la naturalidad de las respuestas del generador de lenguaje natural que se contempla. La sustitución de los actuales textos enlatados utilizados en los entornos virtuales por texto generados dinámicamente permitirá ampliar la cobertura a mayor número de interacciones distintas con los usuarios. La posibilidad de generar respuestas con estructura narrativa, cuya información está extraída de una ontología que representa el entorno y su combinación con una presentación multimodal utilizando modelos estructurados como ontologías (de contexto, de usuario, etc.) constituirá una mejora significativa de la interacción para el usuario de un sistema de diálogo.

Es de esperar que el proyecto proporcione

resultados tanto en el campo de la generación de lenguaje natural, como en el campo de la representación de entornos virtuales o reales, como en el campo de los sistemas de diálogo. Dentro de la generación de lenguaje natural la exploración sistemática de los mecanismos de generación de indicaciones orientadas a la navegación basados en concepciones semánticas expresadas en forma de ontologías será una aportación valiosa. Para los sistemas de realidad virtual o aumentada, la construcción de representaciones semánticas, y no sólo geométricas, de entornos de trabajo, así como del usuario, sus características y capacidades, y también del contexto o situación y su evolución dinámica permitirá construir sistemas con un mayor grado de inteligencia y capacidad de adaptación. Su integración con mecanismos de generación de lenguaje natural y gestión de diálogo permitirá plasmar las nuevas posibilidades de interacción adaptativa del sistema de una forma muy natural para el usuario. En el contexto del proyecto MILES este objetivo se concreta en dar apoyo al usuario para la navegación a través del entorno, pero estos mecanismos podrían extenderse a otras tareas como la búsqueda de objetos, o la manipulación de elementos del entorno.

Para los sistemas de diálogo se explorará una arquitectura de presentación multimodal que incluya un proceso de decisión que escoja la presentación óptima utilizando modelos expresados mediante ontologías. Además se analizarán las características de alto y bajo nivel que influyen en la elección de movimientos de diálogo y se desarrollará un algoritmo de selección que mejore las tasas de acierto a nivel de análisis dentro de los sistemas de diálogo.

Se espera obtener resultados con posibilidad de transferencia a medio plazo, ya que se está en proceso de integrar las diferentes técnicas y tecnologías desarrolladas en una arquitectura software modular que permita diversos tipos de aplicaciones.

Bibliografía

Amores, J. Gabriel, Pilar Manchón, y Guillermo Pérez. 2010. Humanizing conversational agents: Indisys practical case study in ehealth. En Diana Perez-Marin y Ismael Pascual-Nieto, editores, *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices*.

Cahill, L., R. Evans, C. Mellish, D. Paiva, M. Reape, y D. Scott. 2001. The RAGS reference manual. Informe Técnico ITRI-01-08, ITRI, University of Brighton.

de Amores Carredano, José Gabriel, Guillermo Pérez García, Pilar Manchón Portillo, Carmen del Solar, Jesús González Martí, David Ávila Membrives, Antonio Ávila Membrives, y David Moral Alcázar. 2010. MIMUS: Gestión de diálogo multilingüe y multimodal. Informe técnico, Universidad de Sevilla.

Gamma, Erich, Richard Helm, Ralph Johnson, y John Vlissides. 1995. *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley.

Gervás, Pablo. 2011. UCM Submission to the Surface Realization Challenge. En *Generation Challenges 2011 Session at 13th European Workshop on Natural Language Generation (ENLG 2011).*

Gervás, Pablo, Raquel Hervás, y Carlos León. 2008. NIL-UCM: Most-Frequent-Value-First Attribute Selection and Best-Scoring-Choice Realization. En *Referring Expression Generation Challenge 2008, Proc. of the 5th International Natural Language Generation Conference (INLG'08)*, Ohio, USA, 06/2008.

Hervás, Raquel y Pablo Gervás. 2009. Evolutionary and Case-Based Approaches to REG: NIL-UCM-EvoTAP, NIL-UCM-ValuesCBR and NIL-UCM-EvoCBR. Athens, Greece, 04/2009.

Johnson, R. E. y B. Foote. 1988. Designing reusable classes. *Journal of Object-Oriented Programming*, 1(2):22–35, June/July.

Méndez, Gonzalo. 2008. *Una Arquitectura Software Basada en Agentes y Recomendaciones Metodológicas para el Desarrollo de Entornos Virtuales de Entrenamiento con Tutoría Inteligente*. Ph.D. tesis, Facultad de Informática, Universidad Politécnica de Madrid.

Reiter, Ehud y Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

Demostraciones

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 203-206 recibido 05-05-12 revisado 31-05-12 aceptado 05-06-12

InLéctor: Sistema de lectura bilingüe interactiva*

InLéctor: a system for bilingual interactive reading

Antoni Oliver, Marta Coll-Florit, Salvador Climent
Universitat Oberta de Catalunya
Avda. Tibidabo 39-43 08035 Barcelona
aoliverg,mcollfl,scliment@uoc.edu

Resumen: Este proyecto pretende desarrollar un sistema que genere libros bilingües, con audio e interactivos. El sistema ofrecerá diversos formatos de salida que permitan leer y escuchar los libros en diferentes dispositivos, como libros electrónicos, tabletas y ordenadores. Asimismo, ofrecerá la posibilidad de obtener libros paralelos impresos.
Palabras clave: libro electrónico, aprendizaje de lenguas

Abstract: The aim of this project is the development of a system for the generation of interactive bilingual electronic books with audio support. The system will offer several output formats for reading and listening the books on different devices such as electronic books, tablets and computers. It will also allow printing a parallel bilingual book.
Keywords: electronic book, language learning

1. Introducción

El proyecto InLéctor pretende fomentar la lectura en versión original, ofreciendo libros bilingües, en texto y audio, en un entorno de lectura interactiva. En este proyecto queremos fomentar la lectura en lenguas extranjeras ofreciendo una interactividad que permita conocer a priori las palabras más complicadas de un capítulo; enlaces al párrafo traducido, lo que permitirá al usuario pasar de la versión original a la versión traducida de manera inmediata; y ofrecer en algunas obras el audio correspondiente a una lectura humana.

2. Funcionalidades

2.1. Glosario interactivo

Antes de iniciar la lectura de un capítulo o fragmento de una obra, el usuario podrá especificar su nivel de lengua y obtendrá un glosario de las palabras más difíciles del texto. El objetivo es que el usuario aprenda el significado de estas palabras difíciles para así poder disfrutar de una lectura más ágil y con menos interrupciones.

2.2. Texto bilingüe

Las obras se ofrecerán en su lengua original y en su traducción a otra lengua (para nuestro proyecto el castellano o catalán). El original y la traducción estarán paralelizados a nivel de párrafo. Esto permitirá un cambio rápido de la versión original a la traducida.

2.3. Interacción entre los usuarios

El sistema permitirá compartir comentarios sobre la obra o dudas sobre un determinado fragmento que serán accesibles para todos los usuarios.

2.4. Recursos y herramientas

Las obras que se publiquen bajo este proyecto serán únicamente las que tengan los derechos de autor y de traducción libres, es decir, que estén en dominio público. De esta manera nos aseguramos que la distribución de las obras sea totalmente legal. Así, las obras literarias en versión original y las traducciones se extraerán principalmente de Wikisource[1] y del Proyecto Gutemberg[2].

* Este trabajo se ha llevado a cabo dentro del proyecto Know2 *Language understanding technologies for multilingual domain-oriented information access* (MICINN, TINN2009-14715-C04-04)

[1] http://wikisource.org/
[2] http://www.gutenberg.org/

ISSN 1135-5948

En cuanto a los audios, serán en su mayoría provenientes de la fuente LibriVox[3]. Librivox es un proyecto en el que un gran número de voluntarios leen capítulos de libros que están bajo dominio público, y publican también bajo dominio público los ficheros de audio.

Para generar los diccionarios bilingües se utilizarán fuentes libres como por ejemplo Wiktionary[4], WorNets libres (Bond y Paik, 2012) y los diccionarios de transferencia de Apertium (Forcada, Tyers, y Ramírez, 2009).

En cuanto al procesamiento de los textos, se realizará un etiquetado morfosintáctico con Freeling (Carreras et al., 2004) o Treetager (Schmid, 1994). Para obtener los texto paralelos se utilizará el alineador automático Humaling (Varga et al., 2007) o mAligna (Jassem y Lipski, 2008).

2.5. Formatos

El formato básico de trabajo será el DocBook, con atributos especiales que permitan la relación entre los diferentes párrafos de las dos versiones del texto. A partir de este formato básico se generarán los siguientes formatos de salida: HTML, Mobipocket, Epub y PDF.

2.6. Lenguas de trabajo y obras

Las lenguas de trabajo iniciales de este proyecto serán el inglés, francés y ruso al castellano o catalán dependiendo de la disponibilidad de las traducciones. Las primeras obras que está previsto tratar son: The adventures of Sherlock Holmes (Sir Arthur Conan Doyle), Les Trois Mousquetaires (Alexandre Dumas) y Игрок (El jugador) de Fiódor Dostoyevski con sus respectivas traducciones al castellano.

2.7. Público objetivo

Este sistema está pensado para todas aquellas personas con un nivel medio-avanzado de una lengua extranjera que deseen leer obras en versión original. El sistema puede ayudar a mantener un nivel adecuado de la lengua, de una manera amena. Su uso en docencia es interesante, ya que se puede aplicar tanto en el estudio de lenguas extranjeras, como en el estudio de la literatura e incluso en estudios de traducción.

3. Flujo de trabajo

En esta sección expondremos los detalles sobre el flujo de trabajo que estamos siguiendo para la creación de los libros paralelos. Por el momento estamos en la fase de creación de los enlaces entre el párrafo original y el párrafo traducido. Todavía no estamos trabajando la parte de enlaces entre el texto y el audio correspondiente a la locución humana del texto.

3.1. Descarga de la obra

Este primer paso no entraña ninguna dificultad. El objetivo es disponer del original y la traducción en sendos ficheros de texto plano.

```
THE ADVENTURES OF SHERLOCK HOLMES
Arthur Conan Doyle
ADVENTURE I. A SCANDAL IN BOHEMIA
I.
To Sherlock Holmes she is always THE
woman. I have seldom heard him mention
her under any other...
```

3.2. Creación de los ficheros Docbook del original

A partir del fichero de texto correspondiente a la obra en la lengua original se genera un fichero en formato Docbook. Este formato permite crear documentos en un formato independiente de la presentación final. En el formato Docbook se expresa tanto el contenido como la estructura lógica del documento, pero no su formato final. Para crear el fichero en formato Docbook se puede utilizar cualquier editor de textos que tenga un buen soporte de macros (en nuestro caso hemos utilizado JEdit[5]. Mediante la creación de unas pocas macros la conversión de texto plano a Docbook se puede realizar rápidamente.

[3] http://librivox.org/
[4] http://www.wiktionary.org/
[5] http://www.jedit.org/

```
<book>
<title>THE ADVENTURES OF SHERLOCK HOLMES</title>
<chapter>
<title>ADVENTURE I. A SCANDAL IN BOHEMIA</title>
<section>
<title>I.</title>
<para>To Sherlock Holmes she is always THE
woman. I have seldom heard him mention
her under any other...</para>
...
```

3.3. Creación del Docbook numerado

Mediante un sencillo *script* se numeran los títulos y los párrafos. Esta numeración es la que nos permitirá realizar los enlaces entre la versión original y traducida. Como podemos observar en el siguiente ejemplo, se sigue una numeración independiente para los títulos (ya sean de libro, capítulo o sección) y para los párrafos.

```
<book>
<title xml:id="t1-eng">THE ADVENTURES
OF SHERLOCK HOLMES</title>
<chapter>
<title xml:id="t2-eng">ADVENTURE I.
A SCANDAL IN BOHEMIA</title>
<section>
<title xml:id="t3-eng">I.</title>
<para xml:id="p1-eng">To Sherlock Holmes
she is always THE woman. I have seldom
heard him mention her under any other...</para>
...
```

3.4. Alineación y creación del TMX

Mediante los algoritmos de alineación citados anteriormente se alinean los ficheros de texto correspondientes al original y a la traducción. Aunque la alineación que nos interesa es a nivel de párrafo, la alineación se lleva a cabo a nivel de oración. A partir de la alineación se creará un fichero de memoria de traducción en el formato estándar TMX (*Translation Memory eXchange*).

```
<tu>
<tuv xml:lang="en">
<seg>
To Sherlock Holmes she is always THE woman.
</seg>
</tuv>
<tuv xml:lang="es">
<seg>
Ella es siempre, para Sherlock Holmes, la mujer.
</seg>
</tuv>
</tu>
<tu>
```

3.5. Creación del proyecto de traducción

En este paso creamos un proyecto de traducción con alguna herramienta de traducción asistida. Dado que todos los formatos de fichero son estándar, se puede utilizar cualquier herramienta. En nuestro proyecto utilizamos OmegaT[6] que es una herramienta de software libre. En el proyecto asignamos como documento a traducir el Docbook numerado correspondiente a la obra original y como memoria de traducción el fichero TMX creado en el paso anterior.

3.6. Verificación de la alineación

La verificación de la alineación se ha convertido en una tarea de traducción mediante una herramienta de traducción asistida donde la inmensa mayoría de las oraciones se traducirán directamente mediante la memoria de traducción. El encargado de verificar la alineación sólo tiene que preocuparse de verificar que la propuesta proveniente de la memoria sea la adecuada, lo que ocurrirá en la mayoría de casos. En los casos en que la propuesta no sea correcta o que simplemente no aparezca ninguna propuesta, el encargado podrá consultar el fichero correspondiente a la obra traducida para encontrar la traducción al segmento correspondiente.

3.7. Creación del Docbook numerado correspondiente a la traducción

La creación del Docbook numerado correspondiente a la traducción se reduce a la creación del fichero traducido mediante la herramienta de traducción asistida. Mediante la función de buscar y reemplazar de cualquier editor de textos reemplazaremos las marcas de lengua de la numeración de títulos y párrafos por la marca de lengua correspondiente.

[6] www.omegat.org

```
<book>
<title xml:id="t1-spa">LAS AVENTURAS
DE SHERLOCK HOLMES</title>
<chapter>
<title xml:id="t2-spa">AVENTURA I.
ESCÁNDALO EN BOHEMIA</title>
<section>
<title xml:id="t3-spa">I.</title>
<para xml:id="p1-spa">Ella es siempre,
para Sherlock Holmes, la mujer. Rara
vez le he oído hablar de ella
aplicándole otro ...</para>
...
```

3.8. Creación de los formatos finales

A partir de los ficheros docbook numerado correspondientes al original y a la traducción creamos, mediante un simple *script* el libro en formato html con los enlaces entre los párrafos originales y traducidos.

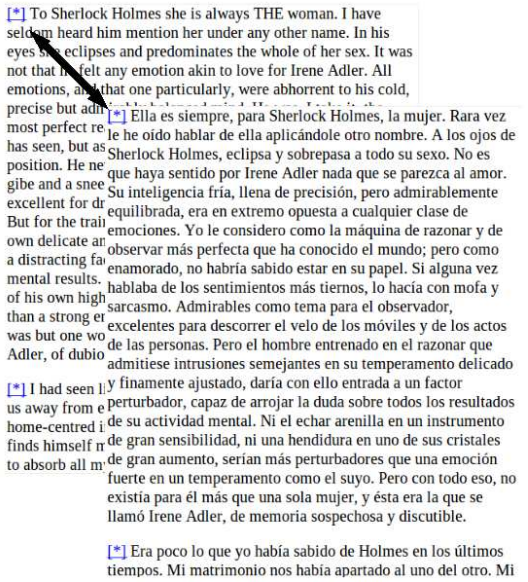
```
<h1><a name="t1-eng"/><a href="#t1-spa">[*]</a>
THE ADVENTURES OF SHERLOCK HOLMES</h1>
<h2><a name="t2-eng"/><a href="#t2-spa">[*]</a>
ADVENTURE I. A SCANDAL IN BOHEMIA</h2>
<h3><a name="t3-eng"/><a href="#t3-spa">[*]</a>
I.</h3>
<p><a name="p2-eng"/><a href="#p2-spa">[*]</a>
To Sherlock Holmes she is always THE woman. I
have seldom heard him mention her under any...
....
....
<h2><a name="t1-spa"/><a href="#t1-eng">[*]</a>
LAS AVENTURAS DE SHERLOCK HOLMES</h2>
<h2><a name="t2-spa"/><a href="#t2-eng">[*]</a>
AVENTURA I. ESCÁNDALO EN BOHEMIA</h2>
<h3><a name="t3-spa"/><a href="#t3-eng">[*]</a>
I.</h3>
<p><a name="p2-spa"/a><a href="#p2-eng">[*]</a>
Ella es siempre, para Sherlock Holmes, la mujer.
Rara vez le he oído hablar de ella aplicándole..
```

A partir de este html con enlaces se utiliza el programa Calibre[7] para transformarlo a los formatos epub y mobipocket. Esta herramienta, que es de software libre, también nos sirve para editar metadatos, añadir una portada, etc.

4. Conclusiones

En este artículo hemos presentado un proyecto para la creación de libros digitales bilingües interactivos con soporte de audio. El objetivo es facilitar la lectura en lengua extranjera y mejorar el nivel de lengua. Actualmente el mercado del libro se

[7] http://calibre-ebook.com/


[*] To Sherlock Holmes she is always THE woman. I have seldom heard him mention her under any other name. In his eyes she eclipses and predominates the whole of her sex. It was not that he felt any emotion akin to love for Irene Adler. All emotions, and that one particularly, were abhorrent to his cold, precise but admi

[*] Ella es siempre, para Sherlock Holmes, la mujer. Rara vez le he oído hablar de ella aplicándole otro nombre. A los ojos de Sherlock Holmes, eclipsa y sobrepasa a todo su sexo. No es que haya sentido por Irene Adler nada que se parezca al amor. Su inteligencia fría, llena de precisión, pero admirablemente equilibrada, era en extremo opuesta a cualquier clase de emociones. Yo le considero como la máquina de razonar y de observar más perfecta que ha conocido el mundo; pero como enamorado, no habría sabido estar en su papel. Si alguna vez hablaba de los sentimientos más tiernos, lo hacía con mofa y sarcasmo. Admirables como tema para el observador, excelentes para descorrer el velo de los móviles y de los actos de las personas. Pero el hombre entrenado en el razonar que admitiese intrusiones semejantes en su temperamento delicado y finamente ajustado, daría con ello entrada a un factor perturbador, capaz de arrojar la duda sobre todos los resultados de su actividad mental. Ni el echar arenilla en un instrumento de gran sensibilidad, ni una hendidura en uno de sus cristales de gran aumento, serían más perturbadores que una emoción fuerte en un temperamento como el suyo. Pero con todo eso, no existía para él más que una sola mujer, y ésta era la que se llamó Irene Adler, de memoria sospechosa y discutible.

[*] Era poco lo que yo había sabido de Holmes en los últimos tiempos. Mi matrimonio nos había apartado al uno del otro. Mi

Figura 1: Enlace entre párrafo original y traducido

encuentra inmerso en un cambio de paradigma y el paso de formato papel a formato digital. El avance en este cambio es lento, al menos en nuestro país, ya que en pocos casos la edición digital comporta una mejora substancial para el potencial cliente, ni en precio, ni en funcionalidades. En nuestro proyecto tratamos únicamente con obras en dominio público, pero las editoriales que tengan los derechos de autor y de traducción de una obra se pueden plantear seriamente la posibilidad de publicar las obras en este formato. El libro digital se asemejaría a una película en DVD, donde el usuario puede escoger la lengua y los subtítulos y adaptar la visualización a sus preferencias.

Actualmente el proyecto no cuenta con financiación específica por lo que el avance es lento. Actualmente disponemos de unas pocas muestras de textos paralelos en diferentes formatos que se pueden descargar de http://lpg.uoc.edu/InLector. Nuestra intención es obtener financiación externa para mejorar las herramientas de creación de los libros paralelos. También estamos abiertos a colaboraciones externas ya sea en la mejora de las herramientas como en la creación de nuevas obras.

Bibliografía

Bond, Francis y Kyonghee Paik. 2012. A survey of wordnets and their licenses. En *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, páginas 64–71, Matsue, Japan.

Carreras, X., I. Chao, L. Padró, y M. Padró. 2004. Freeling: An open-source suite of language analyzers. En *Proceedings of the 4th LREC*, volumen 4.

Forcada, M. L, F. M Tyers, y G. Ramírez. 2009. The apertium machine translation platform: five years on. En *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, página 3–10.

Jassem, K. y J. Lipski. 2008. A new tool for the bilingual text aligning at the sentence level. *Intelligent Information Systems*, página 279–286.

Schmid, H. 1994. Probabilistic part-of-speech tagging using decision trees. En *Proceedings of international conference on new methods in language processing*, volumen 12, página 44–49.

Varga, D., P. Halácsy, A. Kornai, V. Nagy, L. Németh, y V. Trón. 2007. Parallel corpora for medium density languages. En *Proceedings of the RANLP 2005*, páginas 590–596.

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 209-212 recibido 05-05-12 revisado 29-05-12 aceptado 05-06-12

Prototipo buscador de información médica en corpus multilingües y extractor de información sobre fármacos.

Medical information search and information extraction about drugs prototype on a multilingual corpus.

Daniel Sánchez-Cisneros[1], Sara Lana[2], Antonio Moreno[3], Leonardo Campillos[3]
Paloma Martínez[1], Isabel Segura-Bedmar[1]
[1]Departamento de Informática,
Universidad Carlos III de Madrid
{dscisner, pmf, isegura}@inf.uc3m.es
[2]Universidad Politécnica de Madrid
Carretera de Valencia, Km 7
slana@diatel.upm.es
[3]Departamento de Lingüística General
Universidad Autónoma de Madrid
{antonio.msandoval, leonardo.campillos}@uam.es

Resumen: La investigación y desarrollo de nuevos fármacos ha provocado un crecimiento exponencial de la documentación relacionada con el dominio farmacológico y en la industria farmacéutica. Esto ha supuesto un problema para los profesionales del sector, debido a que tienen que invertir una gran cantidad de tiempo y esfuerzo en la revisión de esta documentación para mantener actualizados sus conocimientos. Este trabajo presenta un prototipo que busca información sobre términos médicos en colecciones divulgativas de medicina multilingües (en inglés, español, árabe y japonés) indexadas según conceptos de UMLS. El prototipo también detecta los fármacos y sus interacciones presentes en los textos.
Palabras clave: Recuperación de Información, Extracción de información.

Abstract: Research and development of new drugs has caused an exponential growth of the documentation related to drug control in the pharmaceutical industry. This is a problem for professionals since they have to invest a lot of time and effort in reviewing this documentation. This paper presents a prototype that is able to search information about medical terms over a multilingual collections of documents (English, Spanish, Arabic and Japanese), indexed with UMLS concepts. The prototype also detects drugs and drug-drug interactions discovered in the texts.
Keywords: Information Retrieval, Information Extraction.

1 Introducción

En los últimos años el tamaño de la documentación científica ha sufrido un crecimiento exponencial debido a los avances de investigación en áreas como la química, la medicina o la biología. Gran parte de esta información se almacena en grandes bases de datos bibliográficas que recopilan estos trabajos científicos. En el ámbito de la farmacovigilancia, todos los días se reportan nuevos efectos adversos, lo que supone un importante problema para los profesionales del sector de la salud, que necesitan invertir gran cantidad de tiempo y esfuerzo en la revisión de toda la documentación publicada sobre efectos adversos, y en particular, sobre interacciones farmacológicas.

En este trabajo se describe un prototipo que permite realizar búsquedas sobre distintas colecciones de documentos, y procesar cada uno de los documentos para detectar los fármacos y las interacciones farmacológicas presentes en ellos. Distintos sistemas como

ISSN 1135-5948

PubMed[1] o PIE[2] permiten realizar búsquedas sobre la base documental MedLine[3] . Nuestra principal aportación es que nuestro prototipo permite realizar búsquedas sobre distintas colecciones multilingües como MedLine, Harrison[4], tuOtroMedico[5], OcuSalud[6], etc. Además, los documentos que utilizamos han sido etiquetados con los conceptos del metatesauro UMLS (Bodenreider, 2004) que aparecen en el texto, basándonos en los recursos de diccionarios biomédicos MeSH[7] y SNOMED[8] para conceptos en inglés y en español respectivamente.

En el ámbito de la extracción de entidades y relaciones en el dominio biomédico, la mayor parte de la investigación se ha centrado en el dominio biológico. Los principales avances se han desarrollado bajo el marco del foro de evaluación BioCreative[9]. En este dominio cabe destacar la herramienta Reflect (Pafilis et al., 2006) que permite reconocer entidades como genes y proteínas, y mostrar información sobre sus interacciones en textos de páginas web. Otro ejemplo es el sistema iHOP (information Hyperlinked Over Proteins) (Hoffmann y Valencia, 2005) que tiene una colección de textos procesados para recuperar de manera sencilla información de interacciones entre proteínas.

Para fomentar el desarrollo de sistemas de extracción de información en el dominio farmacológico, el año pasado organizamos la tarea DDIExtraction (Segura-Bedmar, Martínez y Sánchez-Cisneros, 2011). La mayoría de los trabajos fueron basados en técnicas de aprendizaje automático supervisado, y en particular, los métodos kernels obtuvieron los mejores resultados F1 65%) (Thomas et al., 2011) (Mahbub et al., 2011) (Mahbub y Lavelli, 2011). Desgraciadamente ninguno de los equipos participantes desarrollo un prototipo para que los usuarios del dominio pudieran interpretar la información detectada por sus sistemas.

Por ello, nuestra principal aportación es el desarrollo de una herramienta online[10] que permita a médicos y farmacéuticos acceder de una forma más eficaz a la información relativa a fármacos y sus interacciones. Para cada uno de los fármacos detectados, la herramienta permite mostrar información procedente de distintas fuentes, como DrugBank[11] o Pubchem[12], de una forma integrada. La herramienta también muestra la lista de interacciones farmacológicas

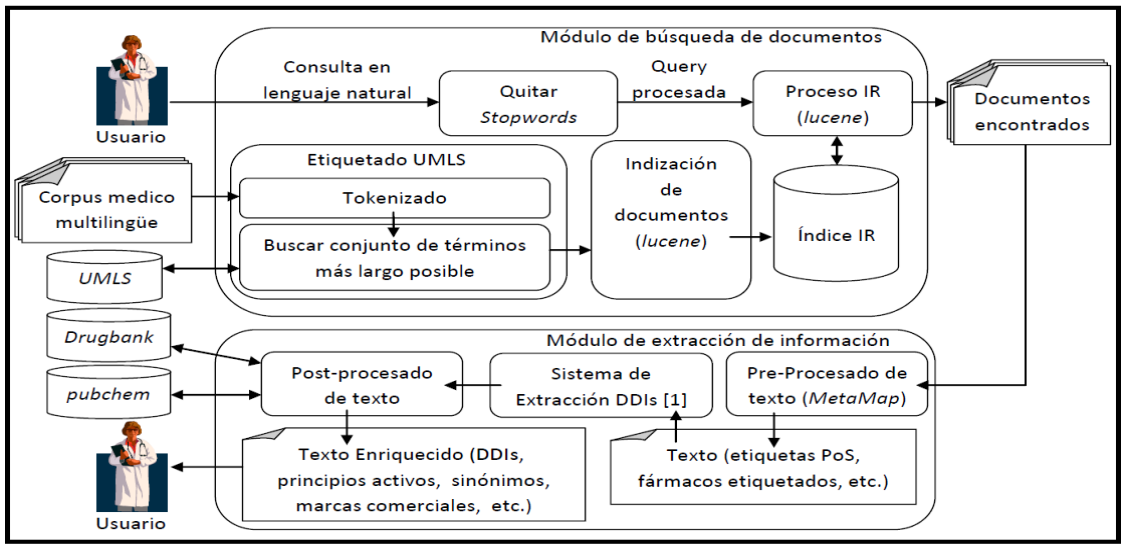


Figura 1: Arquitectura del sistema.

presentes en el texto. Estas utilidades proporcionan un mejor y más ágil acceso a la información, necesaria en la toma de decisiones respecto a la administración de un determinado fármaco.

[1] http://www.ncbi.nlm.nih.gov/pubmed/
[2] http://www.ncbi.nlm.nih.gov/CBBresearch/Wilbur/IRET/PIE/
[3] http://www.nlm.nih.gov/pubs/
[4] www.harrisonmedicina.com
[5] www.tuotromedico.com
[6] http://www.ocu.org/ocu-salud-s501.htm
[7] http://www.nlm.nih.gov/pubs/
[8] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1479961/
[9] http://www.biocreative.org/
[10] 163.117.129.57:8080/newddiextractorweb/
[11] http://www.drugbank.ca/
[12] http://pubchem.ncbi.nlm.nih.gov/

2 Arquitectura del sistema

El sistema está compuesto por la arquitectura general representada en la figura 1, donde se puede diferenciar un módulo de búsqueda y un módulo para la extracción de información. El módulo de búsqueda permitirá al usuario realizar búsquedas de textos biomédicos en un repositorio de colecciones de documentos multilingües. Por otro lado, el segundo módulo permite procesar textos para detectar fármacos y extraer sus interacciones farmacológicas, así como obtener información extendida de cada fármaco identificado en el texto.

2.1 Módulo de Búsqueda.

En primer lugar, el sistema permite buscar documentos en un repositorio multilingüe. Este repositorio ha sido creado por el grupo LLI de la universidad Autónoma de Madrid[13] y contiene colecciones de documentos en los siguientes idiomas:

- Inglés: más de 19.000.000 documentos en inglés de la colección biomédica Medline 2010.
- Español: 4.204 documentos en español de las revistas biomédicas Harrison, OcuSalud y TuOtroMedico.
- Japonés: 4.746 documentos en japonés de un revistas médicas de especialidades variadas.
- Árabe: 43.526 documentos en árabe del portal médico Altibbi[14].

Esto hace un total de 19.051.476 documentos de carácter biomédico, sin embargo la mayoría (19 millones) están en inglés. Para realizar el módulo de búsqueda se ha usado la herramienta IR Apache Lucene[15].

Las colecciones en inglés y español han sido procesadas por el grupo GSI de la Universidad Politécnica de Madrid[16] para etiquetar todos los términos MeSH y SNOMED identificados en los textos. MeSH es un tesauro de términos en inglés que es utilizado para indexar los artículos de MedLine. SNOMED es una terminología clínica que permite representar la información clínica de forma multilingüe.

En este proceso de etiquetado se han tenido que realizar tareas de desambiguación, ya que un término puede tener varias acepciones. Para tratar este problema se ha seguido la siguiente estrategia de prioridades:

1. Buscar los conceptos de MeSH y SNOMED que contengan todas las palabras del término.
2. Buscar los conceptos de MeSH y SNOMED que contengan el conjunto mayor de palabras del término.
3. Buscar los conceptos de MeSH y SNOMED que contengan alguna de las palabras del término.

Con este conjunto de corpus etiquetado con etiquetas MeSH y SNOMED hemos creado un índice sobre el que realizaremos los procesos de búsqueda en nuestra herramienta. Esto permite al usuario realizar búsquedas más avanzadas por conceptos.

Como resultado de estas búsquedas se devuelve un listado de documentos facilitando su identificador de Medline (PMID), titulo, revista, autores, snippet, etc. Finalmente el usuario puede ver el texto de cada documento, así como procesar el texto en busca de relaciones semánticas (interacciones entre fármacos).

2.2 Módulo de Extracción de Información.

El módulo de extracción permite procesar textos, ya sean resultados del módulo de búsqueda o textos introducidos directamente por el usuario. Para ello, en este módulo se ha realizado un trabajo de integración de varios recursos:

- En primer lugar, los textos son analizados semánticamente por MetaMap[17] para identificar los fármacos.
- A continuación, los textos son procesados por el sistema DrugDDI (Segura-Bedmar, 2010) que permite la extracción de interacciones farmacológicas. Dicho sistema está basado en el Shallow Linguistic Kernel (Giuliano, Lavelli y Romano, 2006).
- Finalmente, el sistema busca información para cada fármaco detectado: el nombre de su principio activo, código ATC, nombres comerciales, descripción del fármaco, etc. Para obtener toda esta información, utilizamos bases de datos

[13] http://www.lllf.uam.es/ESP/
[14] http://www.altibbi.com/
[15] http://lucene.apache.org/
[16] http://www.gsi.dit.upm.es/
[17] http://metamap.nlm.nih.gov/

farmacológicas como Drugbank y Pubchem.

Como resultado de este proceso, el sistema devuelve por un lado el texto con los fármacos identificados e información adicional de cada fármaco, y por otro lado un listado de las interacciones identificadas en el texto (ver figura 2).

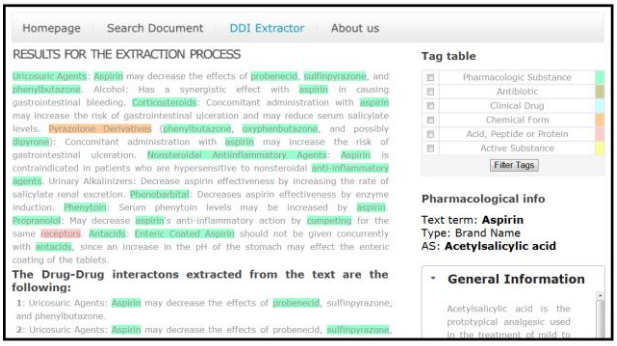


Figura 2: Ejemplo de procesado de texto.

3 *Trabajo en curso.*

El sistema permite la búsqueda de documentos sobre determinados conceptos clínicos (enfermedades, fármacos, síntomas, etc.), y además, el sistema es capaz de procesar los documentos para extraer de forma dinámica y online las interacciones entre fármacos que se describen en el texto. Para ello, procesa información estructurada y no estructurada.

Para ello, se ha incorporado un repositorio multilingüe de colecciones de documentos en diferentes idiomas, al que se ha realizado un etiquetado para identificar los términos MeSH y SNOMED en inglés y español, utilizando sus diccionarios de conceptos médicos.

Como trabajo futuro nos planteamos investigar en la detección de la gravedad de la interacción, su grado de certeza, y otros factores que pueden influir en la interacción como la dosis, tiempo de ingesta entre medicamentos, características individuales del paciente, etc. Toda esta información es vital a la hora de determinar la importancia clínica de una determinada interacción y que el facultativo sea capaz de tomar la decisión correcta respecto a la administración de un determinado fármaco.

4 *Agradecimientos.*

Este trabajo ha sido desarrollado en el marco del proyecto MA2VICMR (S2009/TIC-1542) y MULTIMEDICA[18] (TIN2010-20644-C03-01).

[18] http://labda.inf.uc3m.es/multimedica/

Bibliografía

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res., 267-270.

Giuliano, C., Lavelli, A., Romano, L. (2006). Exploiting shallow linguistic information for relation extraction from biomedical literature. EACL-2006, (págs. 401 - 408).

Hoffmann, R., Valencia, A. (2005). Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics.

Mahbub, M., Ben, A., Lavelli, A., y Zweigenbaum, P. (2011). Two Different Machine Learning Techniques for Drug-Drug Interaction Extraction . DDIExtraction2011: First Challenge Task on Drug-Drug Interaction Extraction 2011, 19 - 26.

Mahbub, M.F., y Lavelli, A. (2011). Drug-drug Interaction Extraction Using Composite Kernels. DDIExtraction2011: First Challenge Task on Drug-Drug Interaction Extraction 2011, 27 - 33.

Pafilis, E., O'Donoghue, S.I., Jensen, L.J., Horn, H., Kuhn, M., Brown, N.P., y Schneider, R. (2006). Reflect: augmented browsing for the life scientist. Nature Biotechnology, 508-510.

Segura-Bedmar, I. (2010). Application of information extraction techniques to pharmacological domain: extracting drug-drug interactions. Universidad Carlos III de Madrid, Departamento de Informática.

Segura-Bedmar, I., Martínez, P. Sánchez-Cisneros, D., (2011). The 1st DDIExtraction-2011 Challenge Task: Extraction of Drug-Drug Interactions from Biomedical Texts. SEPLN 2011, (págs. 1 - 9). Huelva.

Thomas, P., Neves, M., Solt, I., Tikk, D., y Leser, U. (2011). Relation Extraction for Drug-Drug Interactions using Ensemble Learning . DDIExtraction 2011: First Challenge Task on Drug-Drug Interaction Extraction 2011., 11 - 18.

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 213-216 recibido 08-05-12 revisado 31-05-12 aceptado 05-06-12

Servicios de anotación y búsqueda para corpus multimedia

Annotation and Search Services for Multimedia Corpus

David Hernández-Aranda
NLP&IR Research Group
ETSI Informática, UNED,
Madrid, Spain
daherar@lsi.uned.es

Rubén Granados
NLP&IR Research Group
ETSI Informática, UNED,
Madrid, Spain
rgranados@lsi.uned.es

Ana García Serrano
NLP&IR Research Group
ETSI Informática, UNED,
Madrid, Spain
agarcia@lsi.uned.es

Resumen: En este artículo corto se muestra la funcionalidad tanto del servicio anotador de textos desarrollado en el marco del proyecto Buscamedia[1], como del buscador sobre recursos o documentos multimedia anotados.
Palabras clave: Recuperación de información multimedia, Anotación multimedia, Recuperación de información textual, Fusión multimedia, Corpus.

Abstract: This paper shows the textual annotator service developed in the project Buscamedia as well as the search performed on multimedia resources or documents annotated.
Keywords: Multimedia information retrieval, Multimedia Annotation, Text-based Information Retrieval, Multimedia Fusion, Corpus.

1 Introducción

La recuperación de información multimedia (texto, imágenes, audio, vídeo) se aborda con enfoques textuales en la mayoría de las herramientas y sistemas existentes, usando anotaciones y metadatos asociados a las imágenes (Depeursinge and Müller, 2010), al audio o a los vídeos (Hernández-Aranda et al., 2010), o una parte de ellos, como son los segmentos, las instantáneas o *keyframes*, etc (Geurts et al 2005). Por ello, la anotación automática de recursos multimedia, sin intervención humana, está en continua investigación (Feng and Lapata, 2010), (Lombardo and Damiano 2012).

Sin embargo en el proyecto español Buscamedia se afronta el problema con una aproximación netamente multimedia, para lo que se han desarrollado subsistemas que "entienden" y procesan los recursos multimedia como se presenten (identificando por ejemplo los personajes que intervienen, objetos físicos, etc.). Cuando el resultado del análisis de estos subsistemas son anotaciones en forma de texto, se integran en el subsistema textual.

En las secciones siguientes, se presenta brevemente el sistema desarrollado y cómo buscar en el corpus *Deportes20*. A continuación se describe el corpus desarrollado y sus anotaciones provenientes del análisis de los recursos multimedia, como son las transcripciones, los subtítulos, algunos objetos físicos en imágenes, el texto sobreimpreso, los logos y las moscas. Se sigue con una breve presentación del servicio anotador textual y finalmente se muestran algunos ejemplos de búsqueda orientados a la validación del sistema a través de una prueba de concepto.

2 Descripción del sistema de búsqueda

El prototipo desarrollado consta de una interfaz web que permite la búsqueda y la visualización de resultados a partir de una consulta dada, siguiendo las pautas de un buscador común, pero que además permite mostrar todas las funcionalidades desarrolladas ya que los "botones" del interfaz las representan.

La visualización de los resultados se realiza a partir de los *snippets* creados manualmente sobre archivos multimedia de los vídeos o textos (noticias o páginas web) del corpus

[1] http://www.cenitbuscamedia.es/

Deportes20, como pueden ser imágenes o *keyframes* y segmentos de visión con información concreta, o textos multilingües porque los documentos multimedia están en castellano, catalán, euskera o inglés.

Figura 1: Interfaz del sistema de búsqueda

En la figura 1 se muestra el interfaz correspondiente al sistema de búsqueda textual en el que se han realizado los siguientes pasos:

Preproceso. Con la información textual extraída de los documentos multimedia, se realiza un análisis de detección de las entidades nombradas. Dependiendo del idioma, se aplica una herramienta distinta. Para el castellano e inglés, *Stilus* (licencia para investigación proporcionada por Daedalus[2]), y para el catalán, se aplica *Freeling*[3].

Para el caso de los vídeos en catalán, se hace uso además de un recurso externo, el denominado *Thesaure* (con licencia de uso restringida de la corporación catalana de televisión), con el objetivo de enriquecer la anotación en la que aparezcan términos de dicho tesauro.

A continuación se crea un documento único para cada documento multimedia unificando la información contenida originalmente, y además se añaden el idioma, el nombre del documento original, y las entidades nombradas detectadas, o los campos relacionados con la información semántica del tesauro.

Un ejemplo de documento XML único es:

```
<out>
   <idioma>es</idioma>
   <subtitulos>Y el líder en Liga sigue
      siendo el Real Madrid.. Los
      blancos debutan esta semana en
      Copa...</subtitulos>
   <textoSobreimpreso>KOREAN PETRONAS
      Formula1. ronaldo 7. P.LEÓN. The
      next big Audi. ¿Te llevo?. LA
      NOCHE DE CR7. Audi. RONALDO...
      </textoSobreimpreso>
   <logos>PETRONAS Bwin Audi Mahou
      Adidas mahou audi Audi AUDI
      bwinCamiseta Real Madrid C.F. LFP
      bwin RNE Punto pelota Onda cero
      Barça TV TV...</logos>
   <nes> Real_Madrid_Club_de_Fútbol
      MADRID Madrid REAL Real_Madrid
      Comunidad_Autónoma_de_Murcia
      Murcia José_Mourinho Mourinho
      Mou...</nes>
   </out>
```

A partir de estos documentos únicos el pre-procesamiento sigue con los analizadores *SnowBall* implementados para cada idioma en *Lucene*, para efectuar el *stemming*, y con la eliminación de stopwords.

Indexación. El modelo de indexación consiste en la creación de un único índice haciendo uso de *Lucene*, indexando en diferentes campos toda la información de los cuatro idiomas.

Búsqueda. Una vez indexado el corpus *Deportes20*, en la búsqueda de cada consulta se obtendrá una única lista de resultados ordenados por relevancia. La función de *ranking* utilizada es BM25F que extiende a BM25 para documentos estructurados (formados por campos).

En este prototipo se permite la selección del operador con el que se desea hacer la búsqueda (OR o AND). Además se podrán seleccionar los tipos de metadatos, correspondientes a los distintos tipos de anotaciones, del documento único. Y se podrán filtrar los resultados recuperados por el tipo de documento (solo vídeos, solo noticias/páginas web o ambos).

El servicio de búsqueda está disponible para otros investigadores, y previa solicitud de *login* y *password* pueden acceder al prototipo en la dirección siguiente: http://albali.lsi.uned.es/deportes20-1.0.0/.

3 *El corpus Deportes20*

La colección está compuesta por 4 tipos de recursos o documentos multimedia:

[2] www.daedalus.es
[3] http://nlp.lsi.upc.edu/freeling/

Vídeos en catalán (proporcionados por CCMA[4], miembro del consorcio): 21 documentos multimedia en catalán, de los cuales, 10 tienen asociado un documento XML con su descripción, una carpeta con *keyframes* asociados, y los objetos detectados que aparecen en ellos.

De los 11 recursos restantes sí que se dispone de sus vídeos correspondientes, así como de sus transcripciones y *keyframes* asociados. Sin embargo, en este caso, no se dispone de los objetos que aparecen en ellos.

Vídeos en castellano (proporcionados por ISID[5], miembro del consorcio). 10 vídeos en castellano, de los cuales 6 tienen asociado un vídeo, la transcripción del audio y los subtítulos de dicho vídeo. Los 4 vídeos restantes, además de la información anterior, contienen el texto sobreimpreso y los logos y moscas detectados en cada vídeo.

Páginas Web (proporcionadas por Daedalus, miembro del consorcio). Son 34 páginas web en formato HTML, cuya temática está relacionada con los vídeos de los grupos anteriores, de las cuales 30 están en idioma castellano, 3 en catalán y 1 en inglés (selección manual).

Noticias (proporcionadas por *Daedalus*). Conjunto de 62 noticias en formato HTML, de las cuales 30 están en castellano, 30 en catalán y 2 en euskera. Se extrajeron con consultas relacionadas con los documentos de un corpus de 21.632 noticias de 16 periódicos con formatos diferentes.

Con todo ello, se construye una colección anotada de 127 recursos o documentos multimedia correspondientes a vídeos, páginas web y noticias textuales. Este corpus está disponible para la comunidad de investigadores, previa solicitud.

4 Anotacion automática

La herramienta de anotación textual desarrollada en el proyecto permite analizar textos en diferentes idiomas (español, inglés, catalán) y realizar su análisis morfosintáctico, de forma que se obtengan los términos que pertenecen a una categoría morfosintáctica específica, y las entidades nombradas.

Para ello, esta herramienta utiliza módulos intermedios que sirven de *wrapper* para herramientas conocidas, como son *FreeLing*, *TreeTagger*[6], *Stanford NER*[7] y *Stilus* de *Daedalus*, y que serán seleccionados en la llamada al servicio. El servicio web de esta herramienta de anotación se encuentra en: http://albali.lsi.uned.es/DemoAnotadorWS/ y puede utilizarse para investigación, previa petición de *login* y *password*.

En la figura 2 puede observarse una salida del interfaz de este servicio.

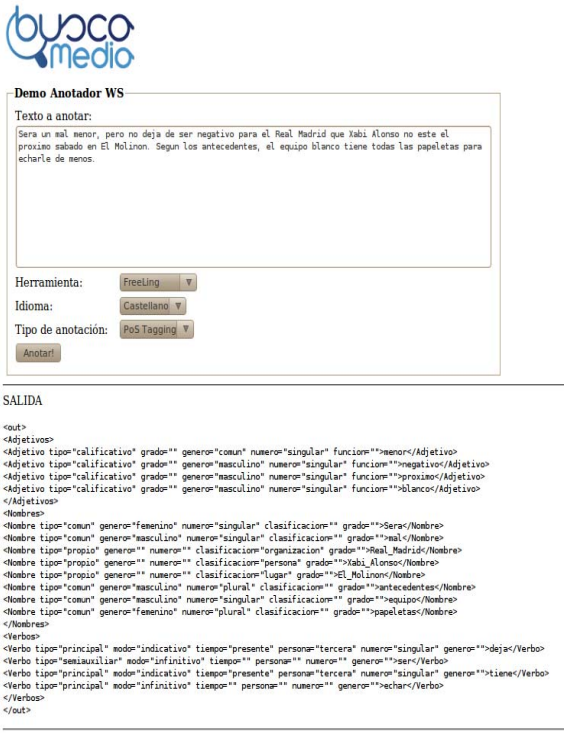


Figura 2: Interfaz del anotador

5 *Pruebas con la colección Deportes20*

El corpus *Deportes20* se complementa con un conjunto de consultas (relacionadas con el objetivo a evaluar) y sus juicios de relevancia (la mayoría con explicaciones detalladas).

A continuación se incluyen algunos ejemplos de prueba, indicando en cada uno de ellos las ventajas alcanzadas con la combinación de anotaciones provenientes de diferentes medias.

Consulta 1: "uso de un zeppelín en eventos deportivos".

Opciones de búsqueda: castellano, OR, Vídeos, objetos

Resultados: En este ejemplo el usuario busca vídeos sobre la aparición de un zeppelín

[4] http://www.ccma.cat/pccrtv/ccrtvSeccio.jsp

[5] www.isid.es

[6] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[7] http://nlp.stanford.edu/software/CRF-NER.shtml

en algún evento deportivo, donde estos suelen usarse con fines publicitarios. Para recuperar el único vídeo del corpus relevante para esta consulta será necesario hacer uso del campo "Objetos", ya que el vídeo buscado está anotado por el servicio de anotación de objetos con la aparición de un zeppelín.

Con este ejemplo queda clara la utilidad de la integración de la salida proveniente del servicio de detección de objetos en vídeos, en forma de información textual. Anotando textualmente el objeto multimedia con la información sobre la identificación del objeto "zeppelín", se consigue posteriormente recuperar el vídeo buscado (correspondiente a un partido de baloncesto en el que aparece un zeppelín).

Figura 3: Interfaz del buscador para la respuesta a la consulta "Cristiano Ronaldo"

Consulta 2: "Información sobre el Casademont".

Opciones de búsqueda: castellano, OR, Vídeos, Thesaure>Detalles

Resultados: La consulta busca vídeos relacionados con un equipo de baloncesto, el Casademont. Si no se hace uso de la información semántica que proporciona el Thesauro (en Detalles), nunca se sabría que se refiere al equipo que actualmente se llama Akasvayu Girona (antiguamente Casademont Girona). Por lo tanto, gracias a la selección de la opción "Detalles" del Thesaure, la búsqueda recupera el vídeo del corpus que trata sobre un partido entre el Barcelona y el Girona.

6 Comentarios finales

Se ha presentado el servicio de anotación de recursos multimedia y unos ejemplos de búsqueda que muestran los beneficios de la anotación.

A continuación se pretende abordar la integración semántica de las anotaciones a través de la información contenida en una ontología multimedia disponible en el proyecto Buscamedia.

7 Agradecimientos

Este trabajo se ha financiado con el proyecto competitivo BUSCAMEDIA (CEN-20091026), financiado por el Ministerio de Industria.

Agradecemos muy especialmente la colaboración de los investigadores de todos los miembros del consorcio, pero muy en particular en esta tarea de creación del corpus a los de *Tecnalia*, *UC3M*, *ISID*, *Bilbomática*, y por supuesto a *Daedalus* y *ATOS*.

Bibliografía

A. Depeursinge and H. Müller, (2010). *Fusion Techniques for Combining Textual and Visual Information Retrieval*. In H. Müller, P. Clough, T. Deselaers, & B. Caputo, Experimental Evaluation in Visual Information Retrieval (Vol. 32). Springer.

A. García-Serrano, R. Granados, D. Hernández-Aranda, V. Fresno y J. Cigarrán, *Anotación para la recuperación de información multimedia: el corpus Deportes20*, Actas del congreso SEPLN 2012, Valencia, 2012.

J. Geurts, J. van Ossenbruggen, L. Hardman, *Requirements for practical multimedia annotation*, Workshop on Multimedia and the Semantic Web, 4-11 2005.

D. Hernández-Aranda, R. Granados, J. Cigarran, A. Rodrigo, V. Fresno, and A. García-Serrano. *UNED at mediaeval 2010: exploiting text metadata for automatic video tagging*. In MediaEval 2010 Workshop. Pisa, Italy, 24 October, 2010.

V. Lombardo and R. Damiano, *Semantic annotation of narrative media objects*, Multimedia Tools and Applications, Volume 59, Number 2, Pages 407-439, 2012.

Y. Feng and M. Lapata. 2010. *Topic Models for Image Annotation and Text Illustration*. In Proc. of the Human Language Technologies at the North American Chapter of the Association for Computational Linguistics, 831-839. Los Angeles, California.

Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 217-220 recibido 07-05-12 revisado 01-06-12 aceptado 05-06-12

Sistema SAGAS: herramienta de soporte al subtitulado para personas sordas

Julio Villena[1], Lourdes Moreno[2], Paloma Martínez[2], José Carlos González[1]
[1]Daedalus – Data, Decisions and Language, S.A.
[2] Grupo LaBDA, Dpto. de Informática, Universidad Carlos III de Madrid
{jvillena, jmartinez, jgonzalez}@daedalus.es, {lmoreno, pmf}@inf.uc3m.es

Resumen: Siguiendo legislación en España, en televisión se deben alcanzar unas cuotas en el servicio de subtitulado para personas sordas, además, los subtítulos deben elaborarse siguiendo normativa. Este marco regulador conlleva una demanda de tecnología que facilite a los radiodifusores y productores de contenido la generación de subtitulado, como es la generación automática de subtitulado a partir de reconocimiento de audio. En este trabajo se presenta "SAGAS, Sistema Avanzado de Generación Automática de Subtítulos", que proporciona subtítulos adecuados a la norma española para contenido vídeo que vaya acompañado de un guión o transcripción.

Palabras clave: subtitulado, subtitulado para personas sordas, reconocimiento automático del habla, ASR, alineamiento

Abstract: Following legislation in Spain specific TV quotas must achieved in Subtitling Service for deaf people and additionally subtitles should be developed according to regulations. This regulatory framework implies a technology demand to facilitate broadcasters and content producers to generate subtitling, like automatic subtitling from Automatic Speech Recognition (ASR). This paper introduces "SAGAS prototype (Advanced System for automatic generation of subtitles)" which provides subtitles according to Spanish standard from an audio and a transcript.

Keywords: subtitling, subtitling for deaf people, Automatic Speech Recognition, alignment

1 Introducción

Las personas con sordera necesitan de subtitulado para poder acceder a los contenidos audiovisuales que se ofrecen en televisión (TV), entre otros medios. La TV es el medio de comunicación más influyente en la sociedad y es esencial asegurar el acceso a todos los ciudadanos. Dicho derecho está legislado en España y hay cuotas de servicio de subtitulado impuestas por ley a los radiodifusores y productores de contenidos de TV (BOE, 2010). Además, este subtitulado debe seguir buenas prácticas en su elaboración cumpliendo con la normativa en España (AENOR, 2003).

Para que haya un cumplimiento de este marco regulador, se debe disponer de tecnología de generación automática de subtitulado, ya que los procesos manuales son inviables por los excesivos costes. Por otro lado, las tecnologías involucradas, como las de reconocimiento del habla, tienen problemas tecnológicos aun sin resolver, sobre todo en el caso de reconocimiento del audio en tiempo real con independencia del locutor. La falta de tecnología que dé soporte a los subtituladores, junto con la gran demanda de contenido subtitulado en TV en España, ha motivado este trabajo.

El trabajo que aquí se presenta, sistema SAGAS, proporciona una herramienta que genera de manera automática subtítulos, adecuados a la normativa española, a partir de un audio y una transcripción con información del audio.

Esta investigación se enmarca en el proyecto Sistema "Avanzado de Generación Automática

ISSN 1135-5948

de Subtítulos, SAGAS" (TSI-020100-2010-184)[1] cofinanciado por el Ministerio de Industria, Energía y Turismo, dentro del Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica 2008-2011.

Este proyecto supone una oportunidad para los agentes involucrados en la comercialización de contenidos digitales, con el objetivo de mejorar sus procesos productivos, simplificando y optimizando los procesos de subtitulado. El consorcio responsable del proyecto es un equipo multidisciplinar de expertos compuesto por: DAEDALUS[2], empresa referente en el sector de la producción audiovisual que coordina el proyecto y se encarga del desarrollo técnico, grupo de investigación LaBDA[3] de la Universidad Carlos III de Madrid encargado de la validación que permite asegurar los mejores resultados en cuanto a la calidad del proceso y la adecuación a normas de subtitulado y estándares de accesibilidad y por último, RTVE con el valioso rol de usuario. RTVE aporta al proyecto su experiencia y archivo de contenidos multimedia, que mantiene como fruto de todos los años de actividad en el sector, su colaboración y visión resulta imprescindible a la hora abordar el proyecto.

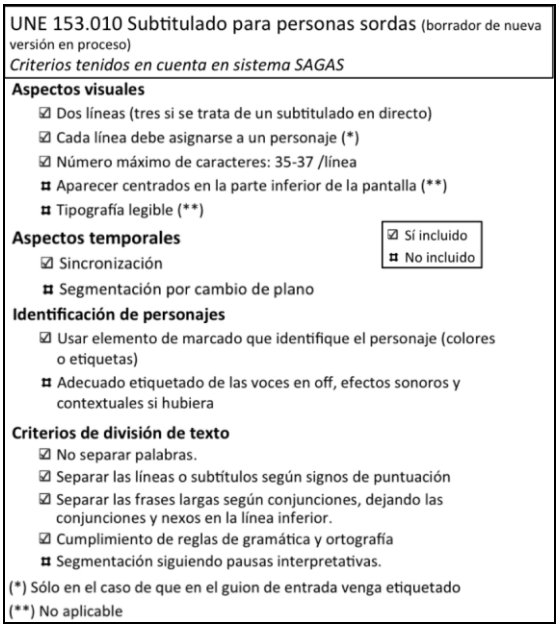
2 Trabajos relativos

Se encuentran sistemas de subtitulado asistidos por motores de reconocimiento automático de audio, algunos de estos parcialmente asistidos por operadores humanos (Boulianne et al., 2006), y otros que obtienen una transcripción del audio sin ningún tipo de asistencia humana (Neto et al, 2008). Estos últimos conllevan problemas como una alta tasa de error en el Módulo de Reconocimiento automático del Habla y retardo en la generación de los subtítulos (Meinedo, Viveiros y Neto, 2008).

En el panorama de TV en España, se encuentran varios sistemas automáticos de generación de subtitulado (Álvarez, del Pozo y Arruti, 2010) (Bordel et al, 2011). Cada uno de ellos utiliza distintos recursos de Tecnologías del Lenguaje para obtener mejores resultados en el Módulo de Reconocimiento. Al igual que SAGAS, se encuentran trabajos relativos que parten de transcripciones de textos más o menos equivalentes a la trascripción del audio, donde se hace un reconocimiento automático del habla en tiempo real para alinear temporalmente el texto y la voz (García et al, 2009).

3 Aplicación de Norma UNE 153.010 de subtitulado

La generación automática de los subtítulos en España no se puede llevar a cabo con otras herramientas internacionales ya que, siguiendo la norma UNE 153.010, hay requisitos propios que las distinguen. La adecuación a la normativa es un valor añadido de este trabajo. Se ha tenido en cuenta la versión del 2003, así como la nueva versión en borrador, que será el nuevo estándar a seguir próximamente. En esta nueva versión se han tenido en cuenta otros escenarios de aplicación y se han incorporado requisitos nuevos. Así se distingue entre aspectos visuales, temporales, de identificación de locutores y criterios de división de texto a cumplir.


UNE 153.010 Subtitulado para personas sordas (borrador de nueva versión en proceso)

Criterios tenidos en cuenta en sistema SAGAS

Aspectos visuales
- ☑ Dos líneas (tres si se trata de un subtitulado en directo)
- ☑ Cada línea debe asignarse a un personaje (*)
- ☑ Número máximo de caracteres: 35-37 /línea
- ⌗ Aparecer centrados en la parte inferior de la pantalla (**)
- ⌗ Tipografía legible (**)

Aspectos temporales

☑ Sí incluido
⌗ No incluido

- ☑ Sincronización
- ⌗ Segmentación por cambio de plano

Identificación de personajes
- ☑ Usar elemento de marcado que identifique el personaje (colores o etiquetas)
- ⌗ Adecuado etiquetado de las voces en off, efectos sonoros y contextuales si hubiera

Criterios de división de texto
- ☑ No separar palabras.
- ☑ Separar las líneas o subtítulos según signos de puntuación
- ☑ Separar las frases largas según conjunciones, dejando las conjunciones y nexos en la línea inferior.
- ☑ Cumplimiento de reglas de gramática y ortografía
- ⌗ Segmentación siguiendo pausas interpretativas.

(*) Sólo en el caso de que en el guion de entrada venga etiquetado
(**) No aplicable

Figura 1. Criterios UNE 153.010 aplicados en Sistema SAGAS

Tal como indica la figura 1, el sistema SAGAS se ha implementado para que los subtítulos generados cumplan con la mayoría de los requisitos de esta nueva versión de la norma.

[1] http://labda.inf.uc3m.es/sagas/esp/
[2] http://www.daedalus.es/
[3] http://labda.inf.uc3m.es/

4 Sistema SAGAS

La contribución consiste en el desarrollo de un prototipo que da soporte al subtitulado automático de contenido audiovisual en diferido para diversos medios: TV, Internet y dispositivos móviles. El contexto es el del subtitulado en diferido, *off-line* (fuera de línea) o enlatado, en cuyo caso no se trata de un proceso en tiempo real sino que se realiza previamente, sin necesidades temporales.

4.1 Arquitectura del sistema

El prototipo de herramienta de subtitulado que aquí se presenta puede ser utilizada bien como una herramienta aislada (*standalone)* o bien como un módulo en una herramienta comercial actualmente utilizada en la producción de subtítulos (como FAB Subtitler, por ejemplo).

A través de la interfaz de usuario, la herramienta toma como entrada el guión del material audiovisual en cuestión y el vídeo de dicho material. La salida está formada por el vídeo de entrada en el que se habrán integrado los subtítulos sincronizados, construidos a partir del guión, en los instantes de tiempo correspondientes, almacenados en formato estándar como EBU o formatos XML.

La arquitectura tiene un diseño modular, de forma que el sistema comprende de manera desacoplada el Módulo de Reconocimiento de Habla y el Módulo de Alineación, además de tener una interfaz para la entrada de datos y otra para la salida de subtítulos. Este diseño permite que cada módulo pueda ser utilizado por separado por aplicaciones externas, simplificando además el mantenimiento de los mismos.

Para el **Módulo de Reconocimiento del habla**, se han generado nuevos modelos en los motores de reconocimiento o ASR que no lo incorporan de manera nativa para obtener mejores resultados en el proceso de reconocimiento. En el prototipo actual, se han utilizado como motores la API del motor de reconocimiento de *MMIndexer* y *Dragon Naturally Speaking*. Este módulo da cómo salida un archivo de índices.

El **Módulo de Alineación** es mostrado en la Figura 2. En este módulo se da un proceso automático de alineación del guión o trascripción con el audio mediante la adición de marcas de tiempo, así como el tratamiento de revisión de errores y la segmentación del texto en subtítulos conformes a la norma en España y criterios de calidad vistos (ver apartado 3).

Tal como muestra la figura 2, como entrada se tiene la transcripción y el archivo de índices obtenido tras el reconocimiento de audio (Resultado Módulo de Reconocimiento).

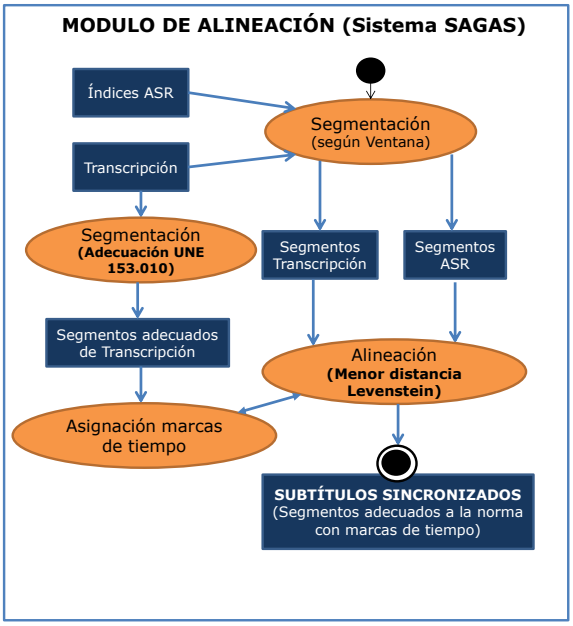


Figura 2: Diagrama de actividades de Módulo de alineación de Sistema SAGAS

De ambos recursos se hace una agrupación de las palabras o segmentación hasta llegar a un determinado tamaño de ventana de comparación. A continuación, se hace una comparación de las ventanas obtenidas por palabras de la transcripción con el de palabras reconocidas del ASR. Estos segmentos se alinean cuando la comparación entre ventanas da una distancia menor, utilizando la distancia de Levenshtein (Levenshtein, 1966).

A esta alineación se le incorporan marcas de tiempo para que los segmentos de subtítulos sigan la UNE 153.010. Para llevar a cabo este último proceso, la transcripción previamente ha sido segmentada siguiendo criterios de la norma como que el número de caracteres de cada segmento no supere el tamaño máximo fijado en 37, aunque se puede configurar con otros valores distintos a la norma.

Como resultado del módulo, se obtiene una estructura de datos que contiene en cada posición: los subtítulos a mostrar, el instante temporal asociado, y el locutor asignado.

Por último, tomando toda esta información, se genera la salida o subtítulos sincronizados en un **Interfaz de salida** tal como se muestra en la figura 3 que reproduce los resultados (reproducción de vídeo con los subtítulos

generados), y permite la descarga de los archivos de subtitulado en distintos formatos.

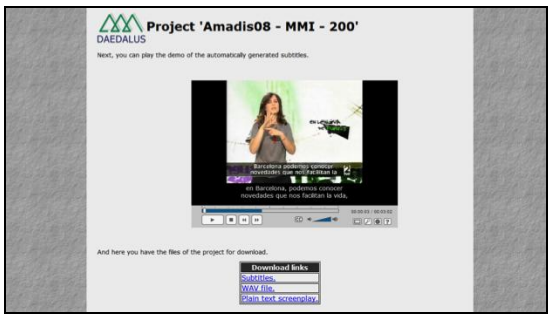


Figura 3: Interfaz de salida con demostrador de los subtítulos generados

5 Conclusiones y lineas futuras

El sistema SAGAS ofrece una herramienta de utilidad a los profesionales involucrados en el proceso de creación de subtítulos siguiendo normativa. Sobre el prototipo actual se ha hecho una evaluación preliminar, detectando algunos aspectos a mejorar en los que se trabaja en la actualidad para incrementar los niveles de calidad del subtitulado generado.

Como líneas futuras, además de mejorar el sistema actual, se quiere integrar una herramienta de autor para que el profesional o subtitulador pueda editar el subtitulado en tiempo de proceso de generación.

Bibliografía

AENOR, 2003. UNE 153.010 Subtitulado para personas sordas y personas con discapacidad auditiva. Subtitulado a través del teletexto.

Álvarez, A., del Pozo, A. and Arruti, A. 2010 APyCA: Towards the automatic subtitling of television content in Spanish. Computer Science and Information Technology (IMCSIT), 567- 574

BOE, 2010. Ley General Audiovisual. Ley 7/2010, de 31 de Marzo, General de la Comunicación Audiovisual se regula la comunicación audiovisual de cobertura estatal y establece las normas básicas en materia audiovisual sin perjuicio de las competencias reservadas a las Comunidades Autónomas y a los Entes Locales en sus respectivos ámbitos.

Bordel, G., Nieto, S., Penagarikano, M., Rodriguez-Fuentes, LJ., Varona, A. 2011. Automatic Subtitling of the Basque Parliament Plenary Sessions Videos. INTERSPEECH 2011. Florence, Italy.

Boulianne, G., Beaumont, F.F., Boisvert, M., Brousseau, J., Cardinal, P., Chapdelaine, C., Comeau, M., Ouellet, P., Osterrath, F. 2006. Computer-assisted closedcaptioning of live TV broadcasts in French, Interspeech 2006, Pittsburgh, USA.

Garcia, JE, Ortega, A, Lleida, E, Lozano, T, Bernues, E and Sanchez, D. 2009. Audio and Text Synchronization for TV news Subtitling based on Automatic Speech Recognition. Broadband Multimedia Systems and Broadcasting, 2009. BMSB '09.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10 (1966):707710.

Meinedo, H., Viveiros, M. and Neto, J. 2008. Evaluation of a live broadcast news subtitling system for Portuguese. Interspeech 2008, Brisbane, Australia, Sep. 2008.

Neto, J.; Meinedo, H.; Viveiros, M.; Cassaca, R.; Martins, C.; Caseiro, D. 2008. Broadcast news subtitling system in Portuguese. 2008. Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, vol., no., pp.1561-1564

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 *24 €* (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	Universitat de Barcelona
Arantza Díaz de Ilarraza	Euskal Herriko Unibertsitatea
Antonio Ferrández	Universitat d'Alacant
Mikel Forcada	Universitat d'Alacant
Ana García-Serrano	UNED
Koldo Gojenola	Euskal Herriko Unibertsitatea
Xavier Gómez Guinovart	Universidade de Vigo
Julio Gonzalo	UNED
José Miguel Goñi	Universidad Politécnica de Madrid

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://www.sepln.org/revistaSEPLN/revistas.php
Las funciones del Consejo de Redacción están disponibles en Internet a través de http://www.sepln.org/revistaSEPLN/edirevista.php
Las funciones del Consejo Asesor están disponibles Internet a través de la página http://www.sepln.org/revistaSEPLN/lectrevista.php

Proyectos

Demostraciones

Información General