Sociedad Española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT Fundación Española para la Ciencia y la Tecnología 2015

ISSN: 1135-5948

Artículos

Tesis

Sociedad Española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT 2015

ISSN: 1135-5948

Comité Editorial

Consejo de redacción

ISSN: 1135-5948
ISSN electrónico: 1989-7553
Depósito Legal: B:3941-91
Editado en: Universidad de Jaén
Año de edición: 2017
Editores: Mariona Taulé Delor Universidad de Barcelona mtaule@ub.edu
M. Teresa Martín Valdivia Universidad de Jaén maite@ujaen.es
Eugenio Martínez Cámara TU Darmstadt camara@ukp.informatik.tu-darmstadt.de
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

Ruslan Mitkov	University of Wolverhampton (Reino Unido)
Manuel Montes y Gómez	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)
Lluís Padró	Universidad Politécnica de Cataluña (España)
Manuel Palomar	Universidad de Alicante (España)
Ferrán Pla	Universidad Politécnica de Valencia (España)
German Rigau	Universidad del País Vasco (España)
Horacio Rodríguez	Universidad Politécnica de Cataluña (España)
Leonel Ruiz Miyares	Centro de Lingüística Aplicada de Santiago de Cuba (Cuba)
Kepa Sarasola	Universidad del País Vasco (España)
Emilio Sanchís	Universidad Politécnica de Valencia (España)
Encarna Segarra	Universidad Politécnica de Valencia (España)
Thamar Solorio	University of Houston (Estados Unidos de América)
Maite Taboada	Simon Fraser University (Canadá)
Mariona Taulé	Universidad de Barcelona (España)
Juan-Manuel Torres-Moreno	Laboratoire Informatique d'Avignoon / Université d'Avignon (Francia)
José Antonio Troyano Jiménez	Universidad de Sevilla (España)
L. Alfonso Ureña López	Universidad de Jaén (España)
Rafael Valencia García	Universidad de Murcia (España)
Felisa Verdejo Maíllo	Universidad Nacional de Educación a Distancia (España)
Manuel Vilares	Universidad de la Coruña (España)
Luis Villaseñor-Pineda	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)

Revisores adicionales

Iñaki Alegría	Universidad del País Vasco (España)
Rodrigo Agerri	Universidad del País Vasco (España)
Daniel Castro Castro	Centro de Lingüística Aplicada de Santiago de Cuba (Cuba)
Victor Manuel Darriba Bilbao	Universidad de Vigo (España)
Víctor Fresno	Universidad Nacional de Educación a Distancia (España)
Manuel Carlos Díaz Galiano	Universidad de Jaén (España)
Salud María Jiménez Zafra	Universidad de Jaén (España)
Eugenio Martínez Cámara	Technische Universität Darmstadt (Alemania)
Fernando Martínez Santiago	Universidad de Jaén (España)
M. Dolores Molina González	Universidad de Jaén (España)
Arturo Montejo Ráez	Universidad de Jaén (España)
Mariluz Morales Botello	Universidad Europea de Madrid (España)
Enrique Puertas Sanz	Universidad Europea de Madrid (España)
Fernando Sánchez-Vega	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)
Fernando Ribadas-Pena	Universidad de Vigo (España)

Sociedad Española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT Fundación Española para la Ciencia y la Tecnología 2015

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 58 de la revista *Procesamiento del Lenguaje Natural* contiene trabajos correspondientes a dos apartados diferentes: comunicaciones científicas y resúmenes de tesis

doctorales. Todos ellos han sido aceptados mediante el proceso de revisión tradicional en la revista. Queremos agradecer a los miembros del Comité asesor y a los revisores adicionales la labor que han realizado.

Se recibieron 37 trabajos para este número de los cuales 33 eran artículos científicos y 4 resúmenes de tesis doctorales. De entre los 33 artículos recibidos 17 han sido finalmente seleccionados para su publicación, lo cual fija una tasa de aceptación del 51,5%.

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 editorial. El criterio de corte adoptado ha sido la media de las tres calificaciones, siempre y cuando hayan sido iguales o superiores a 5 sobre 7.

Marzo de 2017
Los editores

Sociedad Española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT Fundación Española para la Ciencia y la Tecnología 2015

ISSN: 1135-5948

Preamble

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

The subject areas addressed are the following:

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

The 58th issue of the *Procesamiento del Lenguaje Natural* journal contains scientific papers and doctoral dissertation summaries. All of these were accepted by a peer reviewed process. We would like to thank the Advisory Committee members and additional reviewers for their work.

Thirty-seven papers were submitted for this issue of which thirty-three were scientific papers and four dissertation summaries. From these thirty-three papers, we selected seventeen (51.5%) for publication.

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

March 2017
Editorial board

Sociedad Española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT Fundación Española para la Ciencia y la Tecnología 2015

ISSN: 1135-5948

Artículos

Tesis

Información General

Artículos

Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 13-20 recibido 30-10-2016 revisado 13-01-2017 aceptado 12-02-2017

POL: un nuevo sistema para la detección y clasificación de nombres propios*

POL: a new system for named-entity detection and categorisation

Rogelio Nazar, Patricio Arriagada
Instituto de Literatura y Ciencias del Lenguaje
Pontificia Universidad Católica de Valparaíso
Avenida El Bosque 1290, Viña del Mar, Chile
rogelio.nazar@pucv.cl, patricio.arriagada.s@mail.pucv.cl

Resumen: El objetivo de este trabajo es desarrollar una metodología para la detección y clasificación de nombres propios (NP) en las categorías de antropónimo, topónimo y nombre de organización. La hipótesis sobre la que se basa la investigación es que el contexto de aparición de los NP –definido como las n palabras previas– así como los elementos que componen el NP mismo, pueden aportar pistas para predecir el tipo de entidad. Para tal fin, se diseñó un algoritmo de clasificación supervisado que se entrena con un corpus ya anotado por otro sistema, que en el caso de nuestros experimentos fue la suite de analizadores de idiomas FreeLing anotando el corpus de la Wikipedia en castellano. En el entrenamiento, nuestro sistema aprende a relacionar tipos de entidades con palabras del contexto así como las que componen los NP anotados. Se evalúan los resultados en el corpus CONLL-2002 y también con un corpus de geopolítica perteneciente a la revista *Le Monde Diplomatique* en su edición en castellano. Se compara además el desempeño en ese corpus de distintos sistemas de extracción y clasificación de NP en castellano.
Palabras clave: Entidades nombradas, nombres propios, lingüística textual

Abstract: The purpose of this research is to develop a methodology for the detection and categorisation of named entities or proper names (PPNN), in the categories of geographical place, person and organisation. The hypothesis is that the context of occurrence of the entity –a context window of n words before the target– as well as the components of the PN itself may provide good estimators of the type of PN. To that end, we developed a supervised categorisation algorithm, with a training phase in which the system receives a corpus already annotated by another NERC system. In the case of these experiments, such system was the open-source suite of language analysers FreeLing, annotating the corpus of the Spanish Wikipedia. During this training phase, the system learns to associate the category of entity with words of the context as well as those from the PN itself. We evaluate results with the CONLL-2002 and also with a corpus of geopolitics from the journal *Le Monde Diplomatique* in its Spanish edition, and compare the results with some well-known NERC systems for Spanish.
Keywords: Named entities, proper names, text linguistics

1 Introducción

El presente artículo sintetiza una investigación en curso cuyo objetivo es desarrollar una metodología de detección y clasificación de nombres propios (NP) o *named entities*, tal como se conocen generalmente en la comunidad del Procesamiento del Lenguaje Natural (PLN). La tarea se descompone en dos fases que se resuelven de manera independiente pero consecutiva: en primer lugar la detección del NP, que incluye el problema de su delimitación, y luego su clasificación en las categorías de antropónimos, topónimos y nombres de organizaciones. Identificamos como

* Investigación financiada por CONICYT (Gobierno de Chile), Proyecto Fondecyt 11140686: "Inducción automática de taxonomías de sustantivos generales y especializados a partir de corpus textuales desde el enfoque de la lingüística cuantitativa". Agradecemos también a los revisores por sus comentarios.

ISSN 1135-5948

POL a nuestro algoritmo por Persona, Organización, Lugar.

Basándonos en el trabajo previo de Arriagada (2016), la hipótesis en la que la investigación se basa es que el contexto de aparición de los NP –definido como las n palabras previas– además de las palabras que componen el NP mismo, pueden aportar elementos predictores del tipo de entidad. Por ejemplo, un verbo, un adverbio o un adjetivo, pueden ofrecer pistas respecto al tipo de entidad que acompañan cuando están en una relación predicado/argumento con un NP (De Miguel, 1999). Así, cabe esperar que el sujeto de un verbo como *considerar* sea un sujeto humano. Incluso sin análisis sintáctico, se hipotetiza que la mera coocurrencia de estos elementos puede aportar información útil.

El diseño experimental para poner a prueba esta hipótesis se basa en el desarrollo de un algoritmo supervisado, con una etapa de entrenamiento en que recibe texto anotado por otro sistema, y una etapa de prueba en la que detecta y clasifica entidades en texto sin anotar. Durante el entrenamiento, el algoritmo aprende a relacionar palabras que componen los NP y las que aparecen en el contexto de estos con la categoría de entidad asignada por el otro sistema. En el caso de los experimentos descritos en este artículo, ese sistema es la suite de analizadores de idiomas FreeLing[1] (Carreras et al., 2004), pero podría utilizarse otro sistema sin inconveniente y, al menos en teoría, incluso una anotación manual, aunque esto sería poco práctico debido a los volúmenes de datos que se requieren. En los experimentos aquí descritos, se utilizó como corpus de entrenamiento la Wikipedia en castellano.

Nuestro enfoque también asume el supuesto de partida de la lingüística textual de la correspondencia entre un referente y un texto, es decir que todas las menciones de un NP en un texto refieren a una misma entidad. Por tal motivo, el análisis procede con un texto a la vez y no un corpus como aglomeración de varios textos.

Los resultados muestran una tasa de acierto en las tareas de detección y clasificación de entidades comparables a las de los sistemas más conocidos. Cabe destacar, además, la simplicidad de la metodología ya que, al contrario de herramientas como FreeLing, no utiliza ningún tipo de pre ni post procesamiento de los textos. No requiere lematización, etiquetado morfológico ni análisis sintáctico.

En el sitio web que acompaña a este artículo se ofrece una implementación del algoritmo en código abierto, en forma de dos scripts Perl, uno para entrenamiento y otro para clasificación. Los dos scripts son muy sencillos, con menos de 100 líneas cada uno. Además del código, ofrecemos una demo en línea en la que se puede comparar el desempeño con otros clasificadores: http://www.tecling.com/pol

2 Marco teórico

El NP ha sido estudiado por disciplinas tan diversas como la geografía lingüística, la filosofía, la gramática comparada, la lexicología y la traductología, entre otras. Ofrecemos a continuación una breve caracterización del NP y una síntesis del trabajo relacionado que se ha producido en el ámbito del PLN.

2.1 Características de los NP

En la teoría lingüística existe consenso respecto a algunos rasgos prototípicos que caracterizan al NP (Fernández Leborans, 1999; RAE, 2009), entre los que podemos encontrar los siguientes: 1) **Introducción mediante mayúscula**: al menos en castellano y otras lenguas europeas, el NP se distingue gráficamente del nombre común (NC) por medio de este rasgo[2]; 2) **Flexión fija**: el NP se distingue en general del NC por carecer de flexión, salvo en casos muy específicos como los comentados por Coseriu (1982) donde el plural de los NP es indicio de un uso que se asimila al de un NC (ej.: *En la galería se exhiben varios Picassos*); 3) **Monorreferencialidad o unicidad referencial**: sería un defecto de forma que en un mismo texto un NP hiciera referencia a entidades distintas sin advertencia por parte del autor. Esto conlleva el hecho de que el NP no requiera el uso de artículo (definido o indefinido) y que de hecho su uso sin artículo sea normativo en todos los contextos (**El Jorge Luis Borges...*[3]); 4) **Falta de significado léxico**: los NP no poseen significado sino referencia (Russell, 1905) y

[1] http://nlp.lsi.upc.edu/freeling/

[2] Esto no es así en alemán, donde la mayúscula inicial es rasgo distintivo de la categoría gramatical nombre en general (*Das Auto*).

[3] A pesar de esto, el habla coloquial en Barcelona o regiones como Cuyo en Argentina registra el uso de artículo en el NP, posiblemente por influencia del catalán.

por esta razón no aparecen en los diccionarios lexicográficos, a menos que su uso pase a denominar una clase en lugar de un particular, como suele suceder por ejemplo en el caso de las marcas comerciales (*Comprar un kleenex*); 5) **Imposibilidad de traducción exacta**: la traducción o transliteración de los NP es a menudo problemática a menos que se trate de nombres ya convencionalizados (ej. *Miguel/ Michael/ Mijail*).

Estos o algunos de estos rasgos, que son inherentes a los NP, podrían ser objeto de interés para detectarlos y clasificarlos. Otros rasgos, aunque menos estudiados, podrían servir también, como por ejemplo un patrón de coocurrencia característico del NP. Esta es una de las variables que se pretende explorar en este trabajo, como indicador del valor o función referencial de una unidad léxica.

2.2 Detección/clasificación de NP

En el campo del PLN, el reconocimiento y la clasificación de NP tiene una larga historia en aplicaciones de recuperación o extracción de información y en la búsqueda de respuestas, dentro de la especialidad del *named entity recognition & classification* (Manning, Raghavan, y Schütze, 2008). Existen en la actualidad diversos sistemas para detectar y clasificar los NP, que en esta comunidad disciplinaria se denominan de manera genérica "entidades nombradas". Las categorías son principalmente las de antropónimo, topónimo o nombre de organización, aunque en rigor las entidades nombradas incluyen también fechas y signos diversos para designar valores o cantidades. Sin embargo, las clasificaciones pueden ser incluso más finas, como las exhibidas por Nadeau y Sekine (2007).

Un antecedente importante en la historia de este tipo de sistemas es el trabajo realizado en las Message Understanding Conferences (MUC) celebradas desde 1987, donde se demuestra que el reconocimiento de entidades es un componente fundamental de los sistemas de extracción de información, que requiere tanto del análisis léxico, sintáctico y en algunos casos hasta textual (Grishman y Sundheim, 1996; Wilks, 1998). Típicas estrategias utilizadas entonces y en la actualidad son el uso de *trigger-words* o palabras asociadas a un tipo de entidad, como "Inc." en el caso de las corporaciones, y las *gazetteer*, o listados de NP de distinto tipo.

Otro hito particularmente importante en el caso de la lengua castellana fue la celebración de la *Conference on Computational Natural Language Learning* (CoNLL-2002) por la tarea de reconocimiento de entidades utilizando un corpus de la agencia de noticias EFE (Tjong Kim Sang, 2002). Este corpus se convirtió en un estándar para medir el desempeño de los clasificadores que aparecieron después, e impuso la notación BIO (Ramshaw y Marcus, 1995; Tjong Kim Sang y Veenstra, 1999) para determinar el comienzo de un NP (B), su interior (I) o su finalización (0).

Carreras, Màrquez, y Padró (2002) obtuvieron el mejor desempeño en esta competencia con un algoritmo de aprendizaje automático (*AdaBoost*) y utilizando distintas pistas tales como las palabras del contexto, las categorías gramaticales, rasgos ortográficos, así como las *trigger-words* y *gazetteers*. Reportaron un 79 % de precisión y cobertura en la detección y un 81 % también en ambos valores en la clasificación, valores cercanos a los de algunos competidores como Florian (2002) con 78 % de precisión y 79 % de cobertura en la tarea de clasificación. Padró y Padró (2005) intentaron un modelo más económico en términos computacionales, basado en el algoritmo *Causal-States Splitting Reconstruction*, aunque sin llegar a superar esos resultados. Gamallo et al. (2014), presentando el programa CitiusNEC, describen una propuesta más compleja y de alto coste computacional, que integra diversas estrategias y que, sin embargo, tampoco superan los resultados informados por Carreras, Màrquez, y Padró (2002). Agerri, Bermudez, y Rigau (2014) son los que más se acercan, con una propuesta basada en el algoritmo de aprendizaje automático *Maximum Entropy*. Señalamos, sin embargo, el riesgo de evaluar siempre con ese mismo corpus ya que cuando se utilizan corpus distintos, y especialmente de otros géneros, las cifras de desempeño varían significativamente (van Hooland et al., 2015).

Se ha avanzado mucho desde las primeras conferencias MUC, y sin embargo el nicho de investigación continúa abierto hasta la actualidad, pues los sistemas de análisis aún no alcanzan su plenitud de desarrollo. Quedan además algunas lagunas. Si bien disponemos de algunos análisis detallados como el de Tkachenko y Simanovsky (2012) en corpus en inglés, queda por hacer un estudio del

peso relativo que tienen los distintos elementos gramaticales que acompañan al NP, como sería el caso de la colocación verbo-nominal, línea que sigue abierta a pesar de que Grishman y Sundheim (1996) ya la mencionaban. Otro vacío que se puede percibir en la literatura es la escasez de estrategias basadas en el enfoque de la lingüística textual, ya que en general no consideran al texto como unidad comunicativa con sentido completo (van Dijk, 1992). En el sentido más elemental, de esto se desprende por ejemplo que si una entidad es mencionada más de una vez dentro de un texto, cabe esperar que la referencia sea la misma en cada mención.

De todo el trabajo relacionado, el de Solorio (2004) es el que más se parece al presente artículo. Ella también utiliza FreeLing como una "caja negra" con la cual entrenar un clasificador. En su caso, el clasificador es *Support Vector Machines* y las pistas para el entrenamiento son las categorías gramaticales de las dos palabras anteriores y posteriores a la palabra analizada. Se puede decir, por tanto, que en este artículo exploramos y desarrollamos algunas consecuencias y variantes de esta idea.

La Tabla 1 exhibe los sistemas que utilizamos para evaluar el desempeño de nuestro método. Además de FreeLing y CitiusNEC, son Stanford-NERC (Finkel, Grenager, y Manning, 2005), una implementación de un algoritmo basado en el muestreo de Gibbs, y Semantria, un producto comercial de la empresa Lexalytics[4], basado en aprendizaje automático (*Conditional Random Fields*), expresiones regulares y extensas bases de datos de NP de distinto tipo.

Sistema	URL
CitiusNEC	gramatica.usc.es/pln
FreeLing 3.1	nlp.lsi.upc.edu/freeling
FreeLing 4.0	nlp.lsi.upc.edu/freeling
Semantria	semantria.com
Stanford	nlp.stanford.edu/software

Tabla 1: sistemas utilizados en la evaluación

3 Metodología

La metodología se basa en el diseño de un algoritmo de clasificación supervisado que, en una etapa de entrenamiento, recibe un corpus anotado y, en la etapa de prueba, utiliza este aprendizaje para detectar y clasificar NP en corpus sin anotar. Describimos a continuación materiales y procedimiento.

3.1 Materiales utilizados

Los materiales utilizados son un etiquetador de nombres propios, en nuestro caso FreeLing, y un corpus para utilizar como entrenamiento, en este caso la Wikipedia en castellano[5] (aprox. 586 millones de tokens). Cabe aclarar que esto incluye solamente el texto, es decir que se descartan todos los metadatos y la estructura de categorías de este recurso. No es necesario, por tanto, que el algoritmo se entrene únicamente con la Wikipedia, ya que podría ser entrenado con cualquier otro corpus que tenga un tamaño similar. Además de este recurso, también realizamos pruebas con el corpus de enegramas de Google Books (Lin et al., 2012) como una fuente para extraer nombres propios, aunque esto no es indispensable. Recalcamos que estos recursos solo se requieren durante la fase de entrenamiento ya que, una vez entrenado, el sistema funciona de manera completamente independiente.

Para la evaluación utilizamos el ya mencionado corpus CONLL-2002 (53.049 tokens) y un segundo corpus que confeccionamos a partir de una muestra intencionada de 8 artículos del periódico *Le Monde Diplomatique* en su edición en castellano (21.205 tokens). Las razones para utilizar este segundo corpus son diversas. En primer lugar, el corpus CONLL-2002 es menos apropiado para nuestra evaluación porque nuestro sistema está diseñado para analizar textos individuales. Otro motivo es disponer de un corpus de tamaño más reducido que nos permitiera tener mayor control para un análisis cualitativo y un control riguroso de lo que ocurre, ya que el corpus de CONLL-2002 contiene una considerable tasa de error en el etiquetado. Finalmente, un corpus de naturaleza tan distinta al de entrenamiento muestra un escenario más realista. En este caso, un corpus de geopolítica es especialmente exigente por las múltiples referencias a figuras políticas, territorios y organizaciones en situaciones diversas.

3.2 Procedimiento

La metodología se divide primero en fases de entrenamiento y clasificación, y luego cada

[4]https://www.lexalytics.com/

[5]https://dumps.wikimedia.org/eswiki/20160920/

una de estas se divide a su vez en dos.

3.2.1 Fase de entrenamiento

El entrenamiento tiene una primera fase más simple de recolección de NP y una segunda en la que asocia la categoría asignada por el otro sistema a cada NP con sus componentes y con elementos del contexto.

a) Extracción de un listado de NP: el primer paso consiste en obtener un listado de nombres propios, para lo cual utilizamos el corpus de enegramas Google Books (Lin et al., 2012) debido a su gran tamaño y libre disponibilidad. Para extraer los nombres propios se definió un coeficiente r (ecuación 1) que expresa la razón entre la frecuencia de una unidad léxica j escrita con mayúscula inicial (Mj) y la frecuencia total de esa misma unidad (Tj).

$$r_j = \frac{f(M_j)}{f(T_j)} \tag{1}$$

$$r_j \geq u \rightarrow j \in NP \tag{2}$$

Luego, (2), se consideró como perteneciente al conjunto de los NP a cualquier unidad j con un coeficiente r superior a un umbral arbitrario u[6].

b) Registro de las palabras del contexto: el paso anterior permite extraer listados de nombres propios, pero nuestro método requiere también asociar estos nombres a una categoría. Como ya se mencionó, esta parte del aprendizaje consiste en relacionar el vocabulario que aparece en el contexto del NP (en una ventana de n palabras) con los tipos de entidades. Para esta parte del entrenamiento utilizamos el corpus de la Wikipedia etiquetado previamente con el clasificador de entidades de FreeLing. A modo ilustrativo, considérese la Tabla 2 con un ejemplo de un fragmento de texto etiquetado.

Una considerable proporción de las apariciones de *Ottawa* y *Toronto* contendrán la palabra *ciudad* en su contexto inmediato, y lo mismo ocurrirá en el caso de muchas otras ciudades. Esto sugiere que la aparición de esta palabra en el contexto de estos NP servirá como predictor de que dicho NP sea un topónimo.

Forma	Lema	Etiqueta
Su	su	DP3CS0
capital	capital	NCFS000
es	ser	VSIP3S0
la	el	DA0FS0
ciudad	ciudad	NCFS000
de	de	SPS00
Ottawa	ottawa	NP00G00
y	y	CC
la	el	DA0FS0
ciudad	ciudad	NCFS000
más	más	RG
poblada	poblar	VMP00SF
es	ser	VSIP3S0
Toronto	toronto	NP00G00
.	.	Fp

Tabla 2: ejemplo del análisis de FreeLing en la clasificación de entidades

Del entrenamiento se deriva entonces un modelo que registra la frecuencia de aparición del vocabulario en el entorno de cada clase de NP. Podemos representar este modelo como una estructura de datos dividida en las tres claves que hemos considerado: P para la categoría de persona, O para organización y L para lugar (opcionalmente se puede incluir la categoría "otros"). Así, cuando en el entrenamiento el sistema encuentra una instancia de NP a la que se ha asignado una determinada categoría, se registrarán en cada caso las palabras que ocurren en las n posiciones anteriores, si es que estas no pertenecen a la lista de exclusión S, que contiene signos de puntuación y gramemas de alta frecuencia.

La función L (3) registra los elementos contextuales indicadores de lugar, aunque la representación es la misma en las tres categorías.

$$L\,(j) = \sum_{i=1}^{n} \begin{cases} 1 \text{ if } j \in C_i \wedge j \notin S \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

Por cada elemento i del conjunto C de contextos de aparición de los NP de categoría L en el entrenamiento, con $n = \text{card}(C)$, se registra la frecuencia del elemento contextual j, que podría ser por ejemplo la palabra *ciudad*. Por simplicidad, representamos el contexto C_i como un conjunto y no como se-

[6] En la versión original del artículo definíamos un umbral u de 0.8, lo que resulta en un algoritmo muy conservador (Tabla 3). Por sugerencia de los evaluadores, experimentamos con $u = 0$ y obtuvimos un mejor equilibrio entre precisión y cobertura. No eliminamos esta parte del artículo, sin embargo, porque podría darse un escenario donde interese más precisión que cobertura.

cuencia de palabras, ignorando por tanto posición y distancia. Registramos además solo la frecuencia absoluta. No hay normalización, lematización ni etiquetado morfosintáctico. Solo se registran las formas flexionadas de las palabras tal como aparecen en el corpus sin distinguir categorías gramaticales.

c) Registro de componentes del NP: Además de las palabras del contexto, de manera independiente el algoritmo asocia tipo de entidad con los componentes del mismo NP, con la misma mecánica que en el caso de los elementos del contexto. En el caso de un antropónimo, por ejemplo, tendríamos la distinción entre nombres de pila y apellidos. De esta manera, si en el corpus de entrenamiento se ha observado en reiteradas ocasiones que el componente *Vincent* forma parte de entidades que han sido clasificadas como antropónimos, entonces el sistema tendrá indicios para clasificar luego un nombre como *Vincent Heredia*, aunque este último nunca haya aparecido en el corpus de entrenamiento.

3.2.2 Fase de detección y clasificación

A partir de texto no anotado, la primera fase consiste en la detección de los NP, delimitando principio y final, para su posterior clasificación en las categorías ya mencionadas.

Para la primera parte se analiza de manera secuencial cada palabra del corpus de entrada, definida como una pieza léxica entre espacios en blanco o signos de puntuación. Si una palabra aparece con mayúscula inicial y no cumple un patrón de número romano ni adverbio y no ha recibido normalmente etiquetas distintas a la de NP en el corpus de entrenamiento, se declara el comienzo de un posible NP. Si la siguiente palabra comienza con minúscula y no es un gramema, entonces se delimita el final del candidato a NP. Esto permite la detección correcta de secuencias como *Ciudad del Cabo*, *Isla de Pascua*, etc. De estar definido el parámetro u, serían descartados los candidatos que no tengan al menos un componente que haya sido visto formando parte de un NP en el corpus de entrenamiento.

Luego, la etapa de clasificación consiste en dos partes. Primero, por cada componente j del NP, se seleccionará la categoría K_j que registre la frecuencia más alta para ese elemento (4). Así, si j es $Vincent$, entonces P_j registrará probablemente el valor más alto.

$$K_j = \max(P_j, O_j, L_j) \tag{4}$$

La segunda etapa es idéntica a la anterior, con la única diferencia de que no son las palabras del NP mismo las que se consideran sino las que aparecen en las n posiciones anteriores. Es en este punto donde cobra importancia tomar el texto completo como unidad de análisis en lugar de usar un contexto únicamente oracional, porque esto quiere decir que se estudian todas las menciones de la entidad en el texto y se dispone de más elementos para tomar la decisión sobre su categoría.

Para la decisión final se procede por votación simple, lo cual trae aparejeado el problema de los empates. En el caso de un NP de dos palabras, cada componente podría ser clasificado de manera distinta. Y siendo dos clasificadores (contexto y componentes) hay casos en que un mismo NP es clasificado de manera distinta. Una función de desempate calcula la certeza k de cada clasificador como la razón entre su primera y su segunda opción ($k = \frac{O_1}{O_2}$), y otorga precedencia el clasificador que se muestre más seguro.

4 Resultados

Presentamos a continuación una evaluación comparativa de los resultados tomando como referencia los sistemas presentados en otros trabajos. Evaluamos primero midiendo porcentajes de precisión y cobertura en la detección de NP en el corpus CONLL-2002 en comparación con los de sistemas más conocidos, tal como son reportados por Gamallo et al. (2014). La Tabla 3 informa los valores totales que se obtienen con el script *conlleval.pl* que proveen los organizadores de CONLL-2002, y el distinto desempeño según cómo se ajusten los parámetros u y n.

Sistema	precisión	cobertura	F1
CitiusNEC	67.47	66.33	66.89
FreeLing 3.1	75.08	76.60	75.98
POL $u = 0{,}8$	**90.24**	53.56	67.21
POL $\neg u \wedge n = 1$	76.04	81.52	78.68
POL $\neg u \wedge n = 2$	75.95	82.78	79.22
POL $\neg u \wedge n = 3$	76.06	83.28	**79.51**
POL $\neg u \wedge n = 4$	76.03	**83.33**	79.51
OpenNLP	78.96	79.09	79.02

Tabla 3: Evaluación comparativa de los diversos sistemas para la detección de NP en el corpus CONLL-2002

En general se aprecia que los resultados son similares a los informados por otros autores, excepto cuando se define un u alto, que resulta en alta precisión y baja cobertura.

Repetimos los experimentos utilizando el segundo corpus de evaluación, el de la revista *Le Monde Diplomatique*, y comparamos el desempeño en ese corpus con los cinco sistemas NERC mencionados en la Tabla 1. Para ello, los autores detectamos y clasificamos manualmente todos los NP en este corpus, revisando mutuamente nuestra anotación para evitar posibles casos de desacuerdo. Este procedimiento arrojó un total de 537 entidades distintas. La Tabla 4 muestra los resultados de la detección (Pol con $n = 3$). Se aprecia que en general los valores son comparables, con un patrón similar al caso anterior. POL con $u = 0{,}8$ es preciso pero excesivamente conservador, y Semantria muestra esta tendencia de manera incluso más acusada.

Sistema	**precisón**	**cobertura**	**F1**
CitiusNEC	67	88	76
FreeLing 3.1	74	88	80
FreeLing 4.0	77	**88**	**82**
POL $u = 0{,}8$	83	66	73
POL $\neg u$	74	79	76
Semantria	**94**	38	54
Stanford	73	76	74

Tabla 4: Evaluación comparativa de los diversos sistemas para la detección de NP en el corpus de *Le Monde Diplomatique* en castellano

Sistema	**Correctos**	**Totales**	**Precisión**
CitiusNEC	336	477	70
FreeLing 3.1	358	476	75
FreeLing 4.0	359	477	75
POL $u = 0{,}8$	273	358	76
POL $\neg u$	302	429	70
Semantria	179	205	**87**
Stanford	313	409	76

Tabla 5: Evaluación comparativa de los diversos sistemas en la clasificación de las entidades antes detectadas en el corpus de *Le Monde Diplomatique* en castellano

En la Tabla 5 se exhiben los resultados de la tarea de clasificación de las entidades que han sido previamente detectadas por cada sistema. En este caso también, Semantria lleva la delantera en precisión, aunque la cobertura es muy escasa. El resto de los sistemas muestra una precisión similar.

Las limitaciones de espacio impiden la inclusión de más variantes en los parámetros, pero las que probamos, como el uso del contexto derecho, no ofrecen mejores resultados. Tampoco podemos ofrecer un análisis pormenorizado de los errores que pudimos observar en los resultados de los distintos sistemas evaluados, aunque en general podemos resaltar, en la tarea de detección, la dificultad de desambiguar palabras que se introducen con una mayúscula pero no son NP ni ejercen función de sustantivo en la oración, lo que sucede particularmente en CitiusNEC y las dos versiones de FreeLing, estas dos últimas muy similares entre sí.

En cuanto a la tarea de clasificación, son fuente de problemas los casos de polisemia regular, donde resulta difícil distinguir por ejemplo si un NP actúa como un topónimo o nombre de organización. En al caso de nuestro algoritmo, el error más frecuente es clasificar nombres de organizaciones que también aplican a nombres de personas, como *Louis Vuitton*.

5 Conclusiones y trabajo futuro

Hemos presentado una metodología para detección y clasificación de NP por medio de un algoritmo que se entrena con los resultados de otro sistema. La propuesta se caracteriza por su simplicidad y bajo coste computacional al no requerir ningún procesamiento del corpus. El corpus CONLL-2002 se procesa en menos de un segundo en un PC de escritorio, lo cual representa una ventaja importante. En cuanto a desempeño, la evaluación comparativa demuestra que es comparable al de otros sistemas más complejos. Los resultados sugieren que un enfoque cuantitativo de este tipo con entrenamiento y test es adecuado, ya que extrae los datos directamente del corpus y no de la introspección del investigador, como los modelos basados en reglas. La simplicidad del método, además, facilita su portabilidad a otras lenguas con menos recursos.

Como trabajo futuro, debemos continuar experimentando con distintos tamaños y tipos de ventanas de contexto y de corpus de entrenamiento, así como reproducir los experimentos en otras lenguas. Otra vía de investigación es estudiar el poder predictor de las unidades del contexto en función de su categoría gramatical y también en función de las relaciones sintácticas que contraen los elementos del contexto con el NP.

Bibliografía

Agerri, R., J. Bermudez, y G. Rigau. 2014. Ixa pipeline: Efficient and ready to use multilingual nlp tools. En *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, páginas 3823–3828, Reykjavik, Iceland. European Language Resources Association (ELRA).

Arriagada, P. 2016. *Análisis y clasificación de nombres propios en artículos de geopolítica de la revista Le Monde Diplomatique: una aproximación desde la gramática del texto*. Tesis de grado, Pontificia Universidad Católica de Valparaíso.

Carreras, X., I. Chao, L. Padró, y M. Padró. 2004. Freeling: An open-source suite of language analyzers. En *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, páginas 239–242.

Carreras, X., L. Màrquez, y L. Padró. 2002. Named entity extraction using adaboost. En *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, páginas 167–170, Stroudsburg, PA, USA. ACL.

Coseriu, E. 1982. El plural en los nombres propios. En *Teoría del Lenguaje y Lingüística General*. Gredos, Madrid, páginas 261–281.

De Miguel, E. 1999. El aspecto léxico. En I. Bosque y V. Demonte, editores, *Gramática descriptiva de la lengua española*. Espasa Calpe, Madrid, páginas 2977–3060.

Fernández Leborans, M. J. 1999. El nombre propio. En I. Bosque y V. Demonte, editores, *Gramática descriptiva de la lengua española*. Espasa Calpe, Madrid, páginas 77–128.

Finkel, J. R., T. Grenager, y C. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. En *Proceedings of the 43nd Annual Meeting of the ACL*, páginas 363–370.

Florian, R. 2002. Named entity recognition as a house of cards: Classifier stacking. En *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, páginas 1–4, Stroudsburg, PA, USA. ACL.

Gamallo, P., J. C. Pichel, M. Garcia, J. M. Abuín, y T. Fernández-Pena. 2014. Análisis morfosintáctico y clasificación de entidades nombradas en un entorno big data. *Procesamiento del Lenguaje Natural*, (53):17–24.

Grishman, R. y B. Sundheim. 1996. Message understanding conference-6: a brief history. En *16th International Conference on Computational Linguistics*, páginas 466–471.

Lin, Y., J.-B. Michel, E. L. Aiden, J. Orwant, W. Brockman, y S. Petrov. 2012. Syntactic annotations for the google books ngram corpus. En *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, páginas 169–174, Stroudsburg, PA, USA. ACL.

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

Nadeau, D. y S. Sekine. 2007. A survey of named entity recognition and classification. *Journal of Linguisticae Investigationes*, 30(1):1–20.

Padró, M. y L. Padró. 2005. A named entity recognition system based on a finite automata acquisition algorithm. *Procesamiento del Lenguaje Natural*, (35):319–326.

RAE. 2009. *Nueva gramática de la lengua española*. Espasa Libros, Madrid.

Ramshaw, L. A. y M. P. Marcus. 1995. Text chunking using transformation-based learning. En *Third Workshop on Very Large Corpora*, páginas 82–94. ACL.

Russell, B. 1905. On denoting. *Mind*, (14):479–493.

Solorio, T. 2004. Improvement of named entity tagging by machine learning. Informe técnico, Coordinación de Ciencias Computacionales INAOE (No. CCC-04-004), Puebla, México.

Tjong Kim Sang, E. F. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. En *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, páginas 1–4, Stroudsburg, PA, USA. ACL.

Tjong Kim Sang, E. F. y J. Veenstra. 1999. Representing text chunks. En *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, EACL '99, páginas 173–179, Stroudsburg, PA, USA. ACL.

Tkachenko, M. y A. Simanovsky. 2012. Named entity recognition: Exploring features. En J. Jancsary, editor, *Proceedings of KONVENS 2012*, páginas 118–127. ÖGAI.

van Dijk, T. A. 1992. *La ciencia del texto*. Paidós, Barcelona.

van Hooland, S., M. D. Wilde, R. Verborgh, T. Steiner, y R. V. de Walle. 2015. Exploring entity recognition and disambiguation for cultural heritage collections. *DSH*, 30(2):262–279.

Wilks, Y. 1998. Sense and texts. *Computational Linguistics and Chinese Language Processing*, 3(2):1–16.

Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 21-28 recibido 09-11-2016 revisado 17-01-2017 aceptado 06-02-2017

Desambiguación Verbal Automática: un estudio sobre el rendimiento de la información semántica argumental

Verb Sense Disambiguation: a study about the performance of argumental semantic information

José Priego García
Universitat de Barcelona
Gran Via 585, 08007, Barcelona
jpriegga@gmail.com

Irene Castellón Masalles
Universitat de Barcelona
Gran Via 585, 08007, Barcelona
icastellon@ub.edu

Resumen: Una de las tareas fundamentales para la resolución de la ambigüedad en el ámbito del Procesamiento del Lenguaje Natural es la Desambiguación Semántica Automática; especialmente la tarea específica de Desambiguación Verbal Automática (DVA). En la presente investigación se lleva a cabo una tarea experimental con la finalidad de comprobar la viabilidad de una aproximación a la DVA basada en la información semántica de los argumentos verbales. Los buenos resultados obtenidos indicarían la necesidad de tener en cuenta este tipo de información en futuras propuestas de DVA.
Palabras clave: Semántica, procesamiento del lenguaje natural, desambiguación verbal automática, aprendizaje automático

Abstract: One of the key tasks for resolving the ambiguity in the field of Natural Language Processing is Word Sense Disambiguation; especially the specific task of Verb Sense Disambiguation (VSD). In the present study an experimental task is performed in order to test the feasibility of an approach to VSD based on semantic information about verbal arguments. The good results obtained indicate the need to take into account this information in future proposals for VSD.
Keywords: Semantics, natural language processing, verb sense disambiguation, machine learning

1 Introducción

Una de las características intrínsecas del lenguaje y que se hace patente de forma diaria en la comunicación mediante cualquier lengua es su ambigüedad. Todas las lenguas del mundo poseen multitud de palabras polisémicas que pueden hacer referencia a diferentes sentidos en función del contexto de uso, y resulta imprescindible saber identificar a cuál de ellos se está haciendo referencia en cada ocasión si se quiere comprender correctamente un mensaje. Por ejemplo, ante la oración "los participantes partieron la mañana del viernes" el hablante deberá identificar que el sentido activado por el verbo "partir" en esta ocasión es "abandonar un lugar" y no "fragmentar algo". En el campo del Procesamiento del Lenguaje Natural (PLN) la ambigüedad supone un verdadero obstáculo para el correcto desempeño de cualquier aplicación y es por ello que desde sus inicios se ha intentado emular la capacidad de desambiguación humana mediante el diseño y desarrollo de sistemas de Desambiguación Semántica Automática (DSA).

La investigación que se presenta a continuación se encuentra organizada del modo siguiente: en primer lugar, se realizará una breve descripción de la DSA, sus características y posibles aproximaciones, centrando especial atención al caso concreto de la Desambiguación Verbal Automática (DVA). En segundo lugar, se propondrá y evaluará una aproximación a la DVA basada en información semántica acerca de los argumentos verbales.

2 Desambiguación Semántica Automática

La DSA podría describirse como una tarea de clasificación: un programa informático (clasificador) debe asignar automáticamente a una determinada palabra uno de sus sentidos posibles (clases) en un contexto de uso concreto (Navigli, 2009; Pal y Saha, 2015). En tanto que trata de emular una capacidad cognitiva humana,

ISSN 1135-5948

la DSA es considerada una tarea propia del campo de la Inteligencia Artificial. Con el paso de los años, la necesidad de crear técnicas de desambiguación económicas y versátiles ha llevado al surgimiento de tres principales aproximaciones a la DSA que divergen en el tipo de fuente informativa empleada: métodos basados en conocimiento que emplean bases de datos estructuradas como tesauros y lexicones computacionales (Del Corro et al., 2014), métodos supervisados que combinan corpus anotados y sistemas de aprendizaje automático (Zhong y Tou, 2010) y, finalmente, métodos no-supervisados que extraen la información mediante cálculo estadístico a partir de corpus planos (Wang et al., 2013).

2.1 Desambiguación Verbal Automática

Del mismo modo que a la hora de llevar a cabo una tarea de DSA se puede escoger si el objeto a desambiguar serán todas las palabras del texto o únicamente una de ellas, también es posible centrar el objeto de desambiguación en una única categoría morfológica como, por ejemplo, el verbo. El actual interés por la desambiguación verbal responde a dos motivos principales: los escasos resultados conseguidos hasta la fecha en cuanto a la desambiguación de esta categoría y el papel central del verbo en la estructura oracional. Comenzando por el primero de ellos, en la actualidad los mejores resultados obtenidos en una tarea de DVA se hallan en torno a un 82% de verbos correctamente desambiguados (Del Corro et al., 2014). Se trata de unos resultados que se encuentran lejos de lo deseable en este tipo de tareas al dejar todavía un considerable margen de error y que, por lo tanto, invitan a seguir investigando con el fin de mejorarlos. Esta necesidad de mejora se corrobora si se tiene en cuenta, como se señalaba anteriormente, el papel relevante del verbo dentro de la estructura oracional como núcleo del predicado y administrador de los constituyentes de la oración mediante la selección de argumentos, así como la adjudicación de funciones sintácticas y papeles temáticos. Así pues, debido a su centralidad, una mejora en la desambiguación verbal podría suponer de forma colateral una mejora en aquellas tareas de DSA en las que se pretenda desambiguar todas las palabras de un texto y, por lo tanto, en el PLN en general.

3 Información semántica de los argumentos verbales en la DVA

El grueso de las investigaciones previas en el campo de la DSA de corte general acostumbra a emplear una serie de atributos concretos; mayoritariamente información acerca de la categoría morfológica y función sintáctica de las palabras próximas al término a desambiguar. Pese a funcionar de forma correcta en tareas generales, la aplicación de estos recursos a tareas específicas de DVA no parece que haya reportado hasta la fecha unos resultados tan buenos como se podría desear (Buscaldi et al., 2006). Ante esta situación resulta evidente la necesidad de explorar nuevas fuentes de información relacionadas con el verbo que contribuyan a mejorar el rendimiento de las tareas de DVA y, en definitiva, el campo de la DSA en su conjunto. Como se señalaba en la sección anterior, una característica intrínseca del verbo que quizás no ha recibido la atención merecida hasta tiempos recientes es la estrecha relación sintáctico-semántica que este establece con sus diferentes argumentos; una relación que puede condicionar la tipología de los argumentos en función del sentido activado por el verbo en un contexto de uso determinado. Puede suponerse, pues, que el rendimiento de las tareas de DVA podría aumentar si se contemplara en ellas información relativa a esta relación verbo-argumento; especialmente si, además de información morfológica y sintáctica, se contara también con información de tipo semántico. Esta necesidad de explotar la información argumental en el campo de la DVA ya ha sido señalada por recientes investigaciones que la incluyen de diversas formas en sus diseños experimentales; ya sea mediante el uso de preferencias de selección (Ye y Baldwin, 2006), el análisis estadístico y *clustering* de parejas verbo-argumento (Wagner et al., 2009) o mediante la inclusión de información semántica acerca de los argumentos verbales en un corpus anotado (Dang y Palmer, 2005; Dligach y Palmer, 2008).

Siguiendo la tendencia iniciada por estos estudios, en la presente investigación se llevará a cabo una tarea experimental mediante la cual se pretende evaluar el rendimiento de diferentes tipos de información semántica sobre los argumentos verbales en el proceso de DSA de una serie de verbos.

3.1 Fuentes de información empleadas

Debido a que el objetivo final de esta tarea es la evaluación del rendimiento de diversos tipos de información lingüística procedentes de diferentes fuentes, se ha considerado necesario adoptar una aproximación a la DVA de tipo supervisado. Como ya se ha apuntado anteriormente, la aproximación supervisada extrae la información de corpus etiquetados con información lingüística; en este caso, las características de los argumentos de cada verbo y el sentido activado por este en cada oración. De este modo, la aproximación supervisada otorga la posibilidad de realizar diferentes combinaciones informativas según se requiera en cada experimento.

La presente investigación toma como muestra un subconjunto del corpus Sensem (Alonso et al., 2007; Fernández y Vázquez, 2014). Dicho corpus ha sido anotado automáticamente en el nivel morfológico mediante Freeling (Padró, 2011) y de forma manual en los niveles sintáctico y semántico; incluyendo este último la desambiguación léxica de los sentidos verbales y argumentos nominales tomando WordNet como referencia (Castellón et al., 2012; Gonzalez-Agirre et al., 2012). El subconjunto empleado en esta investigación cuenta con un total de 1.033 oraciones correspondientes a 12 verbos que se hallan en un rango de entre dos y cinco sentidos (véase cabecera de las tablas de resultados); lo que supone entre 65 y 100 oraciones por cada uno de los verbos escogidos.

La selección de lemas verbales se ha realizado respetando tanto la frecuencia de aparición de los diferentes sentidos en el corpus como la necesidad de tener un mínimo de ejemplos por sentido a partir de los cuales entrenar eficientemente al clasificador. Así pues, ninguno de los sentidos verbales posee menos de un 6% de los ejemplos recogidos para su lema.

Al tratarse de un corpus anotado, las diversas oraciones que lo componen contienen información relativa a diferentes niveles de análisis lingüístico. A nivel léxico, cada una ha sido etiquetada con el lema correspondiente a su verbo principal; verbo que es, además, el candidato a desambiguar. Así mismo, cada oración cuenta con el sentido específico que toma dicho verbo en ese ejemplo concreto. Siguiendo la línea de investigaciones anteriores, cada oración recoge también información acerca de la categoría morfológica y función sintáctica de cada uno de los argumentos verbales. En lo tocante al nivel semántico, el realmente relevante en esta investigación, los argumentos de cada oración han sido etiquetados según su papel temático y según su clasificación en tres ontologías de uso común en el ámbito del PLN: la Suggested Upper Merged Ontology (Pease et al., 2002), los Supersenses de WordNet (Fellbaum, 2005) y la Top Concept Ontology (Álvez et al., 2008).

4 Metodología

Como se ha señalado anteriormente, la DSA se puede definir básicamente como una tarea de clasificación. Siguiendo la aproximación supervisada, el clasificador toma los ejemplos que constituyen el corpus de entrenamiento, establece relaciones a partir de estos mediante un proceso de aprendizaje basado en un determinado algoritmo y, finalmente, realiza una clasificación de las instancias contenidas en un corpus de evaluación en el que estas carecen de un valor para el atributo que se pretende averiguar. En el caso concreto de esta investigación, el clasificador toma como ejemplo las oraciones contenidas en el corpus anteriormente descrito y, tras abstraer durante el aprendizaje aquellas características que los definen, otorga el sentido más probable, dentro de los posibles, para cada verbo de las diferentes oraciones existentes en el corpus de evaluación. Cada uno de los experimentos se ha realizado siguiendo un proceso de *10-fold cross validation*; de modo que el resultado de cada uno de estos se deriva de la media de diez pruebas realizadas con diferentes corpus de evaluación autogenerados a partir del corpus completo.

Los diferentes experimentos se han llevado a cabo mediante la plataforma de software para la minería de datos y el aprendizaje automático Weka (Witten et al., 2011), desarrollada por la Universidad de Waikato (Nueva Zelanda). Para ello se ha tenido que adaptar la información contenida en el corpus anotado a un formato específico capaz de ser procesado por esta aplicación. Tras esta adaptación, cada una de las oraciones ha sido transformada en una instancia en forma de vector que presenta un único valor categorial para cada uno de los atributos anteriormente descritos.

	[2] Beneficiar	[2] Defender	[2] Facilitar	[3] Detener	[3] Interpretar	[3] Superar	[4] Hallar	[4] Partir	[4] Tratar	[5] Acabar	[5] Actuar	[5] Mover	Corpus
Baseline	61.8	64.6	65.0	52.3	48.3	41.9	36.1	41.7	76.8	40.6	60.0	32.3	52.1
MS	**97.9**	63.4	66.0	89.5	75.8	39.5	59.5	59.3	**97.5**	59.3	72.8	52.3	70.1
MS + PTs	96.9	**95.1**	76.6	**97.6**	76.9	**98.7**	**89.3**	**95.6**	**97.5**	73.6	**88.5**	60.0	87.7
PTs	96.9	**95.1**	**80.5**	96.5	**79.1**	**98.7**	**89.3**	**95.6**	**97.5**	**80.2**	84.2	**67.6**	**89.2**

Tabla 1: Resultados obtenidos a partir de *baseline*, información morfológico-sintáctica (MS) y papeles temáticos (PTs). Para cada verbo se indica entre [ ] su número de sentidos. Los mejores resultados de cada verbo se encuentran resaltados en negrita.

En el caso de esta investigación se ha considerado oportuno el uso de un algoritmo basado en el principio de *support vector machines* (SVM); concretamente, el algoritmo Sequential Minimal Optimization (Platt, 1998). Debido a sus características, un algoritmo basado en vectores de soporte se encuentra limitado a realizar clasificaciones binarias; es decir, solo puede realizar una clasificación entre dos valores distintos para un mismo atributo. De este modo, en la investigación presente, el algoritmo creará un único plano durante la clasificación de verbos con dos sentidos, pero deberá crear (n·(n-1))/2 planos distintos cuando el verbo a desambiguar presente un mayor número de sentidos (n) con tal de realizar todas las combinaciones entre los valores existentes para, posteriormente, emitir un único resultado de clasificación a partir de estas. A pesar de esta particularidad, los algoritmos basados en SVM presentan mejores resultados que otros algoritmos típicamente empleados en DSA como pueden ser los de tipo probabilístico bayesianos como Naive Bayes o los basados en instancias vecinas como k-Nearest Neighbor (Escudero, 2006; Witten et al., 2011). En comparación con estos últimos, presentan, además, una clara ventaja en tanto que no son susceptibles a posibles desviaciones causadas por diferencias en el número de instancias contenidas en cada grupo al basarse únicamente en aquellas dos que actúan como vectores de soporte.

5 Resultados

A continuación, se realizará una descripción detallada de los resultados obtenidos en esta investigación. En primer lugar, se establecerán como punto de referencia los resultados obtenidos mediante el cálculo de la *baseline*, así como los resultados de una clasificación basada únicamente en los atributos de categoría morfológica y función sintáctica. En segundo lugar, se llevará a cabo una serie de experimentos bajo diferentes configuraciones de los atributos semánticos anteriormente descritos. En todo momento se analizarán en primer lugar los resultados obtenidos en las pruebas realizadas de forma individual para cada verbo y en segundo lugar los resultados pertenecientes a aquella prueba realizada con el corpus completo.

5.1 Baseline e información morfológica y sintáctica

La *baseline* de cada verbo se ha obtenido siguiendo el proceso de clasificación *most frequent sense*, en el cual se asigna el valor predominante para el atributo "sentido" en todas las instancias pertenecientes a cada uno de los verbos presentes en el corpus. Los cálculos se han llevado a cabo mediante ZeroR, un algoritmo que permite el cálculo de la *baseline* de forma automática y que se halla implementado en Weka. En la Tabla 1 se puede observar que, como cabría esperar, tanto los resultados individuales (entre el 32,3% y el 76,8% de instancias correctamente clasificadas) como el de la clasificación realizada con el corpus completo (52,1%) son insuficientes para una tarea de este tipo.

En cuanto a las pruebas realizadas con información morfológica y sintáctica (MS), si bien destacan dos verbos que responden excepcionalmente bien a estas informaciones (97 puntos), las pruebas individuales arrojan unos resultados también pobres alrededor de los 60 puntos (con máximos de 89,5 y mínimos de 39,5). De modo similar, la prueba realizada con el corpus completo alcanza los 70,1 puntos. El hecho de que la desambiguación de determinados verbos como "beneficiar" y

	[2] Beneficiar	[2] Defender	[2] Facilitar	[3] Detener	[3] Interpretar	[3] Superar	[4] Hallar	[4] Partir	[4] Tratar	[5] Acabar	[5] Actuar	[5] Mover	Corpus
Ont+MS+PTs	**96.9**	89.0	85.4	93.0	**83.5**	**98.7**	**90.4**	**94.5**	**97.5**	75.8	**87.1**	**70.7**	88.2
Ont+PTs	90.7	**90.2**	**88.3**	**94.1**	**83.5**	97.5	89.3	93.4	**97.5**	**78.0**	80.0	69.2	**88.6**
Ont+MS	**96.9**	60.9	75.7	86.0	81.3	33.3	75.5	62.6	**97.5**	61.5	72.8	58.4	72.0
Ont	56.7	58.5	74.7	47.6	70.3	37.0	39.3	54.9	71.9	37.3	58.5	44.6	55.6

Tabla 2: Resultados obtenidos en la evaluación conjunta de las tres ontologías (Ont) y su combinación con información morfológico-sintáctica (MS) y papeles temáticos (PTs).

"tratar" logre unos resultados tan elevados únicamente con este tipo de información probablemente se deba a la existencia de unas fuertes preferencias de tipo sintáctico, según las cuales cada uno de los sentidos determina claramente la tipología a la que deben ajustarse sus posibles argumentos.

Pese a que la inclusión de información morfológica y sintáctica supone un aumento del rendimiento de estas tareas, lo cierto es que la levedad de este corrobora la necesidad de explorar nuevas fuentes de información como las que se evaluarán a continuación.

5.2 Papeles temáticos

En una nueva prueba, además de la información morfológica y sintáctica de cada argumento, se ha tenido en cuenta el atributo correspondiente al papel temático (PTs) representado por cada uno de los argumentos del verbo a desambiguar. Los resultados obtenidos parecen indicar que la inclusión de los papeles temáticos incrementa de forma considerable el porcentaje de verbos correctamente desambiguados; que en esta ocasión se halla entre los 60 y 98,7 puntos en las tareas individuales y alcanza los 87,7 puntos en la tarea conjunta.

Por otro lado, resultan especialmente interesantes los resultados obtenidos en una prueba basada únicamente en información sobre los papeles temáticos; prescindiendo, por lo tanto, de información morfológica y sintáctica. Como se recoge en la Tabla 1, los resultados obtenidos en esta ocasión muestran mejoras generales respecto a las tareas anteriores de entre 3 y 7 puntos en las tareas individuales y de 1,5 puntos al emplear el corpus completo.

5.3 Clasificación semántica

Mediante la siguiente batería de pruebas se tratará de comprobar hasta qué punto la información semántica de los diferentes argumentos verbales, expresada según la clasificación de diversas ontologías (Ont), puede ayudar a mejorar los resultados obtenidos anteriormente o si bien puede incluso llegar a sustentar una aplicación de DVA por sí sola. Los resultados de estas se muestran en la Tabla 2.

En primer lugar, se ha podido comprobar que la adición simultánea de la información contenida en cada una de las tres ontologías antes presentadas (SUMO, Supersenses de WordNet y TCO) produce resultados heterogéneos si se aplica a la configuración anterior basada en información morfológico-sintáctica y papeles temáticos. El resultado de las tareas individuales tiende a mostrar mejoras de hasta 10 puntos respecto a la configuración anterior mientras que en la tarea con el corpus completo, en cambio, la mejora es de tan solo 0,5 puntos. Como podría esperarse si se tienen en cuenta las pruebas anteriores, si se realiza la misma prueba sustrayendo esta vez la información morfológica y sintáctica puede observarse una leve mejora en los resultados (0,4 puntos con el corpus completo). Por otro lado, si se mantiene esta y se prescinde de los papeles temáticos se produce un claro descenso de los resultados (16 puntos con el corpus completo). Cabe destacar que este descenso es especialmente notable en las pruebas individuales, que llegan a perder de 2 a 65 puntos respecto a la tarea inicial. Finalmente, si se trata de llevar a cabo una tarea de desambiguación basada únicamente en la combinación de estas tres ontologías, los resultados parecen indicar que se trata de algo inviable: en ninguno de los casos se logra mantener el resultado de pruebas anteriores y los resultados individuales sufren caídas de entre 11 y 61 puntos mientras que la prueba con el corpus completo cae en 33 puntos.

	[2] Beneficiar	[2] Defender	[2] Facilitar	[3] Detener	[3] Interpretar	[3] Superar	[4] Hallar	[4] Partir	[4] Tratar	[5] Acabar	[5] Actuar	[5] Mover	Corpus
PTS+SUMO	96.9	93.9	83.4	96.5	83.5	97.5	88.2	92.3	97.5	**80.2**	**84.2**	75.3	89.3
PTs+Ss	96.9	**95.1**	83.4	94.1	**87.9**	97.5	85.1	**95.6**	97.5	79.1	82.8	**81.5**	90.1
PTs+TCO	96.9	93.9	83.4	96.5	84.6	97.5	**91.4**	93.4	97.5	78.0	82.8	76.9	**90.3**
PTs+TCO+Clust	96.9	**95.1**	**88.3**	**97.6**	**87.9**	97.5	**91.4**	93.4	97.5	78.0	80.0	**81.5**	90.1

Tabla 3: Resultados obtenidos en la evaluación individual de las ontologías SUMO, Supersenses de Wordnet (Ss) y TCO junto a los papeles temáticos (PTs). En último lugar, se indican también los resultados obtenidos al incluir a esta última los *clusters* (Clust) basados en TCO.

La heterogeneidad de los resultados obtenidos en la serie de pruebas anterior, así como la diferente naturaleza de las tres ontologías empleadas, hace que sea prudente examinar el rendimiento de cada una de las ontologías de forma individual. Así pues, se ha realizado de nuevo la serie de pruebas anterior empleando esta vez las ontologías de forma aislada y no al unísono. Los resultados proporcionados por esta serie de pruebas han permitido establecer dos conclusiones principales. En primer lugar, se mantiene la gradación de rendimiento de las configuraciones empleadas: de nuevo, la configuración basada en papeles temáticos y una determinada ontología resulta la más adecuada; seguida de su variante que incluye información morfológico-sintáctica, a continuación por la que emplea información morfológica y sintáctica pero no papeles temáticos y, finalmente, por aquella que no emplea otra información que la recogida en la ontología. La segunda conclusión que se deriva de esta serie de pruebas es que, pese a que en las tareas individuales cada verbo parece recibir un mayor beneficio por parte de una determinada ontología, los resultados tanto de la prueba con el corpus completo como del conjunto de las tareas individuales parecen establecer una jerarquía de rendimiento entre las diferentes ontologías empleadas: TCO supera ligeramente a los Supersenses de WordNet en la mayoría de ocasiones y estos dos superan con claridad a SUMO. La diferencia de rendimiento entre las diferentes ontologías puede observarse en los resultados representados en la Tabla 3.

5.4 Aumentando el rendimiento de TCO

Los resultados obtenidos hasta el momento dejan patente que la inclusión de TCO produce un mayor rendimiento que los Supersenses de WordNet y, especialmente, SUMO. Sin embargo, como se apuntaba anteriormente en la descripción de la metodología de esta investigación, el hecho de que el formato requerido por Weka admita un único valor para cada atributo limita en gran medida la utilidad de TCO si se tiene en cuenta su naturaleza jerárquica y componencial. De este modo, la representación compactada de los rasgos de TCO lleva al clasificador a entender como valores diametralmente distintos todos aquellos que se diferencien por un solo rasgo aun compartiendo parte o la totalidad de los demás.

Esta situación lleva a preguntarse si sería posible generalizar de algún modo la información sobre los argumentos verbales expresada según esta ontología, de forma que se pudiera establecer una relación entre los diferentes argumentos encontrados para cada sentido de un determinado verbo y, de este modo, caracterizar las preferencias de selección de cada sentido verbal. Para ello se ha optado por realizar la descomposición de la información aportada por TCO en sus categorías semánticas individuales y, posteriormente, llevar a cabo un proceso de *clustering* en tantos grupos como sentidos posee un determinado verbo. El proceso de *clustering* se ha llevado a cabo mediante el algoritmo Simple-K-Means implementado en Weka.

Al realizar una nueva prueba de clasificación donde además de papeles temáticos y TCO se han incluido dichos *clusters* (Clust), se ha observado que los resultados no solo se mantienen respecto a la prueba anterior, sino que además, como se muestra en la Tabla 3, pueden mejorar entre 1 y 5 puntos en las pruebas individuales de ciertos verbos. En la prueba realizada con el corpus completo, en cambio, se ha observado una pérdida de 0,2 puntos. Pese a

este ligero descenso, parece ser que la inclusión de información producto del *clustering* permite salvar, al menos en parte, las limitaciones impuestas por el formato establecido por el clasificador y ayuda a mejorar los resultados obtenidos hasta el momento.

5.5 Discusión de los resultados

Los resultados obtenidos en las pruebas anteriores permiten realizar una serie de consideraciones sobre la utilidad de las diferentes informaciones empleadas. En primer lugar, se ha podido comprobar que el uso exclusivo de información morfológica y sintáctica acerca de los argumentos verbales mejora los resultados proporcionados por la *baseline* pero, a pesar de ello, resulta insuficiente para llevar a cabo una tarea de desambiguación con resultados aceptables. En segundo lugar, se ha comprobado con claridad que la adición de información relativa a los papeles temáticos desempeñados por cada argumento incrementa de forma notable el rendimiento de estas tareas. Es más, de hecho, se ha podido observar que la información sobre papeles temáticos puede ofrecer mejores resultados si prescinde de información morfológica y sintáctica; probablemente debido a que dichos papeles temáticos ya suponen de forma implícita una representación de las características morfológicas y sintácticas de cada argumento verbal. En tercer lugar, se ha podido comprobar también cómo el uso de la información proporcionada por ciertas ontologías mejora la desambiguación de la mayoría de los verbos examinados. En todo caso, los mejores resultados se han obtenido con el apoyo de los papeles temáticos, prescindiendo de la información morfológica y sintáctica, y empleando únicamente una de las tres ontologías señaladas. En cuanto a las diferencias de rendimiento entre estas, se ha podido observar que SUMO se ve claramente superada por los Supersenses y TCO; siendo esta última la más fructífera de las tres. Esto probablemente se deba al hecho de que las 2.302 categorías que conforman SUMO la hacen una ontología demasiado específica; cosa que impediría establecer de forma adecuada generalizaciones a partir de diferentes argumentos relacionados con un mismo sentido verbal. Por otro lado, esto no parece suceder con recursos como los Supersenses o TCO, que con un menor número de categorías (25 y 64, respectivamente) reflejan un mayor grado de abstracción y, por lo tanto, sí permitirían relacionar argumentos diversos. En relación con esta última ontología, cabe destacar que los resultados obtenidos a partir de la adición de información basada en el *clustering* de sus categorías parecen ser el indicador de una posible vía para aumentar su contribución en tareas de desambiguación.

6 Conclusión

En este estudio se ha propuesto una aproximación a la DVA basada en la información semántica de los argumentos verbales representada en forma de los papeles temáticos que interpretan y la clasificación de dichos argumentos según diversas ontologías. Los buenos resultados obtenidos (90% para el corpus completo empleando papeles temáticos y TCO) corroboran la necesidad de tener en cuenta estas informaciones en futuros sistemas de DVA. De cara a validar estos resultados, y como trabajo futuro, sería necesario realizar de nuevo los diferentes experimentos llevados a cabo en este estudio con un corpus de mayor tamaño; no solo en número de verbos sino también en número de sentidos y ejemplos por verbo. Así mismo, resultaría interesante realizar un contraste con otras fuentes de información como WordNet y, posteriormente, comparar el rendimiento de esta aproximación con el de otras propuestas vigentes.

Agradecimientos

Esta investigación se ha llevado a cabo gracias al proyecto ReTeLe (TIN 2015-68955-REDT).

Bibliografía

Alonso, L., J.A. Capilla, I. Castellón, A. Fernández y G. Vázquez. 2007. The Sensem project: Syntactico-semantic annotation of sentences in Spanish. En *Recent Advances in Natural Language Processing IV. Selected papers from RANLP 2005*, páginas 89-98, John Benjamins Publishing Co, USA.

Álvez, J., J. Atserias, J. Carrera, S. Climent, A. Oliver y G. Rigau. 2008. Consistent annotation of EuroWordNet with the Top Concept Ontology. En *Proceedings of the 4th Global WordNet Conference*, University of Szeged, Hungría.

Buscaldi, D., P. Rosso, F. Pla, E. Segarra y E.S. Arnal. 2006. Verb sense disambiguation using support vector machines: Impact of

WordNet-extracted features. En *Computational Linguistics and Intelligent Text Processing*, páginas 192-195, Springer, USA.

Castellón, I., S. Climent, M. Coll-Florit, M. Lloberes y G. Rigau. 2012. Constitución de un corpus de semántica verbal del español: Metodología de anotación de núcleos argumentales. En *Revista de Lingüística Teórica y Aplicada*, número 50, páginas 13-38, Universidad de Concepción, Chile.

Dang, H.T. y M. Palmer. 2005. The role of semantic roles in disambiguating verb senses. En *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, páginas 42-49, Association for Computational Linguistics, USA.

Del Corro, L., R. Gemulla y G. Weikum. 2014. Werdy: Recognition and disambiguation of verbs and verb phrases with syntactic and semantic pruning. En *Conference on Empirical Methods in Natural Language Processing*, páginas 374-385, Association for Computational Linguistics, USA.

Dligach, D. y M. Palmer. 2008. Novel semantic features for verb sense disambiguation. En *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, páginas 29-32, Association for Computational Linguistics, USA.

Escudero, G. 2006. Machine Learning Techniques for Word Sense Disambiguation (tesis), Universitat Politècnica de Catalunya, España.

Fellbaum, C. 2005. WordNet and Wordnets. En *Encyclopedia of Language and Linguistics*, 2ª edición, páginas 665-670, Elsevier, U.K.

Fernández-Montraveta, A. y G. Vázquez. 2014. The SenSem Corpus: An annotated corpus for Spanish and Catalan with information about aspectuality, modality, polarity and factuality. En *Corpus Linguistics and Linguistic Theory*, volumen 10, páginas 273-288, De Gruyter, Alemania.

Gonzalez-Agirre A., E. Laparra y G. Rigau. 2012. Multilingual Central Repository version 3.0. En *8th international conference on Language Resources and Evaluation*, páginas 2525-2529, Turquía.

Navigli, R. 2009. Word sense disambiguation: A survey. En *ACM Computing Surveys*, volumen 41, ACM, USA.

Padró, L. 2011. Analizadores Multilingües en Freeling. En *Linguamatica*, volumen 3, número 1, páginas 13-20, Portugal.

Pal, A. y D. Saha. 2015. Word sense disambiguation: A survey. En *International Journal of Control Theory and Computer Modeling*, volumen 5, número 3, AIRCC, India.

Pease, A., I. Niles y J. Li. 2002. The Suggested Upper Merged Ontology: A large ontology for the semantic web and its applications. En *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*, volumen 28, páginas 7-10, Association for the Advancement of Artificial Intelligence, USA.

Platt, J. 1998. Fast training of support vector machines using Sequential Minimal Optimization. En *Advances in Kernel Methods - Support Vector Learning*, The MIT Press, USA.

Wagner, W., H. Schmid y S.S. Im Walde. 2009. Verb sense disambiguation using a predicate-argument-clustering model. En *Proceedings of the CogSci Workshop on Distributional Semantics beyond Concrete Concepts*, páginas 23-28, Países Bajos.

Wang, X., W. Zuo y Y. Wang. 2013. A novel approach to word sense disambiguation based on topical and semantic association. En *The Scientific World Journal*, Hindawi Publishing Co, UK.

Witten, I., E. Frank y M.A. Hall. 2011. Data Mining: Practical Machine Learning Tools and Techniques, Elsevier, USA.

Ye, P. y T. Baldwin. 2006. Verb sense disambiguation using selectional preferences extracted with a state-of-the-art semantic role labeler. En *Proceedings of the Australasian Language Technology Workshop*, páginas 139-148, Australasian Language Technology Asociation, Australia.

Zhong, Z. y H. Tou Ng. 2010. It Makes Sense: A wide-coverage word sense disambiguation system for free text. En *48th Annual Meeting of the Association for Computational Linguistics*, páginas 78-83, Association for Computational Linguistics, USA.

Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 29-36 recibido 11-11-2016 revisado 12-01-2017 aceptado 06-02-2017

On Evaluating the Contribution of Text Normalisation Techniques to Sentiment Analysis on Informal Web 2.0 Texts*

Evaluación de la Contribución de la Normalización al Análisis de Sentimiento en Textos Informales de la Web 2.0

Alejandro Mosquera
University of Alicante
Alicante, Spain
amosquera@dlsi.ua.es

Yoan Gutiérrez
University of Matanzas
Matanzas, Cuba
yoan.gutierrez@umcc.cu

Paloma Moreda
University of Alicante
Alicante, Spain
moreda@dlsi.ua.es

Abstract: The writing style used in social media usually contains informal elements that can lower the performance of Natural Language Processing applications. For this reason, text normalisation techniques have drawn a lot of attention recently when dealing with informal content. However, not all the texts present the same level of informality and may not require additional pre-processing steps. Therefore, in this paper we explore the results of applying lexical normalisation applied to a sentiment analysis classification task on Web 2.0 texts, shows more than a 2.6 % improvement over average F1 for the most informal data.
Keywords: Informality, normalisation, sentiment analysis, opinion mining

Resumen: El tipo de lenguage empleado en las redes sociales suele incluir elementos informales que pueden afectar el rendimiento de las herramientas de procesamiento del lenguaje natural. El uso de técnicas de normalización léxica es una de las opciones que se han estado usando a la hora de tratar contenidos de la Web 2.0. Sin embargo, no todos los textos requieren dicho pre-procesamiento ya que pueden exhibir diferentes niveles de informalidad. En este trabajo exploramos el impacto de aplicar normalización léxica evaluando los resultados de un sistema de análisis del sentimiento antes y después de la normalización. Los resultados de nuestra investigación muestran una mejora de mas del 2.6 % sobre el F1 para los textos mas informales.
Palabras clave: Informalidad, normalización, minería de opiniones

1 Introduction

Nowadays, Web 2.0 applications are some of the most popular forms of communication between Internet users. Blogs, social networks or short text messaging platforms have become a very important participation channel where users publish their comments and opinions. This valuable source of information contains insights about user opinions and sentiments regarding almost any topic. These can determine the reputation of public companies or figures, mine opinion patterns and measure the popularity of news and events.

Sentiment analysis (SA) is the sub-field of Natural Language Processing (NLP) that extracts and identifies subjective information. A basic task in SA deals with measuring the overall polarity orientation of a document about some topic. When SA is applied to social media comments it can be used to increase the effectiveness of marketing campaigns, discover new market threats and opportunities or react faster to customer issues.

However, the language used in social media websites and applications can contain a

* This work has been partially funded by the European Commission under the 7th Framework Programme for Research and Technological Development through the SAM (FP7-611312) project, by the Spanish Government through the ATTOS (TIN2012-38536-C03-03) and LEGOLANGUAGE (TIN2012-31224) projects, and by the University of Alicante through the project "Explotación y tratamiento de la información disponible en Internet para la anotación y generación de textos adaptados al usuario" (GRE13-15) and "Tratamiento inteligente de la información para la ayuda a la toma de decisiones" (GRE12-44). We also thank our anonymous reviewers for their insightful comments and valuable feedback

ISSN 1135-5948

variable amount of informal elements such as lexical variants, slang or non-standard punctuation (Thurlow, 2003) that can make any NLP task challenging. For this reason, these texts can benefit from a pre-processing step that understands these informal features and replaces them by their formal equivalent (Wang and Ng, 2013).

The use of lexical normalisation to enhance NLP processing is not a new topic and it has been the subject of recurrent research applied to short and noisy texts such as SMS (Aw et al., 2006). The similarities shared by SMS and more recent genres such as microblogs (Han and Baldwin, 2011) have helped to develop similar approaches. Moreover, not all Web 2.0 genres have the same level of informality. For example, micro-blog posts have a character limit that favors contractions and ellipsis while blog entries or product reviews are usually larger and more elaborated (Santini, 2006).

Because of these genre differences not all the Web 2.0 texts would experience the same benefits after a normalisation step. For this reason, in this paper we analyse the effects of replacing informal lexical variants with their canonical version on social media texts. This has been applied to a SA classification task with aim to enhance polarity detection results. In order to do this, we have carried out several polarity classification experiments using English texts with different degrees of informality and evaluated the impact of normalisation in the results. We have also explored the use of informality analysis as a way of measuring the need of pre-processing on each case.

This article is organised as follows: In Section 2 the state of the art is reviewed. Section 3 describes the informality analysis process. The SA systems used in the experiments are explained in Section 4. In Section 5, the text normalisation step is introduced. The corpora used for all the experiments are detailed in Section 6. Section 7 contains the obtained results and their analysis. Finally, our main conclusions and future work are drawn in Section 8.

2 Related Work

Both industry and academic researchers have increased their interests on measuring user sentiments from social media. After the initial works of Pang, Lee and Vaithyanathan (2002) several applications of opinion mining have been developed focused on microblogs (Barbosa and Feng, 2010; Bifet and Frank, 2010) using both machine learning (Turney, 2002) and lexicon-based approaches (Taboada et al., 2011). The real-time nature of tweets provides a large amount of metadata that can be used as a training corpus for opinion mining systems (Pak and Paroubek, 2010) without requiring annotated corpora (Wiebe, Wilson, and Cardie, 2005).

Whilst normalisation is a common preprocessing step in several areas of NLP (Sproat et al., 2001; Adda et al., 1997) the rise of social media has expanded the concept and meaning of this process. Lexical normalisation techniques (Liu et al., 2011; Han, Cook, and Baldwin, 2013) based on the substitution of out of vocabulary (OOV) words have been used in opinion mining systems before (Mukherjee et al., 2012; Gutiérrez et al., 2013; Sidorov et al., 2013) but this process is usually presented as an intermediate transformation step without explicitly detailing the contribution of normalisation to the classification results. In a more recent analysis of the improvements of using text normalisation applied to SA tasks (Mosquera and Moreda, 2013) it has been shown that normalisation can have positive effects on informal genres. On the other hand, there are different genres within the Web 2.0 and they do not have the same level of informality (Mosquera and Moreda, 2012c), so the enhancements obtained after normalisation can be more or less relevant depending on that level.

Regarding the analysis of the formality/informality of documents most of the prior research tried to measure text formality using readability indexes, and the concept of lexical density (Fang and Cao, 2009). There were attempts to create a formality score by using the frequency of deictic words, that are expected to increase with the informality of a text and, conversely, the frequency of non-deictic words should increase with text formality (Heylighen and Dewaele, 1999). While this score can be used to detect deep formality this approach cannot quantify stylistic or grammatical deviations. Regarding approaches measuring informality, the work of Mosquera and Moreda (2012a) uses multidimensional analysis in order to determine the informality level of Web 2.0 texts. This method not only shows information about

what texts are more informal but it also allows the comparison of texts from other corpora or genres by using a set of dimensions (Mosquera and Moreda, 2012c).

For this reason, in this paper we study the cases where Web 2.0 texts benefit from using normalisation techniques. We apply this analysis to a common NLP task such as SA, and evaluate when this pre-processing step is necessary and can really enhance the classification results. In order to to this, informality analysis is used to score and rank the SA corpora by their informality level before and after the normalisation step.

3 Informality Analysis

Because Web 2.0 texts have specific informal features not usually present in more formal genres we have applied informality analysis using the SMILE (Mosquera and Moreda, 2012b) tool, a framework for classifying texts by their informality level based on four dimensions: Complexity, Emotiveness, Expressiveness and Incorrectness. These dimensions are based on aggregated text features such as the presence of slang and offensive words, incorrect capitalisations and punctuation marks, frequency of character repetitions, readability measures, frequency of emoticons or the frequency of SMS-style contractions.

4 Sentiment Analysis

In order to carry out the SA experiments we have used a 3-class (positive, negative and neutral) unsupervised SA classification system based on WordNet (WN)(Fellbaum, 1998) and additional resources. In order to enrich the WN resource, it has been linked with different lexical resources such as WordNet Domains (WND) (Magnini and Cavaglia, 2000) a lexical resource containing the domains of the synsets in WordNet, SUMO (Niles and Pease, 2003) an ontology relating the concepts in WordNet, WordNet Affect (WNA) an extension of WN where different synsets are annotated with one of the six basic emotions proposed by Ekman (1999), SentiWordNet (Esuli and Sebastiani, 2006) a lexical resource where each synset is annotated with polarity, Semantic Classes (SC)(Izquierdo-Bevià, Suàrez, and Rigau, 2007) a set of Base Level Concepts (BLC) based on WN. RST (Gutiérrez et al., 2010) is a method able to disambiguate the senses of the words contained in a sentence by obtaining the Relevant Semantic Trees from different resources. For SA, RST makes use of the polarity information from SentiWordNet that is contained in ISR-WN (Gutiérrez et al., 2010). In order to measure the association between concepts in each sentence according to a multidimensional perspective, RST uses the Association Ratio (AR) measure (Vázquez, Montoyo, and Kozareva, 2007). The purpose is to include the Multidimensional Semantic Analysis into the Opinion Analysis using RSTs (Gutiérrez, Vázquez, and Montoyo, 2011) with WNDs and SCs.

5 Text Normalisation

We have used TENOR (Mosquera, Lloret, and Moreda, 2012), a multilingual lexical normalisation tool for English and Spanish texts in order to transform noisy and informal words into their canonical form. After this step they can be easily processed by NLP tools and applications.

In order to do this, OOV words are detected with a dictionary lookup. TENOR uses a custom-made lexicon built over the expanded Aspell dictionary and then augmented with domain-specific knowledge from the Spell Checking Oriented Word Lists (SCOWL)[1] package.

The OOV words are matched against a phone lattice using the double metaphone algorithm (Philips, 2000) to obtain a list of substitution candidates. With the Gestalt pattern matching algorithm (Ratcliff and Metzener, 1988) a string similarity score is calculated between the OOV word and its candidate list.

Nevertheless, there are acronyms and abbreviated forms that can not be detected properly with phonetic indexing techniques *(lol - laugh out loud)*. For this reason, TENOR uses an exception dictionary with common Internet abbreviations and slang collected from online sources[2].

Moreover, a number transliteration lookup table and several heuristics such as word-lengthening compression, emoticon translation and simple case restoration are applied to improve the normalisation results. Finally, TENOR uses a trigram language model in order to enhance the clean candidate selection.

[1] http://wordlist.sourceforge.net/

[2] http://en.wiktionary.org

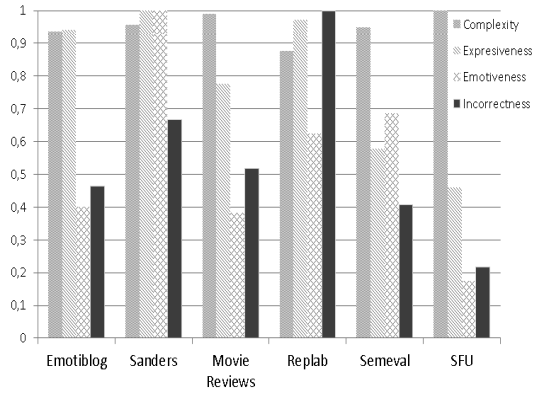

Figure 1: Distribution of informality dimensions in the corpora

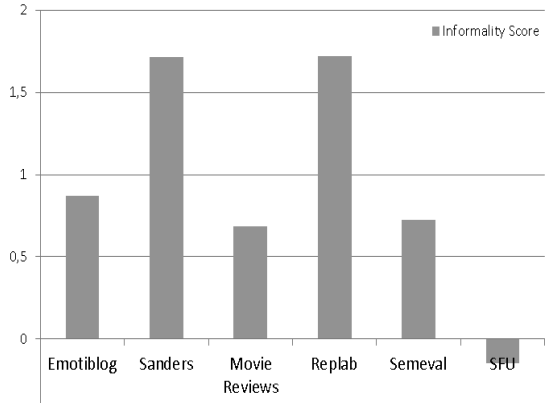

Figure 2: Normalised informality scores

6 Datasets

The polarity classification system has been evaluated using annotated English and Spanish texts from different Web 2.0 genres:

Microblog publications (Sanders): 5513 Twitter messages in English [3].

Blog posts: The Kyoto sub-set of the EmotiBlog corpus [4] corpus comprising 1173 English texts.

Movie reviews: Polarity dataset from movie reviews (Pang and Lee, 2005) containing 10662 sentences.

Microblog publications (RepLab 2013): 70139 polarity-annotated English and Spanish tweets from the RepLab2013 testing dataset (Amigó et al., 2013).

Microblog publications (Semeval 2013): 6434 polarity-annotated English tweets from the Semeval 2013 sentiment analysis training dataset (Nakov et al., 2013).

Online reviews (SFU): The SFU corpus (Taboada and Grieve, 2004) contains 400 online reviews in English for several product categories.

We have obtained the distribution of informality dimensions (see Figure 1) and normalised informality scores for each corpus (see Figure 2) by using informality analysis. These results can be aggregated in three main groups: Very informal (Sanders and RepLab tweets), Informal (Semeval tweets and movie reviews) and Formal (SFU reviews).

[3] http://www.sananalytics.com/lab/twitter-sentiment/sanders-twitter-0.2.zip

[4] http://gplsi.dlsi.ua.es/gplsi11/allresourcespanel

7 Results

A ten-fold cross-validation evaluation of SA classification has been conducted on the previously described linguistic resources before and after the normalisation step using TENOR. The results on Table 1 show how Sanders and Emotiblog texts obtained more than a 4 % and 3.5 % F1 improvement respectively on polarity classification by using the WN-Domain approach. All the F1 scores obtained during the experiments have been checked for statistical significance at 0.95 % confidence level. The aforementioned cases where normalisation contributes the most to SA have a high classification confidence.

Regarding the Semantic Class method, F1 results are in overall lower and they seem to improve after normalisation where texts are very informal (e.g. Sanders, Replab) only. Enhancements in F1 after using TENOR are slightly higher with the WN-domain approach, especially on the corpora with medium informality level. Moreover, if we take into account the average values by informality level (see Table 2) we can appreciate that these differences are just 1.5 percentual points higher for Semantic Class when analysing the most informal texts, obtaining in overall similar results.

To reduce the dependency of the results on the two unsupervised SA approaches, we have also repeated the same experiments using a supervised SA system (Mosquera and Moreda, 2013) with a subset of the original corpora (Sanders, Emotiblog and Movie reviews). Interestingly, the improvements of normalisation on Sanders are considerably higher (6,45 %) while the results on less in-

formal corpora such as Emotiblog show a substantial decrease on F1 (-5,27 %). This supervised SA model performs better with medium-informality corpora and needs a normalisation step in order to obtain comparable scores for the most informal texts.

After the experimental results, we can conclude that text normalisation consistently helps to improve SA classification systems on the most informal Web 2.0 genres and can be useful on some of the less informal texts.

7.1 Discussion

During the evaluation we found several cases of sentences that were correctly labeled without any pre-processing by the SA systems but generated FPs/FNs after being processed by TENOR. We have manually classified these into 4 main categories:

Boosted/Reduced polarity: Normalisation can reduce the polarity of sentences by removing character repetitions. Some interjections can have a higher score when they include repeated characters *Ohhhhh* (disappointment) vs *Oh* (surprise). e.g. Before normalisation: Ohh I think he did and there's so much more which won't be covered now (Negative) And after normalisation: Oh I think he did and there's so much more which won't be covered now (Neutral)

OOV words are usually ignored by dictionary-based SA systems but after normalisation these are now processed. This is usually the desired effect and improves the results. *Hasta luego :) (See you later)* vs *Hasta luego estoy feliz. (See you later I am happy)* It goes from a detected neutral polarity to a high-positive one after the normalisation of the emoticon with the textual equivalent. But in the case of very short sentences it can cause FPs by boosting the polarity when there is not enough context: *Que frio (so cold)* vs *Que frío.* (Neutral to Negative polarity change after normalisation).

Different language: As we have not performed any language detection during the process, the presence on OOV words in another language will result in a poor quality normalisation that will affect negatively the sentiment detection: e.g. *es una version frida superficial , preciosista y sin ningun contenido (a cold and superficial version, without any content)* vs *eyes una viewers and due frida superficial, preciosista why sin knowing and continued.* (Negative polarity before normalisation but Neutral after).

Entity removal: One of the preprocessing steps of TENOR is the removal of all the OOV entities such as URLs, Twitter hashtags and mentions. This eases the processing for NLP tools but sometimes these tags contain semantic information that can be relevant for SA, e.g.: *Office 2014 #Mac #sucks #hate* (Negative) vs *Office 2014* (Neutral after normalisation) *#NWTrue Blood can't wait til the new one comes on tomorrow* (Positive) vs *Blood can't wait tile the new one comes on tomorrow.* (Negative after normalisation)

Incorrect normalisation: We have not found many cases where incorrect normalisation caused a FP or a FN but these may happen when some of the words are incorrectly treated as OOV if they are not present in the normalisation dictionary. One of the limitations of TENOR is the absence of superlatives and diminutives for Spanish as these are not treated as IV when ending with the suffixes -ato or -ito. e.g. *Toma ya! Esto si que es un piropazo!!!!!* (Neutral) vs *Toma ya! Esto si que es un propicio!* (Positive after normalisation)

A more granular analysis of these effects by informality level can be observed in Table 3.

Level	B/R	Lang.	Entity	Badnorm
Informal	7.50	13,16	55,05	24,29
Medium	2.54	7,35	37,13	52.98
Formal	0.00	5,30	1.90	92,80

Table 3: Percentage of misclassification after normalisation by informality type: Boosted/Reduced polarity(B/R), Different language(Lang.), Entity removal(Entity) and Incorrect normalisation(Badnorm)

8 Conclusion

In this paper we have evaluated the use of lexical normalisation techniques with aim to enhance SA classification by replacing informal lexical variants with their canonical version. Our experiments with Web 2.0 English texts consistently show higher average F1 over the original data. However, after using informality analysis we have discovered that these improvements are higher for the most informal texts (2.6 % avg. F1), while in the less informal and formal corpora normalisation only

System	Corpus	Precision	Recall	F1	Diff.	F1 Confidence (+-)
Mosquera-Moreda2013	Sanders	81,20	80,00	80,60		2.020
Mosquera-Moreda2013	Sanders Norm	**85,80**	**85,80**	**85,80**	6,45 %	2.860
Mosquera-Moreda2013	Emotiblog	84,70	86,10	85,40		2.853
Mosquera-Moreda2013	Emotiblog Norm	79,00	82,89	80,90	-5,27 %	2.249
Mosquera-Moreda2013	Reviews	66,70	66,70	66,70		2.861
Mosquera-Moreda2013	Reviews Norm	**67,10**	**67,10**	**67,01**	0,01 %	2.854
WN-Domain	Sanders	59,49	41,95	49,20		0.354
WN-Domain	Sanders Norm	**61,10**	**44,04**	**51,19**	4,04 %	0.352
WN-Domain	Emotiblog	53,46	44,92	48,82		0.354
WN-Domain	Emotiblog Norm	**53,80**	**47,67**	**50,55**	3,54 %	0.353
WN-Domain	Replab	35,42	35,63	35,52		0.894
WN-Domain	Replab Norm	35,34	**36,38**	**35,85**	0,93 %	0.949
WN-Domain	Semeval	37,79	37,89	37,84		0.948
WN-Domain	Semeval Norm	36,55	36,44	36,49	-3,57 %	0.892
WN-Domain	SFU	53,83	49,98	51,80		0.949
WN-Domain	SFU Norm	**54,37**	**50,22**	**52,20**	0,77 %	0.948
WN-Domain	Reviews	55,90	45,10	49,90		1.043
WN-Domain	Reviews Norm	55,70	45,00	49,80	-0,20 %	1.319
WN-Semantic Class	Replab	34,66	34,97	34,80		1.317
WN-Semantic Class	Replab Norm	**34,80**	**35,91**	**35,35**	1,58 %	0.921
WN-Semantic Class	Semeval	36,11	36,55	36,32		1.319
WN-Semantic Class	Semeval Norm	**36,12**	36,55	36,32	0,00 %	1.313
WN-Semantic Class	SFU	52,26	50,48	51,40		1.185
WN-Semantic Class	SFU Norm	**52,60**	**51,01**	**51,80**	0,78 %	1.175
WN-Semantic Class	Sanders	57,41	49,20	52,99		1.176
WN-Semantic Class	Sanders Norm	**58,07**	**51,85**	**54,79**	3,40 %	1.175
WN-Semantic Class	Emotiblog	53,80	53,69	53,75		4.896
WN-Semantic Class	Emotiblog Norm	53,25	53,49	53,37	-0,71 %	4.898
WN-Semantic Class	Reviews	55,00	48,20	51,40		4.895
WN-Semantic Class	Reviews Norm	55,00	**48,50**	**51,50**	0,19 %	4.896

Table 1: Polarity classification results by corpus

Level	DOM	Sem.CLS	DOM.NRM	Sem.CLS.NRM	Diff.DOM	Diff.Sem.CLS
Informal	42,36	43,89	**43,52**	**45,07**	2,74 %	2,68 %
Medium	45,52	47,16	45,62	47,06	0,21 %	-0,20 %
Formal	51,80	51,40	**52,20**	**51,80**	0,77 %	0,78 %

Table 2: Polarity classification results (F1) by informality level using WordNet Domain (DOM) and WordNet Semantic Class (CLS) methods before and after normalisation (NRM)

shows a positive impact in some cases. To understand these possible negative effects introduced by normalisation we conducted a case by case analysis and identified the four main scenarios where normalisation can not only fail to improve but also worsen SA results. Finally, we can conclude that normalisation improves SA techniques and informality analysis can be used to determine which texts could benefit from this pre-processing step.

References

Adda, G., M. Adda-decker, J. luc Gauvain, and L. Lamel. 1997. Text normalization and speech recognition in french. In *Proc. ESCA Eurospeech 1997*, pages 2711–2714.

Amigó, E., J. Carrillo de Albornoz, I. Chugur, A. Corujo, J. Gonzalo, T. Martín, E. Meij, M. de Rijke, and D. Spina. 2013. Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems. In *Proceedings of the Fourth International Conference of the CLEF initiative*, pages 333–352.

Aw, A., M. Zhang, J. Xiao, and J. Su. 2006. A phrase-based statistical model for sms text normalization. *Proceedings of the COLING/ACL*, pages 33–40.

Barbosa, L. and J. Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 36–44, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bifet, A. and E. Frank. 2010. Sentiment knowledge discovery in twitter streaming data. In *Proceedings of the 13th international conference on Discovery science*, DS'10, pages 1–15, Berlin, Heidelberg. Springer-Verlag.

Ekman, P., 1999. *Basic Emotions*, pages 45–60. John Wiley and Sons, Ltd.

Esuli, A. and F. Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2016*, pages 417–422.

Fang, A. C. and J. Cao. 2009. Adjective density as a text formality characteristic

for automatic text classification: A study based on the british national corpus. In O. Kwong, editor, *PACLIC*, pages 130–139. City University of Hong Kong Press.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Gutiérrez, Y., A. González, R. Pérez, J. I. Abreu, A. Fernández Orquín, A. Mosquera, A. Montoyo, R. Muñoz, and F. Cámara. 2013. UMCC_DLSI-(SA): Using a ranking algorithm and informal features to solve Sentiment Analysis in Twitter. *Semeval 2013, Proceedings of the 7th International Workshop on Semantic Evaluations*.

Gutiérrez, Y., A. F. Orquín, A. Montoyo, and S. Vázquez. 2010. Integración de recursos semánticos basados en WordNet. *Procesamiento del Lenguaje Natural*, 45:161–168.

Gutiérrez, Y., S. Vázquez, and A. Montoyo. 2011. Sentiment classification using semantic features extracted from WordNet-based resources. In *Proceedings of the 2Nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA 2011, pages 139–145, Stroudsburg, PA, USA. Association for Computational Linguistics.

Han, B. and T. Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378, Portland, Oregon, USA, June. Association for Computational Linguistics.

Han, B., P. Cook, and T. Baldwin. 2013. Lexical normalization for social media text. *ACM Trans. Intell. Syst. Technol.*, 4(1):5:1–5:27, February.

Heylighen, F. and J.-M. Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants. Technical report, Free University of Brussels.

Izquierdo-Bevià, R., A. Suàrez, and G. Rigau. 2007. A proposal of automatic selection of coarse-grained semantic classes for WSD. *Procesamiento del Lenguaje Natural*, 39:189–196.

Liu, F., F. Weng, B. Wang, and Y. Liu. 2011. Insertion, deletion, or substitution? normalizing text messages without pre-categorization nor supervision. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 71–76, Portland, Oregon, USA, June. Association for Computational Linguistics.

Magnini, B. and G. Cavaglia. 2000. Integrating subject field codes into WordNet. In *LREC*. European Language Resources Association.

Mosquera, A., E. Lloret, and P. Moreda. 2012. Towards facilitating the accessibility of web 2.0 texts through text normalisation. In *Proceedings of the LREC workshop: Natural Language Processing for Improving Textual Accessibility (NLP4ITA) ; Istanbul, Turkey.*, pages 9–14.

Mosquera, A. and P. Moreda. 2012a. A qualitative analysis of informality levels in web 2.0 texts: The facebook case study. In *Proceedings of the LREC workshop: @NLP can u tag #user_generated_content ?!; Istanbul, Turkey.*, pages 23–29.

Mosquera, A. and P. Moreda. 2012b. Smile: An informality classification tool for helping to assess quality and credibility in web 2.0 texts. In *Proceedings of the ICWSM workshop: Real-Time Analysis and Mining of Social Streams (RAMSS)*.

Mosquera, A. and P. Moreda. 2012c. The study of informality as a framework for evaluating the normalisation of web 2.0 texts. In *Proceedings of 17th International conference on Applications of Natural Language Processing to Information Systems (NLDB 2012)*. Springer.

Mosquera, A. and P. Moreda. 2013. Improving web 2.0 opinion mining systems using text normalisation techniques. *Recent Advances in Natural Language Processing (RANLP)*.

Mukherjee, S., A. Malu, B. A.R., and P. Bhattacharyya. 2012. Twisent: a multistage system for analyzing sentiment in twitter. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 2531–2534, New York, NY, USA. ACM.

Nakov, P., Z. Kozareva, A. Ritter, S. Rosenthal, V. Stoyanov, and T. Wilson. 2013. Semeval-2013 task 2: Sentiment analysis in twitter.

Niles, I. and A. Pease. 2003. Mapping WordNet to the SUMO ontology. In *Proceedings of the ieee international knowledge engineering conference*, pages 23–26.

Pak, A. and P. Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Pang, B. and L. Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*.

Pang, B., L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philips, L. 2000. The double metaphone search algorithm. *C/C++ Users Journal*, 18:38–43, June.

Ratcliff, J. W. and D. E. Metzener. 1988. Pattern matching: The gestalt approach. *Dr. Dobb's Journal*, 13(7):46–72, July.

Santini, M. 2006. Web pages, text types, and linguistic features: Some issues. *ICAME Journal*, 30.

Sidorov, G., S. Miranda-Jiménez, F. Viveros-Jiménez, A. Gelbukh, N. Castro-Sánchez, F. Velásquez, I. Díaz-Rangel, S. Suárez-Guerra, A. Treviño, and J. Gordon. 2013. Empirical study of machine learning based approach for opinion mining in tweets. In *Proceedings of the 11th Mexican international conference on Advances in Artificial Intelligence - Volume Part I*, MICAI'12, pages 1–14, Berlin, Heidelberg. Springer-Verlag.

Sproat, R., A. W. Black, S. F. Chen, S. Kumar, M. Ostendorf, and C. Richards. 2001. Normalization of non-standard words. *Computer Speech and Language*, 15(3):287–333.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, 37(2):267–307, June.

Taboada, M. and J. Grieve. 2004. Analyzing appraisal automatically. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, pages 158–161.

Thurlow, C. 2003. Generation txt? the sociolinguistics of young people's text-messaging. *Discourse Analysis Online*, 1(1).

Turney, P. D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vázquez, S., A. Montoyo, and Z. Kozareva. 2007. Word sense disambiguation using extended relevant domains resource. In H. R. Arabnia, M. Q. Yang, and J. Y. Yang, editors, *IC-AI*, pages 823–828. CSREA Press.

Wang, P. and H. T. Ng. 2013. A beam-search decoder for normalization of social media text with application to machine translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 471–481, Atlanta, Georgia, June. Association for Computational Linguistics.

Wiebe, J., T. Wilson, and C. Cardie. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, 39(2-3):165–210.

Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 37-44 recibido 11-11-2016 revisado 12-01-2017 aceptado 07-02-2017

Un detector de la unidad central de un texto basado en técnicas de aprendizaje automático en textos científicos para el euskera*

A Machine Learning based Central Unit Detector for Basque Scientific Texts

Kepa Bengoetxea, Aitziber Atutxa y Mikel Iruskieta
IXA Group. University of the Basque Country
{kepa.bengoetxea,aitziber.atucha,mikel.iruskieta}@ehu.eus

Resumen: En este artículo presentamos el primer detector de la Unidad Central (UC) de resúmenes científicos en euskera basado en técnicas de aprendizaje automático. Después de segmentar el texto en unidades de discurso elementales, la detección de la unidad central es crucial para anotar de forma más fiable la estructura relacional de textos bajo la Teoría de la Estructura Retórica o *Rhetorical Structure Theory* (RST). Además, la unidad central puede ser explotada en diversas tareas como resumen automático, tareas de pregunta y respuesta o análisis del sentimiento. Los resultados obtenidos demuestran que las técnicas de aprendizaje automático superan a las técnicas basadas en reglas a pesar del pequeño tamaño del corpus y de la heterogeneidad de los dominios que éste muestra, dejando todavía lugar para mejoras y desarrollo.
Palabras clave: Unidad central, tópico principal, RST, aprendizaje automático

Abstract: This paper presents an automatic detector of the discourse central unit (CU) in scientific abstracts based on machine learning techniques. After segmenting a text in its elementary discourse units, the detection of the central unit is a crucial step on the way to robustly build discourse trees under the *Rhetorical Structure Theory* (RST). Besides, CU detection may also be useful in automatic summarization, question answering and sentiment analysis tasks. Results show that the CU detection using machine learning techniques for Basque scientific abstracts outperform rule based techniques, even on a small size corpus on different domains. This leads us to think that there is still room for improvement.
Keywords: Central unit, main topic, RST, machine learning

1 Introducción

Saber cuál es el tema principal o la idea global del texto es una tarea relativamente fácil siempre que se domine la lengua; aunque también es cierto que dicha tarea puede complicarse en algunos textos que no exponen la idea principal explícitamente, para conseguir un efecto comunicativo o simplemente porque los textos no están bien redactados.

El tema principal puede ser representado de diferentes formas: *i*) por elementos o palabras clave (desde una única palabra a una lista de palabras), *ii*) por proposiciones u oraciones completas.

Según Iruskieta, Diaz de Ilarraza, y Lersundi (2014) la detección del tema principal o unidad central (UC)[1] es de gran ayuda en la anotación de la estructura retórica, ya que conocer de antemano cúal es la UC permite mejorar el ratio de acuerdo entre anotadores en la Rhetorical Structure Theory (RST) de Mann y Thompson (1988). Teniendo en cuenta esos resultados, pensamos que un analizador discursivo automático podría

* Agradecemos tanto a Kike Fernandez como a Esther Miranda todo el trabajo técnico para poder analizar y visibilizar los resultados de este trabajo. Este trabajo a sido financiado en parte por el siguiente proyecto: TIN2015-65308-C5-1-R (MINECO/FEDER).

[1] La Unidad Central (UC) es un concepto asociado con los árboles de la RST y es la unidad discursiva elemental (UDE) más importante del árbol que tiene la función de ser el principal núcleo del árbol, aunque puede constar de múltiples UDEs en el caso de parataxis.

ISSN 1135-5948

ofrecer resultados más fiables si detectara la unidad central tras la segmentación discursiva automática (Iruskieta y Zapirain, 2015). Además, podría ser utilizado en tareas del Procesamiento del Lenguaje Natural (PLN), aquellas como, resumen automático, análisis del sentimiento o búsqueda de respuestas.

El objetivo de este artículo es construir un detector automático de unidades centrales en textos científicos para el euskera construyendo un clasificador del tipo *Multivariate Bernoulli Naive Bayes*.[2]

Para entrenar y evaluar el detector automático de la unidad central hemos utilizado el corpus[3] *Basque RST Treebank* (Iruskieta et al., 2013), previamente anotado para otros propósitos y tareas (y el único accesible para el euskera).[4]

En el Ejemplo (1) presentamos un texto de ese corpus anotado manualmente: con los segmentos enumerados y la UC en negrita.

(1) [Estomatitis Aftosa Recurrente (I): Epidemiologia, etiopatogenia eta aspektu klinikopatologikoak.]$_1$ ["Estomatitis aftosa recurrente" deritzon patologia, ahoan agertzen den ugarienetako bat da,]$_2$ [tamainu, kokapena eta iraunkortasuna aldakorra izanik.]$_3$ [Honen etiologia eztabaidagarria da.]$_4$ [Ultzera mingarri batzu bezala agertzen da,]$_5$ [hauek periodiki beragertzen dira.]$_6$ [**Lan honetan patologia arrunt honetan ezaugarri epidemiologiko, etiopatogeniko eta klinikopatologiko garrantsitsuenak analizatzen ditugu.**]$_7$ GMB03013[5]

El texto del Ejemplo (1) se ha segmentado en 7 Unidades de Discurso Elementales (UDE)[6] y la unidad central es la última de ellas, la UDE$_7$.

Según Paice (1980) existen algunos indicadores que facilitan detectar automáticamente las ideas principales. Basándonos en esos indicadores y otros que hemos desarrollado en este estudio, la UDE$_7$ muestra los siguientes:

i) *Lan honetan* 'en este trabajo', el nombre *lan* 'trabajo' junto al demostrativo *hau* 'este' junto con el sufijo -*n* (inesivo) de lugar, se refiere al trabajo que el autor presenta en el resumen.

ii) *Garrantzitsuena* 'el más importante', el adjetivo *garrantzitsu* 'importante' y el superlativo −*en*− 'el más' indican que el elemento modificado por el adjetivo está resaltado de alguna forma en la oración.

iii) *Analizatu dugu* 'hemos analizado', el verbo *analizatu* 'analizar' es común para expresar la acción principal que se realiza en trabajos de investigación (Iruskieta, Diaz de Ilarraza, y Lersundi, 2014) y el pronombre adjunto al verbo auxiliar −*gu* 'nosotros', indica que el la acción la han desarrollado los autores del artículo.

Aunque los indicadores[7] por si solos pueden ser ambiguos, ya que pueden utilizarse en otras UDEs que no son unidades centrales, nuestra hipótesis es que podemos detectar la unidad central de resúmenes científicos de una forma aceptable, utilizando adecuadamente todos estos indicadores con técnicas de aprendizaje automático.

En lo que sigue del artículo, explicamos en la Sección 2 los trabajos relacionados en los que nos hemos basado. En la Sección 3 la metodología que hemos empleado para construir el detector de la unidad central. En la Sección 4 presentamos el sistema y en la Sección 5 los resultados obtenidos. Finalmente, exponemos en la Sección 6 las conclusiones y

[2]Los textos utilizados son relativamente complejos teniendo en cuenta la disposición discursiva de la unidad central, ya que la unidad central puede estar en diferentes posiciones en el texto: al principio, en la mitad o al final del texto.

[3]Este corpus puede ser consultado en `http://ixa2.si.ehu.eus/diskurtsoa/`.

[4]Aunque en este trabajo nos hemos basado en la RST, pensamos que la detección de la unidad central podría ser aplicable también en otras teorías.

[5]Texto literalmente traducido: [La Estomatitis Aftosa Recurrente (I): Epidemiología, etiopatogenia y aspectos clínicopatológicos.]$_1$ [La estomatitis aftosa recurrente es una de las patologías orales más frecuentes,]$_2$ [de tamaño, localización y duración variable.]$_3$ [Su etiología es todavía controvertida.]$_4$ [Se caracteriza por la aparición de úlceras dolorosas,]$_5$ [estas recidivan periodicamente.]$_6$ [En este trabajo analizamos las principales características epidemiológicas, etiopatogénicas y clinicopatológicas de esta común patología oral.]$_7$

[6]Las UDEs son los bloques o segmentos más pequeños de los que consta una estructura en árbol discursivo (Carlson, Marcu, y Okurowski, 2001). En general, las UDEs son enunciados independientes o adverbiales.

[7]Otros indicadores en este texto aunque más complejos son: *i*) las palabras o lemas repetidos del título: *epidemiologia* 'epidemiología', *etiopatogenia* y *klinikopatologia* 'clinicopatología', *ii*) los sinónimos como *aspektu* 'aspecto' y *ezaugarri* 'característica', y *iii*) la relación de anafora entre *Estomatitis Aftosa Recurrente* y *patologia arrunt honetan* 'esta patología común'.

el trabajo futuro.

2 Trabajos relacionados

La extración de la unidad más relevante de un texto se ha estudiado con diferentes propósitos y aplicando distintas técnicas. Luhn (1958) hace uso de información estadística sobre una lista de palabras significativas o clave para la extracción de las sentencias más relevantes en resúmenes literarios en inglés. Mientras que Neto et al. (2000) aplica la técnica TF-ISF (*Term Frequency-Inverse Sentence Frequency*) para generar de forma automática resúmenes de textos. En Pardo, Rino, y Nunes (2003) emplean ambas técnicas para extraer la oración más importante de textos científicos tanto en inglés como en portugués de Brasil y obtienen mejores resultados haciendo un ranking de sentencias basado en palabras clave y la posición de la oración.

La unidad central también se puede extraer automáticamente de aquellos analizadores que obtienen la estructura relacional del discurso en forma de árboles jerárquicos. Por ejemplo, se puede extraer del analizador CODRA[8] para el inglés (Joty, Carenini, y Ng, 2015), ya que ésta sería la UDE situada en la raiz del árbol.

Nuestro trabajo es similar al trabajo realizado por Burstein et al. (2001), que emplea un clasificador Bayesiano para identificar la oración temática del texto. El clasificador se sirve como características de la posición, de una lista de palabras clave y ciertas características discursivas basadas en el analizador RST de Marcu (2000). Para extraer la lista de palabras clave, hemos tomado como punto de partida el trabajo de Iruskieta et al. (2015) basado en reglas, para detectar la UC en resúmenes científicos de euskera.

En la sección 5, los resultados del presente experimento en el que se aplican técnicas de aprendizaje automático se comparan con aquellos obtenidos en Iruskieta et al. (2015) a partir de aplicación de reglas.

3 Metodología

3.1 Etapas

Las etapas para desarrollar nuestro detector de UCs basado en técnicas de aprendizaje automático han sido las siguientes:

i. Corpus. Se ha reutilizado el mismo corpus de Iruskieta et al. (2015) que consta de 100 resúmenes científicos en euskera segmentados y con las UCs anotadas manualmente.
ii. Indicadores. Se han utilizado los indicadores de Iruskieta et al. (2015).
iii. Optimización. Se ha elegido y optimizado el algoritmo de aprendizaje automático.
iv. Evaluación. Se ha evaluado el detector automático de UCs.

3.2 El corpus

El corpus sobre el que hemos realizado este estudio está conformado por 100 textos de 5 dominios diferentes (medicina (GMB), terminologia (TERM), ciencia (ZTF), ciencias de la salud (OSA) y de la vida (BIZ)), catalogados por UZEI[9] y la *Udako Euskal Unibertsitatea* (UEU).[10] El corpus de 100 textos contiene 15.168 palabras, cada texto con su unidad central. Presentamos el corpus con mayor detalle en la Tabla 1.

Dominio	Textos	Palabras	UDEs	UCs
GMB	20	2.753	247	29
TERM	20	5.398	523	37
ZTF	20	6.646	548	27
OSA	20	4.964	454	21
BIZ	20	5.407	572	23
Total	**100**	**15.168**	**2.344**	**137**

Tabla 1: Descripción del Corpus

Hemos empleado los dominios GMB, TERM y ZTF para entrenar nuestro sistema y generar el modelo de aprendizaje (incluyendo la selección características y la optimización hiperparamétrica), y los dominios OSA y BIZ para validar los resultados. El corpus de entrenamiento se ha dividido en 10 partes para realizar una validación cruzada. En la Tabla 2 hemos calculado si ambos corpus muestran la misma dificultad en la detección de la unidad central de este modo:

$Dificultad = \frac{UCs}{UDEs}$ cuanto más cerca de 1 es más fácil de determinar la UC.

Corpus	UDEs	UCs	Dificultad
Train	1.318	93	0,07050
Test	1.026	44	0,04288

Tabla 2: Dificultad para elegir la UC

[8] CODRA se puede probar muy facilmente aquí: http://alt.qcri.org/demos/Discourse_Parser_Demo/.

[9] http://www.uzei.eus/.

[10] http://www.ueu.eus/.

Según la información de la Tabla 2 detectar la UC en el corpus de validación (*test*) es más difícil. Los resultados obtenidos en (Iruskieta et al., 2015) también señalan que el resultado fué peor en esa parte del corpus.

El tamaño de este corpus (a nivel de número de textos) es similar al que se ha utilizado en trabajos ya mencionados anteriormente, como el de Paice (1980) con un corpus de 32 textos y el de Burstein et al. (2001) con 100 textos.

3.3 El método de anotación

El corpus fué anotado con la herramienta RSTTool[11] por dos linguistas expertos de RST, en tres fases:

i) Los anotadores segmentaron el texto en UDEs.

ii) Ambos anotadores determinaron cual o cuales de las UDEs formaban la UC.

iii) La anotación de la UC fue evaluada y harmonizada para obtener un *gold standard*.

3.4 Acuerdo entre anotadores

Dos anotadores anotaron manualmente las UDEs y las UCs.[12]

El acuerdo entre el anotador-1 (A1) y el anotador-2 (A2) con el coeficiente Kappa (κ) (Siegel y Castellan, 1988) fue del 0,796 (de un total de 2.344 UDEs). Este grado de acuerdo que está entre los valores del 0,8 κ (acuerdo muy alto) y del 0,6 κ (buen acuerdo) es aceptable, según Krippendorff (2004). También es comparable al acuerdo obtenido en trabajos similares como el de Burstein et al. (2001) con un acuerdo entre dos anotadores de 0,733 κ (de un total de 2.391 oraciones) en un corpus compuesto por 100 textos.[13]

3.5 Extracción de características

El corpus ha sido enriquecido con información morfosintáctica utilizando un analizador morfológico (Aduriz, 2000) y el desambiaguador morfológico (Ezeiza et al., 1998). Se ha creado una lista de palabras clave o significativas para la extracción de la unidad central, una vez que se han analizado las características que mejor indican las UCs en el corpus de entrenamiento. Tomando como referencia el trabajo de Paice (1980), hemos analizado qué verbos, nombres, pronombres y palabras claves (*bonus words*) permiten identificar la UC en nuestro corpus, incluyendo las características que fueran necesarias. Un resumen de las características que se utilizan aprendizaje automático puede verse en la Tabla 3.

Caract.	Descripción
Nombres	Lista de nombres relacionados con la UC
Verbos	Lista de verbos relacionados con la UC
Clave/bonus	Lista de adjetivos y adverbios
Ver. Auxiliares	Lista de verbos con la primera persona del plural
Determinantes	Del tipo *hau* 'este' y *hemen* 'aqui'
Pronombres	Primera persona del plural *gu* 'nosotros'
Combinaciones	Nombres + determinantes, pronombres + nombres y verbos + verbos auxiliares
Verbos principales	Si contiene un verbo principal
Título	Listas de palabras que aparecen en el texto del título
Posición	Posición del segmento en el texto
Posición UDE con verb. aux.	Orden del segmento entre los que incluyen un verbo auxiliar
Condional	Si contiene un verbo condicional
Lista de palabras de parada	Lista de palabras carentes de significado para las UCs

Tabla 3: Características para detectar la UC

3.6 Medidas de evaluación

Para evaluar el detector de la UC, el corpus se ha separado en dos partes. Una parte para el entrenamiento y otra para la prueba final de validación.

Se ha utilizado la misma separación de datos de entrenamiento y validación de Iruskieta et al. (2015) para poder comparar los resultados de ambos trabajos. Los experimentos se han realizado aplicando la técnica de *10-fold cross-validation* sobre los datos de entrenamiento y finalmente se ha evaluado sobre los datos de validación. Para evaluar el sistema se han utilizado las medidas habituales: Exhaustividad (*Recall*), Precisión, y los valores de ambas métricas combinadas en una media armónica denominada valor-F (*F-score* o F_1).

También se ha llevado a cabo un análisis de errores a nivel de texto, para entender como funciona el detector de la UC y ver si hay

[11] http://www.isi.edu/licensed-sw/RSTTool/.

[12] El *gold standard* de estos ficheros pueden ser consultados en http://ixa2.si.ehu.es/diskurtsoa/en/segmentuak.php.

[13] Los desacuerdos más comunes y el proceso de armonización para obtener un gold standard se describen en Iruskieta et al. (2015).

$$log(P(UC|UDE)) = log(P(UC)) + \sum_i \begin{cases} log(P(Ai|UC)/P(A_i)), \\ \text{Si UDE contiene } A_i \\ log(P(\overline{A_i}|UC)/P(\overline{A_i})), \\ \text{Si UDE no contiene } A_i \end{cases}$$

Tabla 4: Fórmula *Bernoulli multivariante*

lugar para mejoras.

4 El detector automático de UCs

Como se ha mencionado previamente, para crear un clasificador que detecte aquellos segmentos de un resumen que tienen mayor probabilidad para ser etiquetados como UC, se ha experimentado con diferentes algoritmos de clasificación como *Multinomial Naive Bayes*, *Multivariate Bernoulli Naive Bayes*, *Support Vector Machines (SVM)* con polinomios de grado 2 y 3, *Radial Basis Functions (RBF)* y *Single Perceptron*, utilizando tanto características basadas en frecuencia como binarias. Finalmente se ha optado por *Multivariate Bernoulli Naive Bayes* por las siguientes razones:

- Los parámetros necesarios para el clasificador se pueden estimar con corpus de entrenamiento pequeños.
- Ha sido utilizado con éxito en tareas similares: para identificar oraciones temáticas (Burstein et al., 2001) o para clasificar textos cortos (McCallum y Nigam, 1998).
- Puede ser empleado tanto como modelo predictivo como descriptivo.
- La aplicación de este clasificador es la que mejores resultados nos ha brindado sobre el corpus de entrenamiento.

La distribución de Bernoulli a la hora de clasificar tiene en cuenta tanto la ausencia como la presencia de las caraterísticas. Para enriquecer el modelo, hemos valido de las características que se muestran en la Tabla 3.

Empleando la fórmula de la Tabla 4, *Bernoulli multivariante*, se obtiene la probabilidad logarítmica que tiene una UDE para pertenecer a la clase UC. El rendimiento mejora si utilizamos el estimador de Laplace para hacer frente a los casos en que las estimaciones de probabilidad de ciertas características que son iguales a cero.

En la fórmula de la Tabla 4: *i*) $P(UC)$ es la probabilidad a priori para que una UDE pertenezca a la clase UC, *ii*) $P(A_i|UC)$ es la probabilidad condicional para que una UDE que pertenece a UC tenga la característica A_i, y *iii*) $P(A_i)$ es la probabilidad a priori para que una UDE contenga la característica A_i, *iv*) $P(\overline{A_i}|UC)$ es probabilidad condicional de que una UDE que pertenece a UC no tenga la característica A_i, y *v*) $P(\overline{A_i})$ es la probabilidad a priori para que una UDE no contenga la característica A_i.

4.1 Elección de un subconjunto de características usando un método Wrapper

Como los algoritmos ingenuos de Bayes sufren con las características reduntantes o correlacionadas, después de seleccionar el algoritmo de aprendizaje con todas las características de entrada, hemos aplicado un wrapper que nos permite seleccionar el mejor subconjunto de características para el clasificador seleccionado.

Para aplicar wrapper necesitamos definir los siguientes criterios:

- Operaciones en el Espacio de Búsqueda. Las operaciones puede ser "añadir característica" o "eliminar característica" o ambas. El término de "selección hacia delante" se refiere a realizar la búsqueda usando el operador "añadir característica", mientras que el término "selección hacia atrás" se refiere a realizar la búsqueda usando el operador "eliminar característica". Mientras que término "*step-wise*" usa ambos operadores. En nuestros experimentos hemos usado únicamente el operador "eliminar característica".
- Estimador de exactitud. Para medir la exactitud de cada operación hemos usado *ten-fold cross-validation* con la función de estimación *F-score*.
- El algoritmo de búsqueda. Para condu-

cir la búsqueda se puede usar diferentes algoritmos. En nuestro experimentos hemos usado el algoritmo de búsqueda *hill-climbing* con la "selección hacia atrás". El algoritmo empieza con todo el conjunto de características y progresivamente elimina una característica y en cada iteración genera sucesores del mejor nodo (aquel que ha obtenido el mayor *F-score*). La condición de terminación será cuando todos los sucesores de la iteracción actual no mejoren el valor de *F-score* de la iteracción anterior.

El wrapper resuelve que el subconjunto óptimo de características que mejor resultado ha obtenido es el siguiente: nombres, verbos, *bonus*, determinantes, pronombres, palabras del título, posición, verbos auxiliares y 3 combinaciones (nombres + determinantes, pronombres + nombres y verbos + verbos auxiliares).

4.2 Post-proceso estadístico

Finalmente, se ha realizado un post-proceso estadístico para los casos en los el clasificador no elija ninguna UDE como UC. En este caso, el post-proceso selecciona el primer candidato más probable de todos ellos, ya que el clasificador nos devuelve un valor de probabilidad para cada UDE.

4.3 Demo para detectar la UC

Una vez realizadas estas tareas, hemos desarrollado una demo, para que pueda ser utilizada por la comunidad científica. De esta forma, la demo pide un texto plano de entrada y ofrece dos formatos de salida diferentes: *i*) Formato web, para utilizar en tareas de PLN. *ii*) Formato RSTTool (RS3), para poder correguir la segmentación o la unidad central y seguir con la tarea manual de la anotación de las relaciones RST en euskera. La demo que puede ser consultada en `http://ixa2.si.ehu.es/CU-detector`.

5 Resultados

En la Tabla 5 se muestran varios resultados: *i*) *Rule Based*. En la primera fila se presenta el mejor resultado registrado en Iruskieta, Antonio, y Labaka (2016) utilizando métodos basados en reglas y aplicando la mejor heurística. *ii*) *ML*. En la segunda fila se pueden ver los resultados obtenidos con el clasificador Bernoulli Naive Bayes utilizando todas las características. *iii*) *ML + Wrap*. En la tercera fila aparecen los resultados obtenidos después de emplear el wrapper, y aplicando el mejor subconjunto de características obtenido. *iv*) *ML + Wrap + Post*. Y finalmente, en la cuarta fila se presentan los resultados después de aplicar el post-proceso estadístico. Obteniendo los mejores resultados en *F-score* de 0,54 con *10-fold cross-validation* y 0,57 con los datos de validación.

Sistema	Datos	Prec.	Rec.	F_1
Rule Based	Dev	0,43	0,51	0,47
	Test	0,70	0,40	0,51
ML	Dev	0,47	0,48	0,48
	Test	0,46	0,54	0,50
ML+Wrap	Dev	0,58	0,46	0,51
	Test	0,46	0,59	0,51
ML+Wrap+Post	Dev	0,56	0,53	**0,54**
	Test	0,48	0,70	**0,57**

Tabla 5: Tabla de resultados

5.1 Análisis de errores

Los diferentes tipos de acuerdos y desacuerdos que hemos observado en el análisis global (texto por texto) de errores que describimos en la Tabla 6 son los siguientes:

- Acuerdo total (coincidencia). El detector solamente ha etiquetado como UC, aquella UDE que se determina como UC en el *gold standard*.
- Acuerdo en UC, pero con falsos candidatos (exceso). Además de las UCs determinadas, el detector ha etiquetado otras UDEs que nos son UCs en el texto.
- Acuerdo parcial en UCs múltiples (falta). El detector ha detectado alguna UC del texto, pero ha dejado otras UCs sin etiquetar.
- Desacuerdo total (desacuerdo). El detector no ha detectado bien ninguna UC del texto.

	Coinc.	Exc.	Falta	Desac.
ML+Wrap	13	13	0	14
ML+Wrap+Post	16	13	2	9

Tabla 6: Análisis de errores

Si comparamos los resultados obtenidos con el método *ML+Wrap* y con el ML+Wrap+post de la Tabla 6, observamos que el postproceso mejora los resultados; ya que, hay mayor número de acuerdos: *i*) hay mayor 'coincidencia' y *ii*) hay mayor número de 'falta', que son acuerdos parciales, ya que

por lo menos una de las UCs ha sido etiquetada adecuadamente.

Hemos podido observar que las causas de los errores cometidos por el sistema en los resultados del post-proceso, son los siguientes:

- 'Exceso'. En 13 ocasiones se ha detectado la UC y otro falso candidato. En 10 ocasiones la primera UC detectada por el sistema es el único válido y en 7 de ellas es la UDE con más indicadores. En las otras 3 ocasiones, el sistema debería decantarse por el segundo candidato detectado con también con más indicadores.
- 'Falta'. En 2 ocasiones se ha detectado una sola UC de las UCs múltiples anotadas manualmente.
- 'Desacuerdo'. En 9 ocasiones el detector no ha sabido establecer correctamente la UC. En 2 ocasiones el texto no cuenta con indicadores suficientes para su detección. En otras 5 ocasiones la unidad central no se presenta como tema principal, sino como una definición o se anuncia mediante una catáfora. En las otras 2 el sistema ha fallado, porque no se han definido alguna otra característica, como por ejemplo la de darle importancia a que algunas caracteríticas estén unas detrás de otras.

Observando estos datos pensamos que hay lugar para mejorar resultados desarrollando técnicas para seleccionar candidatos en el postproceso basándonos en reglas.

6 Conclusiones y trabajo futuro

La mayor aportación de este trabajo es que se ha creado el primer detector de la unidad central (UC) de textos científicos para el euskera, que primero segmenta los textos en UDEs y después determina la UC utilizando únicamente técnicas de aprendizaje automático.[14] La UC se puede extraer del análisis automático que realizan otros analizadores de la RST, como por ejemplo del analizador CODRA, que está entrenado con textos periodísticos en inglés y no para abstracts científicos.

Ahora mismo estamos estudiando si es posible mejorar los resultados obtenidos de las siguientes formas:

- Combinando otras técnicas de aprendizaje automático.
- Combinando diferentes sistemas basados en reglas y en aprendizaje automático.

En el futuro también queremos medir la utilidad de este detector en tareas del PLN y adaptar este detector a otras lenguas y géneros textuales.

- Utilizar en tareas de búsqueda de respuestas (Aldabe et al., 2013) para preguntar sobre el tema principal del texto.
- Aplicar en tareas de análisis del sentimiento en euskera, ya que mejora resultados según Alkorta et al. (2015).
- Adaptar el detector a otras lenguas y evaluarlo con corpus anotados con RST, como pueden ser:
  - La *Spanish RST Treebank* (da Cunha et al., 2011) con 267 textos anotados.
  - La *RST Treebank* en inglés (Carlson, Okurowski, y Marcu, 2002) con 385 textos anotados.

Bibliografía

Aduriz, I. 2000. *EUSMG: morfologiatik sintaxira murriztapen gramatika erabiliz.* Ph.D. tesis, Euskal Herriko Unibertsitatea, UPV/EHU, Donostia.

Aldabe, I., I. Gonzalez-Dios, I. Lopez-Gazpio, I. Madrazo, y M. Maritxalar. 2013. Two approaches to generate questions in basque. *Procesamiento del Lenguaje Natural*, (51):101–108.

Alkorta, J., K. Gojenola, M. Iruskieta, y A. Perez. 2015. Using relational discourse structure information in Basque sentiment analysis. En *5th Workshop RST and Discourse Studies", in Actas del XXXI Congreso de la Sociedad Española del Procesamiento del Lenguaje Natural (SEPLN 2015)*, Alicante.

Burstein, J., D. Marcu, S. Andreyev, y M. Chodorow. 2001. Towards automatic classification of discourse elements in essays. En *Proceedings of the 39th annual Meeting on Association for Computational Linguistics*, páginas 98–105. Association for Computational Linguistics.

[14] Este detector se puede probar en http://ixa2.si.ehu.es/CU-detector.

Carlson, L., D. Marcu, y M. Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. En *2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001*, página 10, Aalborg, Denmark, 1-2 September. Association for Computational Linguistics.

Carlson, L., M. E. Okurowski, y D. Marcu. 2002. *RST discourse treebank*. Linguistic Data Consortium, University of Pennsylvania.

da Cunha, I., J.-M. Torres-Moreno, G. Sierra, L.-A. Cabrera-Diego, y B.-G. Castro-Rolón. 2011. The RST Spanish Treebank On-line Interface. En *International Conference Recent Advances in NLP*, Bulgaria, 12-14 September.

Ezeiza, N., I. Alegria, J.-M. Arriola, R. Urizar, y I. Aduriz. 1998. Combining stochastic and rule-based methods for disambiguation in agglutinative languages. *Proceedings and 17th International Conference on Computational Lingustics*, 1:380–384.

Iruskieta, M., J. Antonio, y G. Labaka. 2016. Detecting the central units in two different genres and languages: a preliminary study of brazilian portuguese and basque texts. *Procesamiento de Lenguaje Natural*, (56):65–72.

Iruskieta, M., M. Aranzabe, A. Diaz de Ilarraza, I. Gonzalez, M. Lersundi, y O. L. de la Calle. 2013. The RST Basque TreeBank: an online search interface to check rhetorical relations. En *4th Workshop "RST and Discourse Studies"*, Brasil, October 21-23.

Iruskieta, M., A. Diaz de Ilarraza, G. Labaka, y M. Lersundi. 2015. The Detection of Central Units in Basque scientific abstracts. En *5th Workshop RST and Discourse Studiesïn Actas del XXXI Congreso de la Sociedad Española del Procesamiento del Lenguaje Natural (SEPLN)*, Alicante.

Iruskieta, M., A. Diaz de Ilarraza, y M. Lersundi. 2014. The annotation of the central unit in rhetorical structure trees: A key step in annotating rhetorical relations. En *COLING*, páginas 466–475, Dublin. Dublin City University and ACL.

Iruskieta, M. y B. Zapirain. 2015. Euseduseg: a dependency-based edu segmentation for basque. *Procesamiento del Lenguaje Natural*, (55):41–48.

Joty, S., G. Carenini, y R. T. Ng. 2015. Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.

Krippendorff, K. 2004. *Content analysis: An introduction to its methodology*. Sage.

Luhn, H. P. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Mann, W. C. y S. A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Marcu, D. 2000. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26(3):395–448.

McCallum, A. y K. Nigam. 1998. A comparison of event models for naive bayes text classification. En *AAAI-98 workshop on learning for text categorization*, volumen 752, páginas 41–48.

Neto, J. L., A. D. Santos, C. A. Kaestner, y A. A. Freitas. 2000. Generating text summaries through the relative importance of topics. *Advances in Artificial Intelligence*, páginas 300–309.

Paice, C. D. 1980. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. En *Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*, páginas 172–191. Butterworth & Co.

Pardo, T., L. Rino, y M. Nunes. 2003. GistSumm: A summarization tool based on a new extractive method. *Computational Processing of the Portuguese Language*, páginas 196–196.

Siegel, S. y N. Castellan. 1988. The Friedman two-way analysis of variance by ranks. *Nonparametric statistics for the behavioral sciences*, páginas 174–184.

Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 45-52 recibido 16-11-2016 revisado 12-01-2017 aceptado 08-02-2017

A Multilingual Multi-domain Data-to-Text Natural Language Generation Approach *

Un enfoque multilingüe y multidominio de datos-a-texto para la generación de lenguaje natural

Cristina Barros and Elena Lloret
Department of Software and Computing Systems, University of Alicante
Apdo. de Correos 99 E-03080, Alicante, Spain
{cbarros,elloret}@dlsi.ua.es

Resumen: La investigación en enfoques multidominio innovadores y flexibles puede ser un paso significativo en el área de Generación del Lenguaje Natural. En este sentido, el objetivo de este artículo es presentar un enfoque estadístico centrado en la fase de realización. Este enfoque permite la generación de oraciones que cumplan un propósito dado por una "característica semilla" de entrada, la cual se encargará de guiar el proceso de generación. Este enfoque ha sido probado en el ámbito de generar automáticamente oraciones que expresan opiniones para reseñas de películas y, además, el enfoque también ha sido probrado en el ámbito de generación del lenguaje para tecnologías de apoyo a problemas relacionados con el lenguaje. Dados los resultados obtenidos, este enfoque es capaz de generar oraciones para dos dominios diferentes con un rendimiento similar en dos idiomas diferentes, obteniendo buenos resultados y cumpliendo los requisitos especificados para cada dominio.
Palabras clave: Generación de lenguaje natural, "característica semilla", modelos de lenguaje factorizados, realización

Abstract: Research in innovative and flexible multi-domain approaches may be a significant step forward in the area of Natural Language Generation. In light of this, the aim of this paper is to present a statistical approach focused on the surface realisation stage. This approach allows the generation of sentences oriented to meet the purpose given by an specific input seed feature, that will guide all the generation process. Our approach was tested to automatically generate opinionated sentences in the domain of movie reviews and was also tested in the domain of Natural Language Generation for assistive technologies. Based on the results obtained, the approach has proved to be able to generate sentences in two different domains with similar performance and for two different languages, obtaining good results and fulfilling the requirements specified for each domain, which opens the door to be applied in new domains and applications.
Keywords: Natural language generation, seed feature, factored language models, surface realisation

1 Introduction

Currently, with the advance of the technology and the increase of the available content, human-computer communication and interaction needs to be as sound, precise and natural as possible (Jacko, 2012).

Much of this information can be given in a non-textual form, being difficult to interpret by humans. Sensor information and data obtained from electronic medical devices or visual numeric ratings and symbols (like stars) with little information about their scoring origins, are clear examples of these kind of information.

For example, in Figure 1, two types of

* This research work has been funded by the University of Alicante, Generalitat Valenciana, Spanish Government and the European Commission through the projects PROMETEOII/2014/001, TIN2015-65100-R and TIN2015-65136-C2-2-R, respectively.

ISSN 1135-5948

movie reviews can be seen, one with only numeric ratings and the other with numeric ratings and some text. These movie reviews differ on the quantity of information given, where a user has more information to make decisions and to know on what basis the movie was scored with the presence of text in the second movie review.

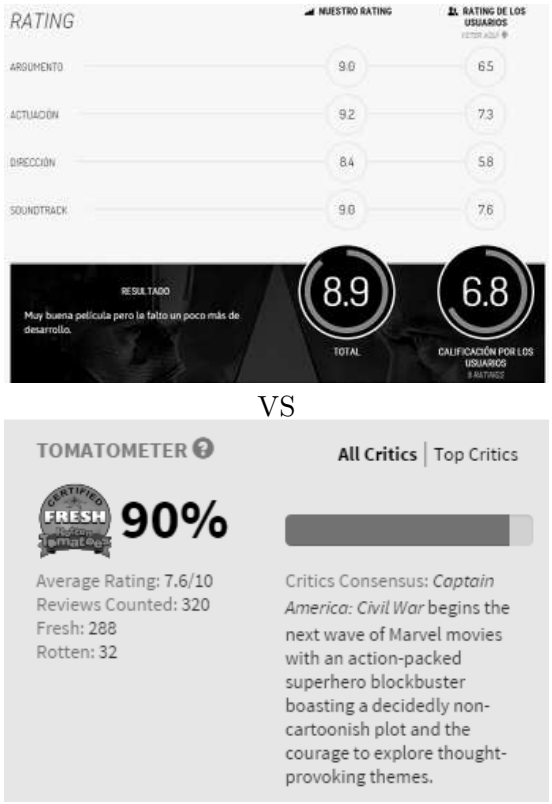

Figure 1: Example of two different types of movie reviews

The area of Natural Language Generation (NLG) aims to automatically develop techniques to produce human utterances, that can be materialised through text or speech. In these terms, NLG techniques can be useful to be used together with non-linguistic elements for generating texts to explain for example the symbols or ratings mentioned above, or another kind of data difficult to interpret such as the one obtained from sensors, among other applications.

In this research area, the development of versatile NLG approaches is still a challenge. Existing NLG systems are designed for very specific domains (Ramos-Soto et al., 2015) and languages (Ballesteros et al., 2015), as well as for particular predefined purposes (Ge et al., 2015), where the cost of adapting these systems can be very high. The research of flexible, multi-domain and multilingual techniques would be a breakthrough in the NLG area.

Towards the advance of such a big challenge, the objective of this paper is to present an almost-fully language independent statistical data-to-text NLG approach that can generate text for different domains, thanks to the concept of an input seed feature which guides all the generation process. Within our scope, this seed feature can be seen as an abstract object (e.g., a rating, a sentiment, a polarity, a phoneme) that will determine how the final sentence will be in relation to its vocabulary or the word categories that this new sentence must contain. We tested our approach in the context of two different domains, that will be explained in section 4, for the English and Spanish languages, in order to show its appropriateness to different non-related scenarios.

2 Related Work

The task of NLG comprises a wide range of subtasks which extend from an action planning until its execution. Therefore, starting from non-linguistic data or text, there are many decisions to be made such as the structure of the message and its content, the rhetorical structure at several levels, the syntactic structure and the correct words choice or the final text arrangement (Bateman and Zoch, 2003). These subtasks can be grouped into a pipeline of three broad stages: document planning, microplanning and surface realisation (Reiter and Dale, 2000). In the document planning stage, the system must decide what information should be included in the text and how to organise it into a coherent structure, leading to a document plan. From this document plan, in the microplanning stage, a discourse plan will be generated, where appropriated words and references will be chosen supplying them with a linguistic structure. Finally, the surface realisation stage generates the final text with the concrete information and structure selected.

This NLG process is commonly addressed from either statistical and knowledge based approaches, where the former are based on the calculus of the probability of certain words to appear together; and the latter resort to linguistic techniques in order to generate text. The main difference between these two approaches is that statistical approaches

are more flexible than knowledge based ones in terms of language and domain.

Traditionally, statistical approaches have been based on Language Models (LM), whose probabilities are extracted from a text. Due to this, these approaches are highly adaptable to different domains and languages. Factored language models (FLM) are an extension of LM proposed in Bilmes and Kirchhoff (2003) which permit a greater flexibility and adaptability. In this model, a word is viewed as a vector of k factors such that $w \equiv \{f^1, f^2, \ldots, f^K\}$, where these factors can be anything, including the Part-Of-Speech (POS) tag, stem or any other lexical, syntactic or semantic feature. Once a set of factors is selected, the main objective of a FLM is to create a statistical model $P(f|f_1, \ldots, f_N)$ where the prediction of a feature f is based on N parents $\{f_1, \ldots, f_N\}$. These models have been widely employed in several areas of Computational Linguistics, mainly in machine translation (Crego and Yvon, 2010). Furthermore they have been used to a lesser degree in NLG, such in the BAGEL system (Mairesse and Young, 2014), where FLM are used to predict the semantic structure of the sentence to generate, or in Novais and Paraboni (2012) where FLM are used to rank sentences in Portuguese.

Moreover, there are several approaches focused on the generation of reviews such as the one presented in Gerani et al. (2014), where an abstractive summarisation for products reviews is generated taking advance of their discourse structure. However, to the best of our knowledge, there is no previous research work focused on generating opinionated sentences employing FLM, and, furthermore, with the restriction of having words related with a concrete seed input features (a specific polarity in our case). In addition, our approach is also novel in the sense that it can be applied to different domains and language with minimal adaption.

3 A Flexible Multi-Domain Natural Language Generation Approach

We propose a statistical approach focused on the surface realisation stage and based on over-generation and ranking techniques employing FLMs.

This technique allows the approach to be almost-fully language independent since it is necessary to adapt some resources (e.g., semantic features) for the language-specific part. This input seed feature concept introduced will permit us to make the generated text flexible regarding its domain and purpose.

This approach first generates several sentences which then will be ranked as will be explained below.

3.1 Generation

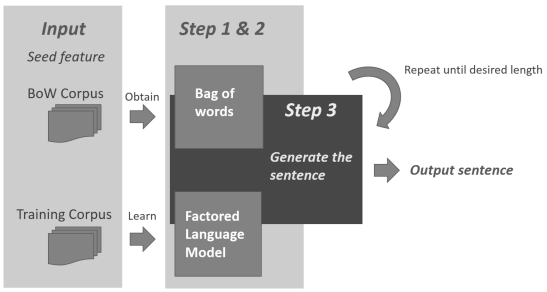

Figure 2: Our proposed approach

For a specific input seed feature (e.g., "positive" polarity), multiple sentences are generated, taking into account: i) a training corpus, ii) a corpus from where a bag of words is obtained (BoW corpus), and iii) the seed feature. The generation approach consists of three major steps, as can be seen in Figure 2:

1. **Step 1: Generate the language model**. A FLM is firstly trained over a corpus (i.e., the training corpus, a collection of texts from where the FLM is trained) in order to obtain the probabilities of the factors of appearing together.
2. **Step 2: Generate the bag of words**. A bag of words containing words related with the input seed feature and their frequency is obtained from the BoW corpus (i.e., a different collection of texts from where the bag of words is gathered). For instance, in the case that we want to generate a sentence with positive polarity, the bag of words could include words such as "great", "good", "outstanding", "excellent", etc.
3. **Step 3: Generate the sentence**. Then, a sentence is generated based on the FLM and the bag of words previously obtained. The generation algorithm follows an iterative process that will finish when the desired length of

the sentence or a full stop are reached. This will allow us to decide the length of the sentence depending on the final application (e.g., a tweet or a sentence to be integrated in a long review) In this iterative generation process, starting in the first iteration from the token start of the sentence, the following words are selected according to the highest probabilities from the FLM, prioritising the selection of words from the bag of words. In this manner, the process guarantee that the generated sentence will contain the maximum number of words related with the input seed feature.

3.2 Ranking

When several sentences are generated for a specific seed feature, the aim of this stage is to decide which one would be finally selected. Only one sentence is selected during this stage. The ranking is performed in order to select one correct sentence based on its probability and the number of words related with the seed feature. In the case that the ranking was not applied, several sentences would be generated, and the user will have to manually select the one that s/he prefers. The sentence probability is computed by the chain rule where the probability of a sentence can be calculated as the product of the probability of all the words: $P(w_1, w_2...w_n) = \prod_{i=1}^{n} P(w_i|w_1, w_2...w_{i-1})$.

The probability of a word is then calculated, as it is suggested in Isard, Brockmann, and Oberlander (2006), such the linear combination of FLMs, where a weight λ_i was assigned for each of them: $P(f_i|f_{i-2}^{i-1}) = \lambda_1 P_1(f_i|f_{i-2}^{i-1})^{1/n} + \cdots + \lambda_n P_n(f_i|f_{i-2}^{i-1})^{1/n}$, where f the selected factors from the different FLMs employed, being the total sum of the weights 1. The final selected sentence would be the one containing the maximum number of words related to the seed feature and which probability is above the average.

4 Domains

We primarily focused our experiments in the domain of generation sentences with the positive and negative polarity in the context of movie reviews. Our final application would be to provide supporting sentences to visual or numeric ratings, so that reviews could be complemented with more information, thus becoming more informative. Furthermore, in order to verify the flexibility and multi-domain of this approach, we also tested this approach in the context of NLG or assistive technologies as it will be explained in section 4.2.

4.1 Opinionated NLG

Within our first domain, the experimentation was focused on the generation of opinionated sentences with a specific polarity (positive or negative), using this polarity as the input seed feature. The main objective is to create meaningful sentences containing words with a specific polarity.

A large portion of the web is dedicated to sites where people express their opinions (such as TripAdvisor[1] or RottenTomataoes[2]), so the generation of this kind of polarity sentences could serve this type of platforms to generate sentences from visual numeric ratings (like stars). So, in a first instance, we focused the generation on the context of movie reviews, where an illustrative example of the sentences we want to generate can be seen in Figure 3. The generation of this kind of sentences can be very useful when an user uses Webpages as the one shown in section 1, where in the review there are only symbols or numbers without any type of explicative or informative associated text.

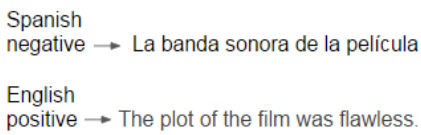

Figure 3: Illustrative example of opinionated NLG sentences (*Translation: The film's soundtrack was awful*)

Given the context seen above, we have employed the Spanish Movie Reviews corpus[3] and the Sentiment Polarity Dataset (Pang and Lee, 2004) as our corpora for Spanish and English, respectively. The approach was tested with the positive and negative polarities using the ML-SentiCon (Cruz et al., 2014) files and the polarity words from (Liu, Hu, and Cheng, 2005) to identify the polarity of a word in Spanish and English, respectively.

[1] https://www.tripadvisor.es/
[2] https://www.rottentomatoes.com/
[3] http://www.lsi.us.es/ fermin/corpusCine.zip

4.2 NLG for assistive technologies

This was a completely different scenario that was used to test and verify the flexibility of our proposed NLG approach. Within our second domain, the experimentation was focused on story generation to help children with dyslalia, a disorder in phoneme articulation. Based on this domain, a phoneme is selected as the seed feature, where the main objective is to generate meaningful sentences containing the maximum number of words in the sentences related to that concrete phoneme. This type of sentences can be useful in dyslalia speech therapies in order to reinforce the phoneme pronunciation through reading and repeating words (Rvachew, Rafaat, and Martin, 1999).

Some illustrative examples of an input phoneme and generated sentences meeting the requirements can be seen in Figure 4.

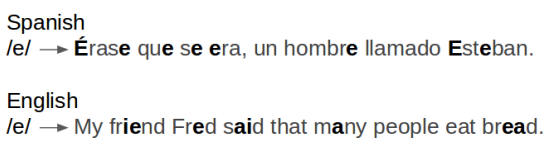

Figure 4: Illustrative example of NLG sentences for assistive technologies (*Translation: Once upon a time, a man named Esteban*)

Therefore, for this scenario, the employed approach is the same as in the first domain specified in section 4.1, where the only difference lies in the seed feature (in this domain, a phoneme) and the corpus used. Consequently, a collection of 158 Hans Christian Andersen tales in two languages (English and Spanish) was chosen as corpora, being the vocabulary contained in it suitable for a young audience. In addition, the approach was tested with all the English and Spanish phonemes.

5 Experiments

With the domains previously mentioned, we conducted an experiment where, using phonemes and polarity as the seed feature in each domain, we automatically generated sentences in Spanish and English. From these generated sentences, a raking was performed over those ended by a full stop according to the linear combination explained in section 3.2. With this experimentation, we wanted to test to what extent the generated sentences were classified as positive and negative (in the case of the first domain) and if they have words with an specific phoneme (in the case of the second domain).

During the experimentation, we used several tools that will be further explained.

Each file of the corpus previously described was processed with Freeling (Padró and Stanilovsky, 2012) in order to obtain information about the selected factors of the FLM. In our case, these factors were the word itself (token), the POS-tag and the lemma. Freeling is a language analyser at a lexical, syntactic and semantic level that works for multiple languages, including English and Spanish.

In order to evaluate the polarity of the sentences, we employed the sentiment analysis classifier described in Fernández et al. (2013).

Finally, we trained the FLM with SRILM (Stolcke, 2002), a software which allows building and applying statistical language models, which also includes an implementation of factored language models.

Taking into account the different factors, Spanish and English sentences were automatically generated using trigram FLM with LEMMA+POSTAG (which proved to works better than other configurations), and subsequently these sentences were ranked with a linear combination of three FLM, as explained in section 3: $P(w_i) = \lambda_1 P(f_i|f_{i-2}, f_{i-1}) + \lambda_2 P(f_i|p_{i-2}, p_{i-1}) + \lambda_3 P(p_i|f_{i-2}, f_{i-1})$, where f can be can be either a lemma and a word, p refers to a POS tag, and λ_i are set $\lambda_1 = 0.25$, $\lambda_2 = 0.25$ and $\lambda_3 = 0.5$. These values were empirically determined.

6 Evaluation and Results

The evaluation of NLG approaches are difficult since there is not a an unique good output (gold-standard) as in other Computational Linguistic fields. In addition,there is no automatic manner to discern the meaningfulness of a given generated text or sentence in an automatic manner. In view of the above, the manual evaluation is the most currently type of assessment used in NLG (Resnik and Lin, 2010). On this basis, we performed a manual evaluation of the generated sentences in order to verify the meaningfulness of the automatic generated sentences.

This manual evaluation was performed by three different evaluators considering a sentence meaningful when: i) the sentence

Surface Realisation Domain		Meaningful generated sentences	Newly meaningful sent. (not in corpus)	Meaningful sent. with seed features
EN	Opinionated sentences for reviews	100%	50%	50%
	Assistive technologies	95%	70%	82.5%
ES	Opinionated sentences for reviews	100%	100%	100%
	Assistive technologies	88.89%	40.74%	88.89%

Table 1: Comparative table of the two domains

is meaningful by itself, ii) the sentence becomes meaningful by adding some punctuation marks, and iii) the sentence becomes meaningful by adding a preposition that usually follows the main verb. In order to measure the agreement between the evaluators the kappa statistic (Randolph, 2008) was employed, obtaining a very good agreement in both domains (an overall agreement of 1 for the opinionated NLG domain in both, English and Spanish; and an overall agreement of 0.83 for the assistive technologies domain in English and an overall agreement of 0.78 in Spanish).

On the other hand, these sentences were automatic evaluated to discern if they met the objective for each domain. This was carried out evaluating the polarity of the sentences with the sentiment analysis classifier mentioned before, in the first domain; and calculating the percentage of words containing the phonemes regarding the total length of the sentence in the second domain.

Table 1 shows the results of the approach once the whole NLG approach is applied (over-generation and ranking), where multiple sentences were generated for a concrete seed feature and subsequently ranked in order to obtain only one sentence for that seed feature. The statistics of the table were calculated based on the total number of selected sentences once the ranking was employed (being the maximum of sentences to generate the two polarities, one sentence for the positive polarity and one for the negative in the first domain; and the total number of phonemes in each language, being 44 phonemes for English and 27 phonemes for Spanish, in the second domain).

As it can be seen in the table, good results were obtained in the meaningful generated sentences in both domains, being almost the half of them not explicitly included in the corpus. Furthermore, we also obtained good results on those meaningful sentences containing words related with the seed feature, fulfilling the characteristics specified in Section 4 in both domains. We checked this using the sentiment analysis classifier mentioned in section 5 in the first domain, where we found that the polarity obtained for the generated sentence was the right polarity we specified as input. In the second we performed a manual evaluation of the words with the phonemes, where the sentences contained an average of 3 words out of 8, that was the average length of the sentences, with the specific phoneme in both languages.

Examples of the generated sentences in English and Spanish are shown in Figure 5.

Opinionated NLG
Polarity: Positive **Sent:** *The good work in this respect.*
Polarity: Negative **Sent:** *The acting be horrible .*
Polarity: Negative **Sent:** *Su falta de imaginación. Trans: Their lack of imagination.*

NLG for assistive technologies
Phoneme: /m/ **Sent:** *My mother be asleep.*
Phoneme: /b/ **Sent:** *I be bear in the book of fairy tale.*
Phoneme: /k/ **Sent:** *Cantar el canción popular. Trans: Sing the popular song.*

Figure 5: Example generated sentences

In view of these results, this approach obtains similar performance in the generation of sentences in both domains for English and Spanish, where the flexibility of the proposed method is demonstrated. The main problem of our approach though is that, due to the use of lemmas as factors, the words in the generated sentences are not inflected, and, in some cases, this affects the readability of the sentence. We are investigating possible methods

to tackle this issue, such as the definition of rules, or the definition of a model to automatically learn the inflections.

7 Conclusions and Future work

In this research work, a multilingual and multi-domain statistical NLG approach which relies on an input seed feature to generate a sentence was presented. This approach allows the generation of sentences oriented to meet the purpose given by an specific seed input feature. This approach was first tested for generating opinionated sentences, where the input seed feature was the desired polarity of the sentence. This type of sentences may be useful for reviews generation based on ratings to support and provide evidences for the numeric values or symbols. Then, to verify that the same approach could be applied to other domains and scenarios, it was applied to the generation of sentences that can be useful in several speech therapies, having a phoneme as the seed feature.

Through the experimentation conducted, the approach has proved to be able to generate sentences in two different domains with similar performance and for two different languages, obtaining good results and fulfilling the requirements specified for each domain.

Although the obtained results are good, we need to add more syntactic and semantic information in order to guarantee the generation of meaningful sentences in all the cases. Consequently, in the future we will study different factors to be included in the FLM, and also, we will analyse to what extent the inclusion of deep learning techniques or word embedding-based method may be beneficial to the approach.

In the short term, we would like to improve the readability of the sentences, a well as to widen and conduct a more exhaustive evaluation of the generated sentences using crowdsourcing platforms.

Furthermore, there are three issues to be improved and research as the next steps. As mentioned before, the inflection of the words of the generated sentences is one the issues to be further investigated. In this respect, we first need to research in the types of transformations that can be applied to the words. For example, we could employ dictionaries containing the inflections and variants of the words,which could be combined with some kind of grammar or structure in order to finally obtain a infected sentence. An example of how could be the inflections of the example generated sentences seen above is shown in Figure 6.

Opinionated NLG
Original Sent: *The good work in this respect.*
Inflected Sent: *The good work was done in this respect.*

NLG for assistive technologies
Original Sent: *My mother be asleep.*
Inflected Sent: *My mother was asleep.*

Figure 6: Example inflections of the generated sentences

On the other hand, another issue that can be further researched is the generation of several sentences with cohesion between them in order to build a larger text. This sentences would need to have related topics to ensure the text coherence. This goal could be achieved, in a first approach, by including in the sentences the same subject or by taking as the subject of the sentence the direct object of the previous sentence.

In addition, for the story generation domain, it could be interesting if the seed feature could be composed, for example phoneme+polarity or phoneme+sentiment in order to generate stories with sentiments.In this case, we would need to adapt the input of the approach to have multiple seed features. For example, if the text to be generated has to help people feeling depressed, it would be necessary to generate an optimistic and happy text, so in this context, the seed feature would be for instance the concept *optimistic*, and, the words selected during the generation process would be related with this concept (e.g. the words *cheerful*, *joy* or *favorable* are related with the optimistic concept). In order to obtain the words related with a concept, lexicons or synsets such as WordNet-Affect (Strapparava and Valitutti, 2004) or word embedding techniques could be employed.

References

Ballesteros, M., B. Bohnet, S. Mille, and L. Wanner. 2015. Data-driven sentence generation with non-isomorphic trees. In *Proceedings of the NAACL*, pages 387–397. Association for Computational Linguistics.

Bateman, J. and M. Zoch. 2003. *Natural Language Generation*. Oxford University Press.

Bilmes, J. A. and K. Kirchhoff. 2003. Factored language models and generalized parallel backoff. In *Proceedings of the Human Language Technology Conference (HLT/NAACL)*, pages 4–6. Association for Computational Linguistics.

Crego, J. M. and F. Yvon. 2010. Factored bilingual n-gram language models for statistical machine translation. *Machine Translation*, 24(2):159–175.

Cruz, F. L., J. A. Troyano, B. Pontes, and F. J. Ortega. 2014. Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, 41(13):5984 – 5994.

Fernández, J., Y. Gutiérrez, J. M. Gómez, P. Martínez-Barco, A. Montoyo, and R. Muñoz. 2013. Sentiment analysis of spanish tweets using a ranking algorithm and skipgrams. *Proc. of the TASS workshop at XXIX Conference of the Spanish Society for Natural Language Processing (SEPLN 2013)*, pages 133–142.

Ge, T., W. Pei, H. Ji, S. Li, B. Chang, and Z. Sui. 2015. Bring you to the past: Automatic generation of topically relevant event chronicles. In *ACL*, pages 575–585. Association for Computational Linguistics.

Gerani, S., Y. Mehdad, G. Carenini, R. T. Ng, and B. Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1602–1613. Association for Computational Linguistics.

Isard, A., C. Brockmann, and J. Oberlander. 2006. Individuality and alignment in generated dialogues. In *Proceedings of the INLG*, pages 25–32. Association for Computational Linguistics.

Jacko, J. A. 2012. *Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications, Third Edition*. CRC Press, Inc.

Liu, B., M. Hu, and J. Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM.

Mairesse, F. and S. Young. 2014. Stochastic language generation in dialogue using factored language models. *Comput. Linguist.*, 40(4):763–799.

Novais, E. M. and I. Paraboni. 2012. Portuguese text generation using factored language models. *Journal of the Brazilian Computer Society*, 19(2):135–146.

Padró, L. and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the 8ht International Conference on Language Resources and Evaluation*. European Language Resources Association.

Pang, B. and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, pages 271–278. Association for Computational Linguistics.

Ramos-Soto, A., A. J. Bugarín, S. Barro, and J. Taboada. 2015. Linguistic descriptions for automatic generation of textual short-term weather forecasts on real prediction data. *IEEE Transactions on Fuzzy Systems*, 23(1):44–57.

Randolph, J. J. 2008. Online kappa calculator [computer software]. retrieved from http://justus.randolph.name/kappa.

Reiter, E. and R. Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

Resnik, P. and J. Lin, 2010. *Evaluation of NLP Systems*, pages 271–295. Wiley-Blackwell.

Rvachew, S., S. Rafaat, and M. Martin. 1999. Stimulability, speech perception skills, and the treatment of phonological disorders. *American Journal of Speech-Language Pathology*, 8(1):33–43.

Stolcke, A. 2002. Srilm - an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing, vol 2.*, pages 901–904.

Strapparava, C. and A. Valitutti. 2004. Wordnet affect: an affective extension of wordnet. In *LREC*, volume 4, pages 1083–1086.

Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 53-60 recibido 18-11-2016 revisado 12-01-2017 aceptado 10-02-2017

Propuesta y desarrollo de una aproximación de generación de resúmenes abstractivos multigénero

Definition and development of a multi-genre abstractive text summarisation approach

Alberto Esteban, Elena Lloret
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Apdo. de Correos 99
E-03080, Alicante, Spain
{aesteban,elloret}@dlsi.ua.es

Resumen: En este trabajo se propone el análisis de técnicas adecuadas para el diseño y desarrollo de un enfoque de generación de resúmenes multigénero, tomando como partida distintas fuentes de datos pertenecientes a distintos géneros textuales. El objetivo principal es combinar todos estos géneros y producir un resumen abstractivo, es decir un nuevo texto coherente que capte las ideas fundamentales sobre un tema recogidas en las fuentes de datos originales. Concretamente, para este trabajo hemos utilizado información de reseñas mediante TripAdvisor y de microblogs, mediante Twitter, debido a que los resúmenes generados se han aplicado en un contexto turístico, para proporcionar a los usuarios los aspectos más y/o menos favorables sobre hoteles y restaurantes. La evaluación del método propuesto a través de usuarios reales indica que los resúmenes generados tienen una calidad suficiente y aceptable para ser utilizados en aplicaciones reales.
Palabras clave: Procesamiento de lenguaje natural, generación de resúmenes, multigénero, resúmenes abstractivos, turismo

Abstract: In this paper, the design and development of a multi-genre abstractive summarisation approach is proposed, taking into account information sources belonging to different textual genres. The main objective is to combine the information in all of these genres and produce an abstractive summary, that is, a new coherent text that captures the main ideas about a topic. Specifically, in this research work, information from reviews extracted from TripAdvisor and microblogs gathered via Twitter was used, since the generated summaries were applied to the tourism sector, to provide users with the most and/or less favourable aspects concerning hotels and restaurants. The evaluation carried out with real users shows that the quality of generated summaries meets the standards and therefore, the summaries can be used in real applications.
Keywords: Natural language processing, text summarisation, multi-genre, abstractive summarisation, tourism

1 Introducción

El gran volumen de datos con el que contamos hoy en día, es evidente en cualquiera de nuestras acciones en Internet. Tanto si buscamos información sobre el último dispositivo móvil puesto a la venta, como si la búsqueda es para encontrar una receta de cocina nos encontraremos con miles de resultados donde elegir. Concretamente, en el dominio turístico, el auge de Internet y las Tecnologías de la Información han hecho que la forma de buscar un hotel o un restaurante cambie de manera drástica. El método tradicional era consultar en una agencia de viajes o a nuestros familiares si conocían un buen hotel o restaurante en el destino de nuestras próximas vacaciones. En la actualidad esa tarea se ha simplificado mucho, podemos visitar varias páginas especializadas para tener toda la información a nuestro alcance. Sin embargo, nos enfrentamos a un problema diferente al de antaño, y es que si bien es

ISSN 1135-5948

cierto que antes teníamos menos información de la deseada, ahora es justo lo contrario; existe una ingente cantidad de información la cual necesita de mucho tiempo para ser analizada. Por ejemplo, si se consulta una página especializada para conocer qué hotel podríamos escoger para hospedarnos en Alicante, tendríamos más de 2500 reseñas solamente para el *Hotel Meliá Alicante* y considerando únicamente una fuente de información.

En la planificación de unas vacaciones o una estancia fuera de casa siempre queremos tomar una buena decisión en la elección del hotel donde nos hospedaremos y/o del restaurante que visitaremos. En la práctica, esto conlleva la lectura de las opiniones de los antiguos clientes y como podemos apreciar, el volumen de opiniones hace que no podamos leerlas todas, dado que es inviable en términos de tiempo y coste la lectura de todas y cada una de las opiniones, siendo difícil recordar cada uno de los aspectos citados en cada opinión. Además, en muchas ocasiones cuando visitamos un portal de reservas de hoteles o restaurantes se muestra un breve resumen o gráfica de las características del establecimiento, pero este resumen o gráfica puede presentar un alto nivel de subjetividad para enfatizar las mejores características, lo que impediría a los usuarios obtener información objetiva sobre el mismo.

En base a este contexto, el objetivo principal de este trabajo es el análisis, propuesta y desarrollo de un método para la generación automática de resúmenes abstractivos multigénero. Mediante el enfoque propuesto, los usuarios podrán conocer de manera concisa lo mejor y lo peor de cada establecimiento en base a los comentarios realizados por los propios clientes en distintas fuentes de información (concretamente en reseñas – TripAdvisor y en microblogs – Twitter), evitando así la falta de objetividad que podría tener leer un resumen realizado por el propio establecimiento para así tomar la mejor decisión de donde alojarse o que restaurante visitar.

2 Trabajo relacionado

Generar un resumen de forma automática no es una tarea trivial y la mayoría de métodos existentes se centran en generar resúmenes extractivos, en los que el sistema simplemente determina la información relevante y la extrae (Lloret, 2012). De ahí que uno de los principales retos sea la generación de resúmenes abstractivos donde se genera un nuevo texto coherente a partir de la información relevante, pero también combinando esta información o añadiendo otra nueva no explícitamente presente en la fuente de origen (Saggion, 2011; Carenini, Cheung, y Pauls, 2013; Li, 2015; Khan, 2015). Si a esto le sumamos la heterogeneidad de géneros textuales (sobre todo los que han aparecido en la Web 2.0) y la diversidad de información, nos encontramos ante otro reto mayor y necesario, que es el de abordar la generación de resúmenes multigénero, hasta ahora apenas tratado por la comunidad científica (Lloret y Boldrini, 2015).

De manera similar a los métodos de resúmenes que trabajan con textos de la Web 2.0, nuestro enfoque utiliza un motor de análisis de opiniones para poder distinguir información con polaridad positiva, negativa y neutral, así como técnicas de clasificación para agrupar la información y heurísticas para determinar la relevancia de la misma. Sin embargo, la principal diferencia de nuestro enfoque radica en la generación del resumen a partir de varias fuentes de información heterogénea, puesto que cada género textual tiene sus propias peculiaridades. Por tanto, se requiere de un proceso de tratamiento y comprensión de la información que permita maximizar la coherencia del resumen final, siendo esencial la generación de resúmenes abstractivos en este contexto.

3 Fases para la generación de resúmenes abstractivos multigénero

A continuación, se van a detallar cada una de las fases involucradas en el proceso de generación de resúmenes. Como entrada, se dispone de un conjunto de documentos procedentes de varias fuentes de información heterogéneas sobre hoteles y restaurantes (en nuestro caso reseñas generadas por usuarios y tuits) sobre el mismo tema. Como salida, se obtiene un breve resumen abstractivo.

3.1 Preparación de los textos

Para esta investigación inicial, hemos optado por considerar la oración como la unidad básica de información para el proceso de generación de resúmenes. Para ello, se ha ll-

evado a cabo la segmentación de oraciones de forma automática usando la herramienta Stanford CoreNLP propuesta por Manning et al. (2014). En el caso de las reseñas, la tarea de la división automática de sentencias es muy complicada dado que las reseñas han sido escritas por todo tipo de usuarios por lo que la informalidad o las faltas de ortografía son muy comunes, lo que puede influir en el rendimiento de la herramienta automática seleccionada. Como el tamaño de un tweet no puede superar los 140 caracteres, y en su mayoría el tweet suele ser una única sentencia, no se ha llevado a cabo ningún proceso de división del texto para este género textual.

Por otro lado, nos interesa saber cuáles son los términos más comentados de cada establecimiento y fuente de información con el objetivo de poder conocer qué tema/s predominan (por ejemplo, habitaciones, servicio, limpieza, etc.). Para obtener dichos términos, llevamos a cabo un proceso de tokenización, eliminación de palabras auxiliares, lematización de las palabras y la identificación de la categoría léxica. Como resultado de este proceso, consideramos los 100 lemas más frecuentes con categoría sustantivo de cada establecimiento y cada fuente de información. Esta información nos ayudará en etapas posteriores para determinar la información relevante y componer el resumen final.

3.2 Análisis y clasificación de polaridad

En esta fase, vamos a determinar la polaridad asociada a cada frase/tuit. Debemos conocer si una frase es positiva, negativa o neutra para conocer si el usuario está opinando o no, y si lo hace, saber si es para dar información sobre aspectos buenos o malos. Por ejemplo, la frase *"fantástica habitación"* tendrá una polaridad positiva, mientras que la frase *"la ubicación del hotel era muy mala"* se clasificará como negativa. Para lograr este objetivo se ha utilizado la herramienta propuesta por Fernández et al. (2015). Además, esta herramienta nos dará también la intensidad de opinión, por lo que nos guardaremos ese valor para etapas posteriores.

3.3 Agrupación de las frases/tuits

El objetivo de esta fase es agrupar las frases/tuits con el fin de conocer las semejanzas que hay entre ellas. Para ello en primer lugar agrupamos las frases/tuits según su polaridad mediante la herramienta propuesta por Fernández et al. (2015). Una vez hechos los grupos por polaridad, se ha calculado la similitud entre frases usando el método del coseno a través de la librería Simmetrics[1], determinando empíricamente el umbral de similitud, de tal manera que frases con una similitud de al menos un 30% serán agrupadas en el mismo conjunto. La razón por la que este umbral es más permisivo se debe a que más que buscar frases idénticas o muy similares, lo que queremos es identificar frases que puedan estar hablando de lo mismo. Además, cabe indicar que los grupos de frases son conjuntos disjuntos, es decir, una frase no puede pertenecer a dos conjuntos de frases. Esta decisión se tomó para disminuir la complejidad computacional del proceso de generación de resúmenes.

3.4 Determinación de la relevancia de cada frase

El objetivo de esta fase es calcular la relevancia de cada frase para establecer un ranking entre las frases de cada grupo obtenido de la fase anterior. Para ello se puntúa cada frase teniendo en cuenta las siguientes heurísticas:

- **Relevancia de la frase según el tipo de fuente de información a la que pertenece.** Para esta heurística, diferenciamos entre la fuente de información de reseñas y de microblogs, considerando más importante las reseñas en una proporción 70/30, dado que se trata de un tipo de género textual más especializado en opiniones.
- **Relevancia de la frase respecto a su propia fuente de información.** En este caso comparamos los lemas de cada frase pertenciente a una reseña o tuit con el listado de los 100 lemas más utilizados del establecimiento en todas las reseñas o tuits recopilados, respectivamente; si existe el lema en el listado, la puntuación aumentará en base a la frecuencia del lema, del tal manera que se irán sumando las frecuencias de los lemas de esa frase para calcular la relevancia de esa frase. Si el lema no está entre los 100 lemas más frecuentes el valor para ese lema será 0.

[1] https://github.com/Simmetrics/simmetrics

"La gran mayoría de las opiniones destacan su excelente servicio" *(establecimiento: hotel - aspecto: servicio - puntuación: 9 - sentimiento: positivo)*
"Las opiniones indican que el servicio es correcto" *(establecimiento: hotel - aspecto: servicio - puntuación: 8 - sentimiento: negativo)*
"Los clientes lo utilizan para hospedarse en viaje de negocios" *(establecimiento: hotel - aspecto: null - puntuación: null - perfil del cliente: de negocios)*
"Los clientes recomiendan este restaurante por su magnífica gastronomía" *(establecimiento: restaurante - aspecto: comida - puntuación: 9.5 - sentimiento: positivo)*
"Según la valoración de los clientes es uno de los mejores de la ciudad" *(establecimiento: restaurante - aspecto: valoración general - puntuación: 10 - sentimiento: positivo)*

Tabla 1: Ejemplo de frases de enlace generadas

- **Intensidad de la frase.** Este valor se obtiene del análisis de polaridad realizado en etapas anteriores y lo consideraremos para dar preferencia a las frases que expresen opiniones de una forma más intensa.

- **Complementos del sustantivo.** La puntuación de los complementos del sustantivo se realiza mediante un análisis morfológico donde intentamos puntuar la aparición de adjetivos seguidos por sustantivos o viceversa con el objetivo de seleccionar las frases que sean más descriptivas. Mediante esta heurística, se valorará positivamente que las frases contengan adjetivos calificando a los sustantivos con el fin de aportar mayor expresividad al resumen.

3.5 Preparación de frases para el resumen

En esta fase llevamos a cabo un proceso de conversión de las frases a un modo más impersonal. Al tratar con información procedente de la Web 2.0 existe mucha variabilidad en la manera de escribir, y por tanto es necesaria una forma de homogeneizar la escritura para aportar mayor legibilidad al texto y que no sea un compendio de frases sueltas. Para ello, diseñamos un proceso basado en reglas para convertir la frase a tercera persona del plural a partir de la extracción de información morfológica presente en la propia frase. Por ejemplo, *"Nos alojamos en la planta superior, desde la cual podíamos ver la la inmensidad del mar"* pasaría a *"Se alojaron en la planta superior, desde la cual podían ver la inmensidad del mar"*.

3.6 Creación de frases de enlace

En este trabajo, la propuesta de resúmenes abstractivos es una propuesta mixta que combina frases extraídas de las fuentes de información con la creación de un conjunto de frases predefinidas que permiten aportar mayor naturalidad a los resúmenes generados de forma automática, basándonos en la idea sugerida por Gerani et al. (2014). Esta opción sería la más adecuada teniendo en cuenta que trabajamos a nivel de oraciones, dado el tipo de textos tratados. Para conseguir un mayor nivel de abstracción y generar un resumen puramente abstractivo, se necesitaría trabajar con una granularidad más fina, como por ejemplo, cláusulas, o tripletas, tal y como se realiza en (Li, 2015; Khan, 2015) pero en primer lugar sería necesario verificar que las herramientas existentes para este propósito se comportan de la misma manera para textos informales de cara a minimizar la introducción de errores en el proceso.

Se generaron un total de 176 frases predefinidas para los hoteles y 160 para los restaurantes. Dichas frases se clasificaron dependiendo del aspecto al que hicieran referencia, puntuación otorgada sobre el aspecto, perfil de usuario, tipo de opinión, etc. teniendo en cuenta los metadatos que podíamos extraer de las fuentes de información utilizadas. La tabla 1 muestra algunos ejemplos de frases predefinidas. Además se crearon algunas oraciones con el objetivo de introducir las frases elegidas que formarán parte del resumen, tales como *"De lo más comentado por los clientes es que"* o *"Las opiniones de los clientes enfatizan que"*.

3.7 Composición del resumen

Debido a la gran cantidad de información y opiniones existentes para cada tipo de establecimiento, como muestra la tabla 2, decidimos generar tres resúmenes por cada establecimiento: i) resumen mixto, que mostrará los aspectos más importantes para bien y para mal del establecimiento; ii) resumen positivo, que mostrará únicamente las características más relevantes que han gustado más a los clientes; y iii) resumen negativo que sólo mostrará los aspectos más importantes que no han gustado a los clientes. El proceso de composición para cada tipo de los resúmenes se detalla a continuación.

Para generar un **resumen mixto**, se llevan a cabo los siguientes pasos:

- Introducir frase predefinida respecto a la valoración general del establecimiento con sentimiento positivo.
- Introducir dos frases predefinidas sobre distintos aspectos del establecimiento y/o perfil de usuario (por ejemplo, limpieza, servicio) con sentimiento positivo.
- Introducir frase predefinida para introducir las opiniones de los clientes.
- Añadir las frases más relevantes del conjunto de frases neutras, positivas y negativas (1 frase por cada grupo, controlando que la similitud de la frase con mayor puntuación no supera un umbral predefinido para evitar así redundancia y contradicciones (caso por ejemplo, de que a un usuario le haya gustado un aspecto del hotel y lo haya valorado positivamente, y a otro el mismo aspecto no le haya gustado y por tanto lo haya valorado negativamente).

La generación del **resumen positivo** y del **resumen negativo**, se realiza de forma análoga a la del resumen mixto pero escogiendo las frases del conjunto de frases clasificadas con polaridad positiva y negativa, respectivamente.

4 Experimentación

Para la generación de resúmenes, elegimos como fuentes de información: la página especializada de reseñas TripAdvisor[2] y la red social Twitter[3], debido por una parte al gran volumen de datos que podíamos obtener y por otra, a que se trataba de fuentes heterogéneas pertenecientes a distintos géneros textuales. A partir de estas fuentes de información, recopilamos un corpus con los comentarios y tuits de restaurantes y hoteles de 6 ciudades nacionales e internacionales (Alicante, Barcelona, Londres, Madrid, Roma y Valencia), utilizando técnicas de crawling y reglas para asegurar de que los comentarios y/o tuits estuvieran en castellano y relacionados con opiniones o comentarios sobre los establecimientos. Además, para el caso particular de los tuits, se eliminaron las menciones (p. ej: @perfil) y los enlaces (p. ej. http://www.melialicante.com). La tabla 2 muestra las estadísticas del corpus recopilado.

Número de hoteles	180
Número de hoteles por ciudad	30
Número de restaurantes	180
Número de restaurantes por ciudad	30
Número de comentarios TripAdvisor	91505
Número de comentarios de TripAdvisor por hotel	235
Número de comentarios de TripAdvisor por restaurante	275
Número de tuits	78713
Número de tuits por hotel	200
Número de tuits por restaurante	240

Tabla 2: Estadísticas de la información extraída

Para cada hotel y restaurante de cada ciudad, se generaron 3 resúmenes, creando finalmente un total de 1080 resúmenes de forma automática. La tabla 3 muestra, a modo de ejemplo, los tres tipos de resúmenes generados para el hotel *NH Valencia*.

5 Evaluación, resultados y análisis

Respecto a la evaluación, se llevó a cabo una evaluación manual con 41 usuarios reales para valorar los resúmenes generados de manera cualitativa. Mediante una pequeña interfaz que contenía todos los resúmenes generados, y para no restringir el resumen que el usuario debía de valorar, se permitió a cada usuario evaluar el tipo de resumen del hotel/restaurante que quisiera sin establecer ningún requisito adicional. Simplemente se les pedía que anotaran el establecimiento elegido y el tipo o tipos de resúmenes consultados. La tarea consistía en puntuar con una nota de 0 a 10, siendo un 0 la peor nota y 10 la mejor cada uno de los siguientes criterios: i) coherencia del resumen: para de-

[2]https://www.tripadvisor.es/
[3]https://twitter.com/

Resumen mixto
El hotel NH Valencia Las Artes según la valoración de los clientes es un buen hotel para hospedarse. La mayoría de las opiniones recomienda el hotel por la buena relación calidad-precio, además los clientes destacan que es excelente para venir en pareja. Entre lo más comentado por los clientes, se encuentra que hotel ubicado en la zona de la ciudad de las artes & ciencias de valencia, lobby pequeño y habitación individual chica pero con un plus inmejorable, una terraza enorme con vista a la ciudad de valencia donde vale la pena apreciar el atardecer y amanecer. Desde la llegada trato e informacion humana muy acogedora las habitaciones modernas, espaciosas y luminosas con hermosas vistas, baños confortables tiene instalaciones para ejercicios, piscina y solarium el bufet para desayuno variado y rico. Por el contrario las habitaciones un poco pequeñas pero cuidadas, pedid mejor las habitaciones que dan a la plaza interior porque sino son algo ruidosas las exteriores.
Resumen positivo
El hotel NH Valencia Las Artes según la valoración de los clientes es un buen hotel para hospedarse. El hotel es recomendable para venir en pareja, además el hotel destaca por tener una gran relación calidad-precio. Las opiniones destacan que habitación normal, hotel cerca de la ciudad de las artes y la playa de la malvarrosa La limpieza de la habitación era inexistente, polvo acumulado por toda la habitación, sin embargo el baño estaba limpio. Otro aspecto importante es que hotel muy céntrico y con muy buen servicio, las habitaciones están muy bien y con buenas vistas, el buffet del desayuno muy completo y todo muy bueno, el personal de servicio muy agradable y se puede descansar muy bien en las habitaciones, muy tranquilo para estar en el centro de valencia. Algo a destacar es que a nuestro parecer este hotel fue realmente perfecto, su ubicación es inmejorable, al lado de la ciudad de las ciencias, al lado de un centro comercial grande, al lado de unos restaurantes, se gustó la taberna de maria, tampoco tan alejado del centro.
Resumen negativo
El hotel NH Valencia Las Artes según la valoración de los clientes es un hotel correcto para pasar unos días. El hotel cuenta con un servicio del que no debemos preocuparnos según la mayoría de las opiniones, además la limpieza del hotel es aceptable. Las opiniones destacan que la ubicación no puede ser mejor, en la puerta de la ciudad de las artes y las ciencias, cerca del parque, dos centros comerciales y. Adicionalmente las habitaciones y el desayuno muy bien pero es caro la atenxion a la hora de *habrir* las habitaciones comunicadas entre si mal. Además algo lejos del centro pero de fácil acceso desde allí, lo peor la limpieza de la habitación, que dejaba algo que desear.

Tabla 3: Resúmenes automáticamente generados para el hotel NH Valencia (como se observa, debido a que las opiniones han sido extraídas de usuarios reales, a veces podemos encontrar faltas de ortografía (como ocurre en el caso de *habrir*)

terminar si el texto generado mediante las técnicas propuestas tenía sentido como un todo y estaba bien estructurado y enlazado; ii) utilidad: para conocer si el resumen leído le podría servir al usuario o no; y iii) errores ortográficos: para conocer la presencia de faltas de ortografía en los resúmenes, derivados del uso de contenido generado por otros usuarios.

En conjunto, los usuarios evaluaron un total de 332 resúmenes (171 procedentes de hoteles y 161 de restaurantes), variando de 1 a 10 usuarios, el número de personas que evaluaron el mismo resumen. Las figuras 1 y 2 muestran los resultados medios globales y lectura de tipo de resúmenes por tipo de establecimiento, respectivamente. A continuación, realizamos un análisis más detallado de los mismos.

En líneas generales, los resultados obtenidos son buenos, teniendo en cuenta los retos que nos planteábamos con el método propuesto: generación de resúmenes abstractivos multigénero y multidocumento, puesto

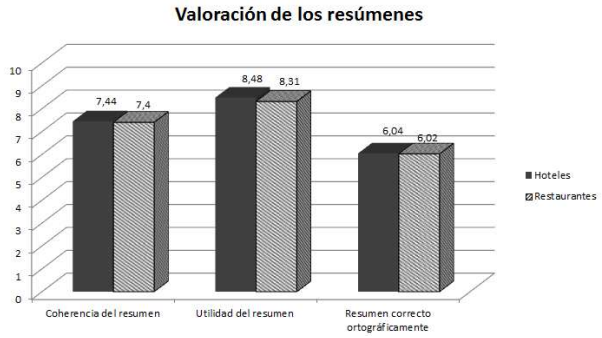

Figura 1: Resultados globales de los resúmenes generados para los hoteles y restaurantes

que trabajamos con varias reseñas y/o tuits de un mismo establecimiento. Si nos centramos en la figura 1 podemos observar que los resúmenes sobre hoteles han sido mejor valorados en todos las apartados descritos, aunque debemos decir que las diferencias respecto a los resúmenes de los restaurantes son mínimas. Este resultado es positivo ya que demuestra que el enfoque propuesto funciona bien para dos dominios distintos y esto nos lleva a pensar que se podría aplicar directamente a otros tipos de productos (películas, libros, móviles, ordenadores,...) y/o servicios (eventos, atracciones,...).

Concretamente, el apartado mejor valorado por los usuario fue la utilidad de los resúmenes alcanzando una nota media por encima de 8. Respecto a la coherencia también se obtuvieron buenos resultados, con notas superiores a 7,40. Sin embargo, el apartado más criticado en la generación de los resúmenes fue la existencia de errores ortográficos y gramaticales en los resúmenes (como el que se ha ilustrado en el resumen negativo de la tabla 3). Es por esto que la nota media otorgada por los usuarios para este criterio se queda cercana al 6, no alcanzando así el notable. Esto nos indica que se deberían refinar la fase de recopilación y extracción de información de fuentes heterogéneas, intentando identificar informalidades y faltas de ortografía. Otra propuesta podría ser añadir un factor de errores gramaticales en la ponderación de las frases para evitar que las frases con faltas de ortografía formaran parte de los resúmenes generados.

Fijándonos en los resúmenes que han obtenido peores valoraciones, identificamos una serie de mejoras que se podrían realizar para mejorar la calidad del método propuesto, algunas de ellas relacionadas con el uso de las herramientas externas utilizadas, como es el caso de la herramienta de clasificación de opiniones. Si la polaridad de la frase se etiqueta de manera errónea, el resumen generado puede no cumplir la finalidad con la que se creó (por ejemplo, se ha observado que muchas veces los resúmenes negativos contienen frases con polaridad positiva y por tanto, el resultado no es un resumen enfatizando los peores aspectos del establecimiento en cuestión). Para mitigar este error, se podría hacer uso de otra herramienta adicional con el mismo cometido y si ante un mismo texto las dos herramientas obtienen el mismo resultado almacenar el sentimiento de la frase, de esta forma evitaríamos el uso exclusivo de una única herramienta. También se ha observado que el paso de algunos verbos a tercera persona ha comportado faltas de ortografía, que se añadirían a las posibles faltas cometidas por los usuarios autores de las reseñas/tuits. Esto sucede especialmente cuando el verbo es irregular, se debería utilizar una lista completa de los verbos irregulares conjugados, para que en el caso de detectarse la existencia de uno de ellos, se haga una sustitución simple, sin comportar ningún tipo de procesamiento que diera lugar a errores.

Por otro lado, analizamos también qué tipo de resúmenes preferían consultar los usarios. Los resultados obtenidos se muestran en la figura 2, donde se observa que el resumen mixto es el más consultado, seguido del positivo y finalmente, el menos leído es el resumen negativo. Y esto ocurre tanto para los hoteles como los restaurantes, indicando que al parecer los usuarios prefieren disponer de un resumen donde se expongan tanto los aspectos positivos y negativos sobre un establecimiento y no tanto sólo un tipo de ellos.

6 Conclusión y trabajo futuro

En este trabajo de investigación se ha presentado un método para la generación automática de resúmenes abstractivos multigénero y multidocumento para el dominio turístico. Las herramientas lingüísticas utilizadas junto con las heurísticas propuestas nos han permitido obtener un método capaz de producir tres tipos de resúmenes abstractivos dependiendo de la perspectiva que queramos ofrecer (sólo lo mejor, sólo lo

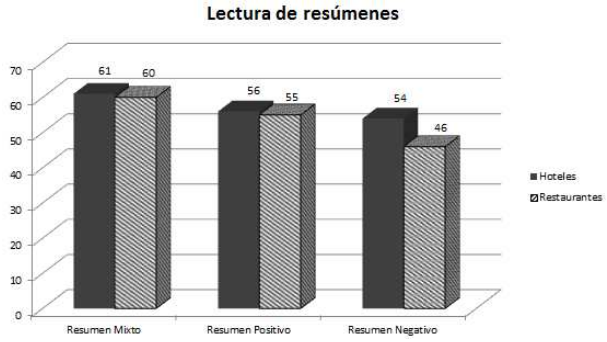

Figura 2: Resultados sobre el tipo de resumen más consultado

peor, o ambos).

Los resultados obtenidos a partir de la evaluación mediante usuarios reales han demostrado la viabilidad y utlidad del método para ser integrado o aplicado a escenarios reales, si bien es cierto que primero se deberían mejorar algunos aspectos, como es el de la corrección ortográfica de los resúmenes producidos que sería la primera mejora que abordaríamos de cara a trabajos futuros. Adicionalmente, nos planteamos integrar técnicas de generación de lenguaje para mejorar la fase de creación de frases de enlace y probar el método de resúmenes en otros contextos, puesto que no sólo tendría aplicación en el dominio turístico. Por ejemplo, en multitud de portales Web, los clientes expresan su opinión acerca de los productos adquiridos, lugares visitados, servicios consumidos, etc. Podríamos utilizar esta información junto con información factual provista por el fabricante del producto, lugar, o servicio como fuentes de información y aplicar el método propuesto para elaborar un resumen que se centre en los mejores y peores aspectos. La mayor dificultad radicaría en el pre-procesamiento de las fuentes de información y el grado de informalidad que éstas contuvieran, ya que el núcleo del método se podría utilizar directamente.

Agradecimientos

Esta investigación ha sido financiada por la Generalitat Valenciana a través del proyecto DIIM2.0: Desarrollo de técnicas Inteligentes e Interactivas de Minería y generación de información sobre la web 2.0 (PROMETEOII/2014/001), y por el Gobierno de España (MINECO) a través de los proyectos TIN2015-65100-R, TIN2015-65136-C2-2-R.

References

Carenini, G., J. C. K. Cheung, y A. Pauls. 2013. Multi-document summarization of evaluative text. *Computational Intelligence*, 29(4):545–576.

Fernández, J., Y. Gutiérrez, J. M. Gómez, y P. Martínez-Barco. 2015. Social rankings: análisis visual de sentimientos en redes sociales. *Procesamiento del Lenguaje Natural*, 55:199–202.

Gerani, S., Y. Mehdad, G. Carenini, R. T. Ng, y B. Nejat. 2014. Abstractive summarization of product reviews using discourse structure. En *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, páginas 1602–1613, Doha, Qatar, October. Association for Computational Linguistics.

Khan, Atif y Salim, N. y. J. K. Y. 2015. A framework for multi-document abstractive summarization based on semantic role labelling. *Appl. Soft Comput.*, 30(C):737–747, Mayo.

Li, W. 2015. Abstractive multi-document summarization with semantic information extraction. En *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, páginas 1908–1913, Lisbon, Portugal, September. Association for Computational Linguistics.

Lloret, E. y E. Boldrini. 2015. Multi-genre summarization: Approach, potentials and challenges. En *eChallenges e-2015 Conference*, páginas 1–9, Nov.

Lloret, Elena y Palomar, M. 2012. Text summarisation in progress: A literature review. *Artif. Intell. Rev.*, 37(1):1–41, Enero.

Manning, C. D., M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, y D. McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. En *Association for Computational Linguistics (ACL) System Demonstrations*, páginas 55–60.

Saggion, H. 2011. Learning predicate insertion rules for document abstracting. En *Computational Linguistics and Intelligent Text Processing - 12th International Conference, CICLing 2011, Tokyo, Japan, February 20-26, 2011. Proceedings, Part II*, páginas 301–312.

Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 61-68 recibido 29-11-2016 revisado 12-01-2017 aceptado 07-02-2017

Cross-Document Event Ordering through Temporal Relation Inference and Distributional Semantic Models*

Ordenación de eventos multidocumento usando inferencia de relaciones temporales y modelos semánticos distribucionales

Estela Saquete, Borja Navarro-Colorado
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Carretera San Vicente s/n 03690 Alicante España
{stela,borja}@dlsi.ua.es

Abstract: This paper focuses on the contribution of temporal relations inference and distributional semantic models to the event ordering task. Our system automatically builds ordered timelines of events from different written texts in English by performing first temporal clustering and then semantic clustering. In order to determine temporal compatibility, an inference from the temporal relationships between events –automatically extracted from a Temporal Information Processing system– is applied. Regarding semantic compatibility between events, we analyze two different distributional semantic models: LDA Topic modeling and Word2Vec word embeddings. Both semantic models together with the temporal inference have been evaluated within the framework of SemEval 2015 Task 4 Track B. Experiments show that, using both models, the current State of the Art is improved, showing significant advance in the Cross-Document Event Ordering task.
Keywords: Temporal information, event coreference, temporal inference, distributional semantics, event ordering

Resumen: Este artículo se centra en estudiar la contribución que la inferencia de relaciones temporales y los modelos semánticos distribucionales hacen a la tarea de ordenación de eventos. Nuestro sistema construye automáticamente líneas de tiempo con eventos extraídos de diferentes documentos escritos en inglés. Para ello realiza primero una agrupación temporal y posteriormente una agrupación semántica. Para determinar la compatibilidad temporal se realiza una inferencia sobre las relaciones temporales entre los eventos extraídos de una sistema automático de procesamiento de información temporal. Para la compatibilidad semántica entre eventos hemos analizado dos modelos semánticos distribucionales distintos: LDA Topic Modeling y Word2Vec Word Embeddings. Ambos modelos semánticos junto con la inferencia temporal han sido evaluados bajo el marco de evaluación de SemEval 2015 Task 4 Track B. Los experimentos muestran que, usando ambos modelos se mejora el estado del arte actual, implicando un avance importante en la tarea de ordenación de eventos multidocumento.
Palabras clave: Información temporal, correferencia de eventos, inferencia temporal, semántica distribucional, ordenación de eventos

1 Introduction

Cross-document event ordering was the topic of the latest SemEval-2015 Task4 (Minard et al., 2015), called "TimeLine: Cross-Document Event Ordering". It consists of, first, extracting events involving a particular target entity among different documents, and, then, ordering them chronologically in a timeline.

Considering one specific entity as the target entity, all the events related to the target entity are extracted from several documents and arranged in a timeline.

The approach to cross-document event ordering presented in this paper is based on the idea that two or more event mentions corefer if they have not only temporal compatibility but

* This paper has been partially supported by the Spanish government, project TIN2015-65100-R, project TIN2015-65136-C2-2-R and PROMETEOII/2014/001.

ISSN 1135-5948

also semantic compatibility. In order to determine temporal compatibility, our approach uses temporal relationships between events extracted from a Temporal Information Processing system. Regarding semantic compatibility, in this paper we analyze two different distributional semantic models: LDA Topic Modeling (Blei, Ng, and Jordan, 2003) and Word2Vec Word Embeddings (Mikolov et al., 2013). We use this Semeval Timeline task as an evaluation and discussion framework.

This paper is organized as follows: next section (Section 2) presents the State of the Art in cross-document event ordering. Then, in Section 3, we explain our approach: how we formalize temporal and semantic compatibility, and how we apply distributional semantics to find coreferential events. Section 4 is devoted to the analysis, evaluation and discussion of the different approaches that have been implemented, and finally the main conclusions are presented in Section 5.

2 State of the Art

The most recent conferences about temporal information processing and temporal relation extraction were part of the SemEval challenges: *TempEval-1*, Temporal Relation Identification (Verhagen et al., 2007); *TempEval-2*, Evaluating Events, Time Expressions And Temporal Relations (Verhagen et al., 2010); and *TempEval-3*, Temporal Annotation (UzZaman et al., 2013). Challenges like the 6th i2b2 NLP Challenge (Sun, Rumshisky, and Uzuner, 2013) also focused on temporal relations but within a clinical context. All of theses challenges mainly focused on temporal relations of events, in order to: a) discover which of them occur before, simultaneously or after the others, and b) annotate all this temporal information (events, timex and relations) using the TimeML annotation scheme.

Regarding cross-document event coreference Bagga and Baldwin (1999) proposed one of the first approaches in this area. Ji et al., (2009) worked on a timeline task using the ACE 2005 training corpora. Bejan and Harabagiu (2014) performed cross- and within-document approaches using a rich set of linguistic features to model the event structure: lexical features such as head words and lemmas, class features such as PoS or event class, semantic features such as WordNet sense or semantic-roles frames, etc. Their proposal follows an unsupervised approach based on a non-parametrical Bayesian model. In the work presented by Li et al., (2011) the goal was to provide an event-fusion approach to obtain the most complete event possible by combining a set of coreference event mentions from different documents which were crawled from Websites. Another cross-document approach is proposed by Lee et al., (2012) introducing a novel coreference resolution system that models entities and events jointly. Cybulska and Vossen (2013) apply an event model based on four components: location, time, participant and action. They avoid the use of machine-learning methods in order to analyze how event components influence event coreference. Goyal et al., (2013) use a syntax-based distributional semantic approach on event coreference resolution. Lu and Ng (2016) present an event coreferent resolution system based on several sieves as similar lemmas, similar modifiers, hypernyms, etc. Finally, Yang et al., (2015) present a hierarchical distance-dependent Bayesian model for within- and cross-document event coreference resolution, concluding that it is a powerful approach to resolve the task in comparison to other state-of-the-art approaches.

Most recently, SemEval-2015 included the task that tried to combine temporal processing and event coreference in order to obtain a timeline of events related to a specific given entity from a set of documents (Minard et al., 2015). They proposed two different tracks on the basis of the data used as input. Track A, for which they provided only raw text sources, and Track B, for which they also made gold event mentions available. Track A had two participants: WHUNLP team and SPINOZAVU team. WHUNLP team processed the texts with Stanford CoreNLP[1] (Manning et al., 2014) and applied a rule-based approach to extract target entities and their predicates. They also performed temporal reasoning. The SPINOZAVU system (Caselli et al., 2015) is based on a pipeline developed in the NewsReader project. It addressed entity resolution, event detection, event-participant linking, coreference resolution, factuality profiling and temporal relation processing, first at document level, and then at cross-document level, in order to obtain timelines. Track B participants were the Heideltoul team and the GPLSIUA team. The Heideltoul approach (Moulahi et al., 2015) uses the HeidelTime tool for temporal information processing and the Standford CoreNLP for event coreference resolution. The GPLSIUA approach (Navarro-Colorado and Saquete, 2015) uses the OPENER language anal-

[1] http://stanfordnlp.github.io/CoreNLP/

ysis toolchain for entity detection, the TIPSem tool (Llorens, Saquete, and Navarro-Colorado, 2012) for temporal processing and two different approaches to detect event coreference: lexical semantics and a Topic Modeling algorithm over WikiNews corpus. Later works such as Laparra et al., (2015) showed that explicit temporal relations are not enough to obtain a full time-anchor annotation of events and evidenced the need for a temporal analysis at document level.

In this paper we reanalyze the impact of combining temporal inference with two of the (current) most important distributional models: LDA Topic Modeling (Blei, Ng, and Jordan, 2003) and Word2Vec Word Embeddings (Mikolov et al., 2013) in the cross-document event ordering task. In next section, we explain how both distributional semantic models have been adapted to event coreference resolution.

3 Our approach

Our approach to Cross-Document Event Ordering is based on the idea that two events *e1* and *e2* are coreferent if they have not only temporal compatibility but also semantic compatibility (they refer, in some way, to the same facts) (Navarro-Colorado and Saquete, 2016). Formally:

$$coref(e1, e2) \rightarrow (e1_t = e2_t) \wedge (e1_s \simeq e2_s)$$

From this idea, our system is structured in four steps:

1. Event and temporal relation extraction using Temporal Information Processing;
2. Target Entity Filtering in order to select the events related to the target entity;
3. Temporal clustering through temporal compatibility inference;
4. Semantic clustering using two different distributional semantic models: LDA Topic Modeling and Word2Vec Word Embeddings.

Each step is explained in depth in the next subsections.

3.1 Temporal Information Processing

The first module of the proposed architecture performs Temporal Information Processing using TIPSem system. It automatically annotates all the temporal information according to TimeML standard annotation scheme (Saurí et al., 2006), which means annotating all the temporal expressions (TIMEX3), events (EVENT) and links between them.

3.2 Target Entity Filtering

Considering that not all the events annotated by the previous module are necessary to build the timeline, but only the ones related to a target entity, a Target Entity Filtering needs to be performed in order to avoid those events that are annotated but not related to the given entity.

The Target Entity Filtering requires resolving name entity recognition and entity coreference resolution. Since this is not the main challenge of our research, this task is performed using an external tool. That is why the OPENER[2] web services were integrated in our proposal. More specifically, the NER and the coreference resolution component.

To determine if an event must be part of the timeline or not, this module selects the events in which a target entity (or a target entity coreference) explicitly participates in a *has_participant* relation with the semantic role A0 (i.e. agent) or A1 (i.e. patient), as defined in the Propbank Project (Palmer, Gildea, and Kingsbury, 2005).

3.2.1 Temporal Compatibility Clustering

As we have explained before, a Temporal Information Processing system such as TIPSem works at document level and is able to extract from each document all the explicit temporal information as well as establish temporal relations between times and events or between events. However, in order to establish a cross-document timeline of events, this is not enough. It is necessary to know explicitly the time at which each event occurs, and to perform cross-document event clustering.

One must infer the time-anchoring of all the selected events from the temporal information extracted by TIPSem in each document (within-document temporal inference). Through this inference, the temporal clustering of all the events of the different documents is performed (cross-document temporal inference). As previously stated, we consider two events to be clustered when they are temporally compatible, that is, if they happen at the same time.

Our model infers temporal compatibility in two steps:

[2]http://www.opener-project.eu/webservices

- **Within-document temporal clustering**: For each document, the temporal information of each event is going to be extracted. Each event is anchored to a time anchor[3] when a SIMULTANEOUS temporal link exists between this event and a temporal expression. After this, two events are considered part of the same cluster if they are temporally compatible, meaning that: a)they are anchored to the same time anchor, or b) they have a SIMULTANEOUS temporal link between them.
- **Cross-document temporal clustering**: From a set of documents (related by the same topic), and considering that in the previous step all the events were assigned a time anchor, all the events in the different documents that are temporally compatible, that is, are anchored to the same time anchor, are clustered together.

3.2.2 Semantic Compatibility Clustering: distributional semantics and event coreference

All those events occurring at the same time and being semantically compatible must be part of the same cluster in the timeline of a specific entity. The problem is that it is not exactly clear which components of the event structure are determinant in event coreferent resolution (Cybulska and Vossen, 2013).

Rather than creating a complex feature matrix to represent the semantics of the argument as Bejan and Harabagiu (2014) does, we propose a compact, use-based distributional representation of the semantics of the arguments. Moreover, contrary to Goyal et al., (2013), who use a syntax-based distributional representation, we use the argument structure of the event.

In this regard, when we apply distributional semantic models we are considering the context of the events as the main component that contributes to establish the semantic compatibility and, therefore, the event coreference. Current computational models of distributional semantics are based on the word/document model of Information Retrieval. In order to increase the semantic representativity of the vector space and to resolve data sparseness problems, different models have been proposed such as, among others, LSA Latent Semantic Analysis (Landauer and Dumais, 1997), LDA Latent Dirichlet Allocation (Topic Modeling) (Blei, Ng, and Jordan, 2003) or, recently, Word2Vec Word Embeddings (Mikolov et al., 2013). In this paper we apply Topic Modeling and Word Embeddings to the event coreference resolution task.

LDA Topic Modeling extracts a predefined set of topics from large corpora. Each topic is a distribution over a fixed vocabulary. In order to assign words to topics, LDA uses two values: the topic assigned to a word in other texts and the most frequent topic in the text where the word appears. Through several iterations, in the end the corpus is represented as a word-topic matrix, in which each topic is composed of the weight of each word in it. Through LDA Topic Modeling, a high dimensional vector space is reduced to a k topics vector space (Blei, Ng, and Jordan, 2003).

In contrast, Word Embeddings Word2Vec is a predictive model that works better with very high vector spaces. It learns about distributed representation of words based on neural networks. From the point of view of semantic representation, instead of trying to reduce the dimensionality of the vector space as LSA and LDA do, Word2Vec tries to optimize the representation of the context where a word appears: on one hand, through the continuos skip-gram model, Word2Vec maximizes relevant contexts; on the other hand, through negative sampling it assigns high probability to relevant words and low probability to noise words (Mikolov et al., 2013). Word2Vec does not use a linguistic-motivated context size: it applies a window of size k. LDA Topic Modeling, in contrast, tends to the establish the text as context, considering relevant the most frequent topic of each text to specify the topic of each word (Blei, Ng, and Jordan, 2003).[4]

We use distributional semantic models for event coreference resolution as follows. Firstly, we apply both LDA Topic Modeling and Word2Vec to English Wikipedia. That way, we obtain two distributional knowledge-bases in which each word is represented as a contextual vector in a high dimensional space.

Through LDA Topic Modeling, the distributional knowledge-base is a vector space made up of 500 topics. Each word is, then, represented as a 500-dimension vector in which each value is the weight of the word in each topic. Through Word2Vec the distributional knowledge-base is the embeddings of each word in a space of 1000

[3] A time anchor is always a DATE (as defined in TimeML) and its format follows the ISO-8601 standard: YYYY-MM-DD, being the maximum granularity admitted in the task DAY.

[4] For a more systematic comparison of distributional models, see (Baroni, Dinu, and Kruszewski, 2014).

dimensions.[5]

Secondly, each event structure is represented as a compositional vector. Applying a Part of Speech tagger and a Semantic Role labeling, the event structure is extracted.[6] It is made up of the nouns, verbs and adjectives of the event head and the main arguments[7]. Then the contextual vector of each word is extracted from the distributional knowledge-base (Topic Modeling Knowledge-base on one hand, Word2Vec knowledge-base on the other). Finally, following the additive model (Mitchell and Lapata, 2010), all word vectors are added up into a single compositional vector that represents the distributional meaning of the whole event structure.

Formally, the event structure is represented as a tuple of three elements: two arguments ($A0$ and $A1$) and one event head (H):

$$ES =< A0, A1, H >$$

Each argument is a compositional vector $\overrightarrow{V}(A)$ formed by the sum of the contextual vector $\overrightarrow{V}(w_n)$ of each word in the argument (w_n):

$$\overrightarrow{V}(A) = \sum^{n} \overrightarrow{V}(w_n)$$

The event head H is the contextual vector of a single word. Finally, the compositional vector of the whole event structure $\overrightarrow{V}(ES)$ is:

$$\overrightarrow{V}(ES) = \overrightarrow{V}(A0) + \overrightarrow{V}(A1) + \overrightarrow{V}(H)$$

Finally, in order to detect if two events are coreferential, the system calculates the cosine similarity between both event vectors. If the cosine similarity between two event vectors is higher than 0.9[8], the system concludes that there is a coreference between them and hence they are grouped together in the same cluster. Formally:

$$coref(\overrightarrow{V}(ES_1), \overrightarrow{V}(ES_2)) \Longrightarrow sim(\overrightarrow{V}(ES_1), \overrightarrow{V}(ES_2)) \geq 0.9$$

Event coreference is considered a transitive relation:

$$coref(\overrightarrow{V}(ES_a), \overrightarrow{V}(ES_b)) \wedge coref(\overrightarrow{V}(ES_b), \overrightarrow{V}(ES_c)) \Longrightarrow$$

$$coref(\overrightarrow{V}(ES_a), \overrightarrow{V}(ES_c))$$

4 Experiments and evaluation

In order to evaluate our approach, we have used the dataset provided for Task4 at SemEval 2015.[9] This dataset is composed of articles from Wikinews about three topics: 1) Airbus and Boeing; 2) General Motors, Chrysler and Ford; and 3) Stock Market. All the experiments shown in this section were performed using Track B input[10].

As a baseline, we have implemented a lexical (non-distributional) WordNet-based approach. With this baseline we assume that two events are coreferent if their event heads express the same concept. Therefore, two events are clustered together as coreferential if both event heads are the same word (that is, they have the same lemma), or both event heads are synonyms (that is, they share the same synset in WordNet).

Regarding the distributional semantic models applied to cross-document event ordering, the system has been run with six different configuration:

1. TC+TM0505: Temporal clustering + LDA Topic Modeling Semantic clustering considering the event head and the arguments in the same proportion.
2. TC+TM1000: Temporal clustering + LDA Topic Modeling Semantic clustering considering only distributional similarity between heads.
3. TC+TM0010: Temporal clustering + LDA Topic Modeling Semantic clustering considering only distributional similarity between arguments.
4. TC+W2V0505: Temporal clustering + Word2Vec Words Embedding Semantic clustering considering the event head and the arguments in the same proportion.
5. TC+W2V1000: Temporal clustering + Word2Vec Words Embedding Semantic clustering considering only distributional similarity between heads.
6. TC+W2V0010: Temporal clustering + Word2Vec Words Embedding Semantic clustering considering only distributional similarity between arguments.

The results are shown in Table 1. These data show the performance of the system according to the evaluation measures of Semeval2015-Task 4 Track B. Furthermore, the comparison between our results and the results obtained by

[5] In order to create these distributional knowledge bases, we have used Gensim (`https://radimrehurek.com/gensim/`) both to compute Topic Modeling and Word2Vec, and Wiki2Vec (`https://github.com/idio/wiki2vec`).

[6] We use a Python implementation of Collobert's SENNA (Collobert et al., 2011) (`https://pypi.python.org/pypi/practnlptools/1.0`).

[7] A0 and A1 following PropBank tagset (`https://verbs.colorado.edu/~mpalmer/projects/ace.html`).

[8] After some tests, we have settled a threshold of 0.9 over 1.

[9] http://alt.qcri.org/semeval2015/task4/index.php?id=data

[10] SemEval 2015 Task 4, two different tracks were proposed on the basis of the data used as input: Track A for which they provided only raw text sources, and Track B, for which they also provided available gold events mentions.

other systems can only be done against the GPLSIUA and HEIDELTOUL teams' outcome, since those were the teams that participated in the same Track at SemEval 2015-Task 4. Given that this is a very novel task, no other competition has been held.

Approach	Micro-F1	Micro-P	Micro-R
TC+LCV2 (baseline)	29.95	25.17	36.98
TC+TM0505	30.23	25.56	36.98
TC+TM1000	30.23	25.56	36.98
TC+TM0010	**30.28**	25.64	36.98
TC+W2V0505	**30.41**	25.82	36.98
TC+W2V1000	30.35	25.74	36.98
TC+W2V0010	**30.43**	25.85	36.98
GPLSI_Run1	25.36	21.73	30.46
HEIDELTOUL_Run2	18.34	13.58	28.23

Table 1: Cross-document Event Ordering Results

The best results were achieved by the Word2Vec Word Embeddings model, with an F1-measure higher than 30.40% both in the experiment considering the event head and the arguments in the same proportion, and in the experiment considering only distributional similarity of the event heads. Regarding Topic Modeling approaches, the best results for event ordering were achieved by the TM0010 experiment, in which only the similarity between arguments was taken into account. In any case, F1-score results are quite similar throughout the different distributional semantics models, with the same recall and very slight improvements in precision. The same recall is obtained because all the experiments share the same selection of events and temporal inference. These data show that, in this evaluation framework, there are no great differences between lexical-based and distributional-based models on one hand (only 0.5 points of improvement), and Topic Modeling and Word Embeddings on the other hand. This happens because the task is more focused on event ordering and the event clusters are quite small in general, therefore, the impact of the event coreference resolution modules is not outstanding.

In comparison with other approaches at Semeval2015-Task 4 competition, all our experiments outperform the state-of-the-art systems in all their runs and all metrics (see Table 1), with a remarkable difference in F1-score of 12.09 points with HEIDELTOUL and 5.07 points with GPLSI when we compare the best solutions given by both systems. These data show clearly that the fully distributional models are suitable for the event ordering task.

To conclude, after a qualitative analysis of the results obtained by our approach, with the SEMEVAL 2015 Task 4 corpus and its evaluation framework, we can say that the performance of the distributional (contextual) representation of each word in event coreference is quite similar to lexical representations. However, regarding event ordering, if we compare the results of our distributional approach with the systems presented at SemEval 2015 task 4, the improvement of distributional models is significant. Besides, in this evaluation framework, there are not any significant differences between the performance of LDA Topic Models and Word2Vec Word Embeddings distributional models. Therefore, it is not possible to conclude that for event coreference resolution an optimized local context is better than an (abstract) textual topic model or the other way round. Both contextual representations are similar regarding event coreference resolution in this framework, and further experiments with specific event coreference corpora like ECB and ECB+ are required.

5 Conclusions

The main aim of this paper is to determine what is the contribution of temporal inference and distributional semantic models to the cross-document event ordering task.

Regarding temporal clustering, in order to determine the time-anchoring of the events and to cluster together those which happen at the same time, our approach uses the temporal relationships between events obtained by a Temporal Information Processing system called TIPSem.

In order to analyze the impact of applying distributional semantic approaches to the task, two different methods are analyzed: LDA Topic Modeling and Word2Vec Word Embeddings. Each distributional model has been run with three different configurations: 1. considering distributional similarity between event heads and between arguments in the same proportion, 2. considering only distributional similarity between event heads, and 3. considering only distributional similarity between arguments. All of them include the temporal clustering since it is impossible for two events to be coreferent if they occur at different times.

Regarding cross-document event ordering, the different experiments have been evaluated under the framework proposed at SemEval-2015 Task 4 Track B. Results show that timeline creation is a very challenging task (Best F1-Score of 30.43%) but with our approach we are outper-

forming the results of the state-of-the-art systems (+12.09 points than HEIDELTOUL and +5.09 points than GPLSI in F1-score) and we consider that combining temporal inference with distributional semantic methods is a feasible approach to tackle the event ordering task.

Therefore, with this corpus and these data we can conclude first that fully distributional models are suitable for the event ordering task; second that merely the distributional vector of the event head is enough to represent the distributional meaning of the event structure and, finally, that for this specific evaluation framework, there are not significant differences between Topic Modeling and Word2Vec Word Embeddings in these tasks, as the way they are currently developed.

As Future Work, we plan to compare our system with the state-of-the-art event coreference systems using ECB or ECB+ corpora. Furthermore, we want to analyze if there is some relation between the kind of event and its event structure, and a further study of another distributional models.

References

Bagga, A. and B. Baldwin. 1999. Cross-document event coreference: Annotations, experiments, and observations. In *In Proc. ACL-99 Workshop on Coreference and Its Applications*, pages 1–8.

Baroni, M., G. Dinu, and G. Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, June. Association for Computational Linguistics.

Bejan, C. A. and S. Harabagiu. 2014. Unsupervised Event Coreference Resolution. *Computational Linguistics*, 40(2):311–347.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Caselli, T., A. Fokkens, R. Morante, and P. Vossen. 2015. SPINOZA_VU: An NLP Pipeline for Cross Document TimeLines. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 787–791, Denver, Colorado, June. Association for Computational Linguistics.

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, pages 41–71.

Cybulska, A. and P. Vossen. 2013. Semantic relations between events and their time, locations and participants for event coreference resolution. In *RANLP*, pages 156–163. RANLP 2011 Organising Committee / ACL.

Goyal, K., S. K. Jauhar, H. Li, M. Sachan, S. Srivastava, and E. H. Hovy. 2013. A structured distributional semantic model for event co-reference. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers*, pages 467–473. The Association for Computer Linguistics.

Ji, H., R. Grishman, Z. Chen, and P. Gupta. 2009. Cross-document event extraction and tracking: Task, evaluation, techniques and challenges. In *RANLP*, pages 166–172. RANLP 2009 / ACL.

Landauer, T. K. and S. T. Dumais. 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

Laparra, E., I. Aldabe, and G. Rigau. 2015. Document level time-anchoring for timeline extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015)*, Beijing, China.

Lee, H., M. Recasens, A. Chang, M. Surdeanu, and D. Jurafsky. 2012. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 489–500, Stroudsburg, PA, USA. Association for Computational Linguistics.

Li, P., Q. Zhu, and X. Zhu. 2011. A clustering and ranking based approach for multi-document event fusion. In *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2011 12th ACIS International Conference on*, pages 159–165, July.

Llorens, H., E. Saquete, and B. Navarro-Colorado. 2012. Automatic System for Identifying and Categorizing Temporal Relations in Natural Language. *International Journal of Intelligent Systems*, 27(7):680–703.

Lu, J. and V. Ng. 2016. Event coreference resolution with multi-pass sieves. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Manning, C. D., M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, , and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS 2013)*, volume 26. pages 3111–3119.

Minard, A.-L., M. Speranza, E. Agirre, I. A. adn Marieke van Erp, B. Magnini, G. Rigau, and R. Urizar. 2015. Semeval-2015 task 4: Timeline: Cross-document event ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 778–786. Association for Computational Linguistics.

Mitchell, J. and M. Lapata. 2010. Composition in Distributional Models of Semantics. *Cognitive Science*, 34:1388–1429.

Moulahi, B., J. Strötgen, M. Gertz, and L. Tamine. 2015. Heideltoul: A baseline approach for cross-document event ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 825–829, Denver, Colorado, June. Association for Computational Linguistics.

Navarro-Colorado, B. and E. Saquete. 2015. GPLSIUA: Combining Temporal Information and Topic Modeling for Cross-Document Event Ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 820–824, Denver, Colorado, June. Association for Computational Linguistics.

Navarro-Colorado, B. and E. Saquete. 2016. Cross-document event ordering through temporal, lexical and distributional knowledge. *Knowledge-Based Systems*, 110:244–254, October.

Palmer, M., D. Gildea, and P. Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31.

Saurí, R., J. Littman, R. Knippen, R. Gaizauskas, A. Setzer, and J. Pustejovsky, 2006. *TimeML Annotation Guidelines 1.2.1 (http://www.timeml.org/)*.

Sun, W., A. Rumshisky, and O. Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. In *J Am Med Inform Assoc.*, pages 806–13, September-October.

UzZaman, N., H. Llorens, L. Derczynski, J. Allen, M. Verhagen, and J. Pustejovsky. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA. ACL. ISBN: 978-1-937284-49-7.

Verhagen, M., R. Gaizauskas, M. Hepple, F. Schilder, G. Katz, and J. Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 75–80, Prague. ACL.

Verhagen, M., R. Saurí, T. Caselli, and J. Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden. ACL.

Yang, B., C. Cardie, and P. I. Frazier. 2015. A hierarchical distance-dependent bayesian model for event coreference resolution. *TACL*, 3:517–528.

Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 69-76 recibido 30-11-2016 revisado 13-01-2017 aceptado 08-02-2017

Merging Multiple Features to Evaluate the Content of Text Summary

Combinación varias Caracteristicas para evaluar el contenido del resumen de texto

Samira Ellouze, Maher Jaoua, Lamia Hadrich Belguith
ANLP-RG, MIRACL Laboratory, FSEG Sfax, University of Sfax, Sfax, Tunisia
{Samira.Ellouze, maher.jaoua, l.belguith}@fsegs.rnu.tn

Abstract: In this paper, we propose a method that evaluates the content of a text summary using a machine learning approach. This method operates by combining multiple features to build models that predict the PYRAMID scores for new summaries. We have tested several single and "Ensemble Learning" classifiers to build the best model. The evaluation of summarization system is made using the average of the scores of summaries that are built from each system. The results show that our method has achieved good performance in predicting the content score for a summary as well as for a summarization system.
Keywords: Text summary, content summary evaluation, machine learning

Resumen: En este artículo proponemos un método que evalúa el contenido de un resumen de texto utilizando un enfoque de aprendizaje automático. Este método funciona combinando múltiples Caracteristicas para construir modelos que predicen las puntuaciones PYRAMID para nuevos resúmenes. Hemos probado varios clasificadores individuales y "Ensemble Learning" para construir el mejor modelo. La evaluación del sistema de resumen se realiza utilizando el promedio de las puntuaciones de los resúmenes que se construyen a partir de cada sistema. Los resultados muestran que nuestro método ha logrado un buen rendimiento en la predicción de la puntuación de contenido para un resumen, así como para un sistema de resumen.
Palabras clave: Resumen del texto, Evaluación de resumen de contenido, aprendizaje automático

1 Introduction

In recent years, several automatic summary systems have been developed. The evaluation of these systems is important to determine their ability to perform the assigned summarization task. It is in this context that several studies have been conducted to develop evaluation metrics which are applicable to manual and/or automatic summarization. However, and in order to have a common data set, several evaluation conferences such as SUMMAC, DUC (Document Understanding Conference), TAC(Text Analysis Conference), etc., were held to evaluate the performance of summaries generated automatically. For instance, the TAC conference adopted three manual measures, namely PYRAMID (content score), readability (linguistic quality) and overall responsiveness (score that reflects both content and linguistic quality of text summary) to assess the quality of text summary. Most metrics developed in the field of automatic evaluation of content summaries address the assessment using a surface analysis (lexical or syntactic) since a deep analysis that affects the syntactic and the semantic level requires meta-knowledge for modeling the contents of text summary. It is in this context that we have targeted as a field of study the evaluation of content summary while trying to address some aspects of syntactic and semantic level. So the objective is to build models able to predict manual content metric by combining automatic metrics and features defined on the candidate summary(CS). The choice of combining these features as a strategy has a number of advantages. For instance, one can benefit from the use of content features that operate on different levels of analysis. The combination of features is performed using algorithms based

ISSN 1135-5948

on regression techniques. The remainder of this article is structured as follows. In Section 2, we give an overview of the principal works that have addressed the problem of content summary evaluation. Then in Section 3, we describe the proposed method which operates by means of machine learning techniques. In Section 4, we give the details of each machine learning step. In Section 5, we present our experiments and the obtained results.

2 Previous Works

The summary evaluation task started as a manual and time-consuming evaluation. One of the famous metrics of content summary evaluation is PYRAMID (Nenkova and Passonneau, 2004) which is based on identifying the common ideas between a candidate summary and one or several reference summaries. These ideas are represented as semantic information units called "Semantic Content Units (SCUs)". Because of the time required to evaluate summaries with manual metrics, many studies are conducted to find ways to automatically assess the content of the summary. One of the standards in automatic evaluation is ROUGE (Lin, 2004). It measures overlapping content between a candidate summary and reference summaries. ROUGE metric scores are obtained through the comparison of common words: N-grams. In order to circumvent the limitations of ROUGE metric (Hovy et al., 2006) proposed a new metric called BE (Basic Elements) which is based on the decomposition of each sentence in minimum semantic units called "Basic Elements"(BE). This metric calculates the overlap between a candidate summary and reference summaries using BE units. Later, Giannakopoulos et al. (2008) introduced AutoSummENG metric, which is based on statistical extracting of textual information from the summary. The information extracted from the summary, represents a set of relations between n-grams in this summary. The n-grams and the relations are represented as a graph where the nodes are the N-grams and the edges represent the relations between them. The calculation of the similarity is performed by comparing the graphs of the candidate summary with the graph of each reference summary. Afterwards, the SIMetrix measurement was developed by (Louis and Nenkova, 2013); it assesses a candidate summary by comparing it with the source documents. The SIMetrix computes ten measures of similarity based on the comparison between the source documents and the candidate summary. Among the used similarity measures we cite the cosine similarity, the divergence of Jensen-Shannon(JS), etc. Recently, Cohan (2016) have developed the SERA (Summarization Evaluation by Relevance Analysis) metric, which is designed to evaluate scientific articles. This metric relies on relevant content in common between a candidate summary and reference summaries. Cohan (2016) use an information retrieval based method which treats summaries as search queries and then measures the overlap of the retrieved results.

3 The Proposed Method

The basic idea of the proposed evaluation methodology is based on the prediction of the manual score PYRAMID for a candidate summary. This prediction is obtained by the extraction of features from the candidate summary itself, from comparing the candidate summary with the source documents or with reference summaries. The choice of the prediction of PYRAMID score is motivated by its importance on the one hand and their availability in the manual evaluations of the DUC and TAC evaluation conferences, on the other hand. Since PYRAMID is based on the manual extraction of SCUs by human judges, SCUs cannot be identified from a summary that does not have a good linguistic quality. Thus, it is interesting to include linguistic features to ensure a better prediction of the PYRAMID score.To get the best prediction model, we tried to combine the relevant traits by using multiple regression-based algorithms. In the next section, we will detail the machine learning phase, which represents the mainstay of the proposed method.

4 Machine learning phase

4.1 Features extraction

This first step identifies for each summary the values of all the features. In order to calculate some features related to linguistic quality, we have to use various natural language processing tools such as the Stanford parser (Klein and Manning, 2003), the Stanford Tagger (Toutanova et al., 2003), the Stanford NER (Finkel, Grenager, and Manning, 2005), the Stanford Coref (Lee et al., 2011), the srilm toolkit (Stolcke, 2002), etc. In this

phase we use some new features and other features that are successfully used in the assessment of content. For the linguistic features that have been used, we have tried to cover many linguistic aspects (e.g. grammaticality, non-redundancy, Structure and coherence, etc). In this work, we have included all the classes of features that were used in (Ellouze, Jaoua, and Hadrich Belguith, 2013) and (Ellouze, Jaoua, and Hadrich Belguith, 2016): traditional readability measure features, shallow features, language modeling features, part-of-speech(POS) features, syntactic features, Named Entity based features, local coherence features, ROUGE/BE scores, AutoSummENG scores and Adapted ROUGE scores. Table 1 and Table 2 gives respectively the list of content and of linguistic quality features used in (Ellouze, Jaoua, and Hadrich Belguith, 2013) and in (Ellouze, Jaoua, and Hadrich Belguith, 2016). Furthermore, we have added the features cited subsequently.

4.1.1 Shallow features

We have added to the shallow features cited in Table 3 a set of lexical diversity features which are based on Type/token ratio where tokens refer to the number of words in a summary and types refer to the number of distinct words in a summary. A high score of these features can ensure that the sentences of a summary are less repetitive and have a rich vocabulary. In addition, we have determined for each candidate summary (CS) features based on paragraph length since a short paragraph can be more easily understood and can have fewer problems of co-referencing. Table 3 gives the list of added features.

4.1.2 Part-of-Speech features

We have added same POS features which are related to nouns and verbs which are the most important and essential part of content words for a text summary. This is because a summary must content less description details (i.e., less adjectives and adverbs) and more important actions expressed by nouns and verbs. The added features which are calculated for a CS are cited in the Table 4.

4.1.3 SIMetrix scores features

We have used all the ten scores calculated by SIMetrix (Louis and Nenkova, 2013) such as the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between the source documents (SDs) and the CS (KLInputSummary), the KL divergence between the CS and the SDs (KLSummaryInput), the unsmoothed version of Jensen Shannon divergence (Lin, 1991) between the SDs and the CS (unsmoothedJSD) and the smoothed one (smoothedJSD), the cosine similarity between the SDs and the CS (cosineAllWords), the percentage of the descriptive words of the SDs that appear in the CS (percentTopicTokens), the percentage of the CS composed of the more descriptive words from the SDs. (fractionTopicWords"fTW"), the cosine similarity between the CS and the most descriptive words in the SDs (topicWordOverlap), the probability of uni-grams of the CS given SDs (unigramProb), multinomial probability of the CS given SDs (multinomialProb).

4.1.4 Coreference Features

We have used the "Stanford Coref"(Lee et al., 2011) to allows us identify the different co-reference relations in a summary and the sentences where the co-reference and its antecedent are. From those pieces of information, we have extracted the number of times a pronoun has no antecedent (CorefWithoutAnt), the number of times a pronoun has antecedent (corefWithAnt), whether its antecedent is in the current sentence (AntSameSent), in the previous sentence (AntPrevSent)or not in the same sentence or in the previous sentence (AntOtherSent). In addition, we have determined the ratio between the number of co-references without antecedent to the total number of co-references with antecedent (RatWithAntWithoutAnt) and vice versa (RatWithoutAntWithAnt), the number of pronouns without antecedent to the total number of words (RatWithoutAntNbWord) and the number of pronouns without antecedent to the total number of pronouns (RatWithoutAntNbPron).

4.1.5 Redundancy features

To calculate these features, we compared each sentence in the CS with the other sentences by using a lexical similarity measure. For each measure of similarity, the average similarity between sentences and the average maximum(Max) similarities between each sentence and other sentences of the CS were determined. The following features are calculated for each CS: AVG and Max redundancy with DICE coefficient (RedondAVGdice, RedondMaxDice), with overlap coefficient (RedondAVGover, RedondMaxO-

Feature	Description
ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-SU4, ROUGE-L and ROUGE-W	ROUGE score based on respectively uni-grams, bi-grams, tri-grams, four grams, skip-bigrams and uni-grams, Longest Common Subsequence of n-grams and Weighted Longest Common Subsequence of n-grams
BE	Score based on syntactic units called BE (Basic Elements)
ROUGE-1_{Ad}, ROUGE-2_{Ad}, ROUGE-3_{Ad}, ROUGE-4_{Ad}, ROUGE-5_{Ad}, ROUGE-L_{Ad}, ROUGE-$S4_{Ad}$ and ROUGE-W_{Ad}	ROUGE adapted score based on respectively uni-grams, bi-grams, tri-grams, four grams, five grams, Longest Common Subsequence of n-grams, skip-bigrams and Weighted Longest Common Subsequence of n-grams
AutoSummENG-W_{123}, AutoSummENG-W_{333}, AutoSummENG-W_{253}, AutoSummENG-C_{123}	AutoSummENG with n-grams of words of length between respectively [1..2], [3..3], [2..5] and of characters of length between [1..2] with window size of 3 for all used variants

Table 1: List of content features used previously

Feature	Description
NbDET, NbCC, NbPSC, NbPRP, NbN, NbV, NbADJ and NbADV	Number(NB) of respectively determinants, coordinating conjunctions, prepositions and subordinating conjunctions, personal pronouns, nouns, verbs, adjectives (ADJ) and adverbs(ADV)
AVgDET, AVgCC, AVgPSC, AVgPRP, AVgADJ, AVgV, AVgN, AVgADV	Average(AVG) NB of determinants, coordinating conjunctions, prepositions and subordinating conjunctions, personal pronouns, adjectives, verbs, nouns and adverbs per sentence
Dens_DET, Dens_CC, Dens_PSC, Dens_PRP, Dens_ADJ, Dens_V, Dens_N and Dens_ADV	Density of respectively determinants, coordinating conjunctions, prepositions and subordinating conjunctions, personal pronouns, adjectives, verbs, nouns and adverbs
FleschK_Ind, FleschR_Ind, Aut_Read_Ind and Gun_Fog_Ind	Readability measures of respectively Flesch-Kincaid Index, Flesch Reading Ease, Automated Readability Index and Gunning Fog Index
AVGSyllWord, AVGCarWord,	AVG NB of respectively syllables, characters per word
AVGWordSent	AVG NB of words per sentence
RatWordMaxWord	Ratio between CS size and maximum size allowed by TAC campaign
logSent, logCar, logWord	Logarithm of the NB of respectively sentences, characters and words
AvgNPsent, AvgVPsent, AvgPPsent and AvgSBARsent	AVG NB of respectively noun phrases, verb phrases, prepositional phrases and clauses per sentences
NbNP, NbVP, NbPP and NbSBAR	NB of respectively noun phrases, verb phrases, prepositional phrases and clauses
AVG_Height_PT	AVG height of the parse tree
AVG_NB_dep_sent	AVG NB of dependency relations by sentence
logProbUnigram, logProbBigram, logProbTrigram	Log probability of respectively uni-grams, bi-grams and tri-grams of the CS
pplUnigram, pplBigram, pplTrigram	Measure of perplexity for respectively unigrams, bi-grams and tri-grams normalized by the NB of uni-grams, bi-grams and tri-grams
ppl1Unigram, ppl1Bigram and ppl1Trigram	Measure of perplexity for uni-grams, bi-grams and tri-grams with exclusion of the sentence end tags
NbEnt, DensEnt and AvgEntSent	NB, Density and AVG of Named entities in the candidate summary
AVGLevenDist, AVGCosSim, AVGJacSim, AVGJSDiver, AvgKLdiv, AVGPearCor, AVGdiceInd, AVGoverlapCoef	the AVG of respectively Levenshtein distance, cosine similarity, Jaccard distance, divergence of JS, Kullback-Leibler divergence , Pearson correlation, Dice index and overlap coefficient between adjacent sentences

Table 2: List of linguistic quality features used previously

Feature	Description
Nb_DistWord	NB of distinct words
TTR	Type/token ratio
Rac_Dens_DistW	Root of the density of distinct words
Dens_Corr_DistW	Correct density of distinct words
Bilog_Dens_DistW	Bi-logarithmic density of distinct words
Uber_Index	Uber index
AVGSentParag	AVG NB of sentences per paragraph
AVGWordParag	AVG NB of words per paragraph
Dens_stopWords	Density of stop words

Table 3: List of added Shallow features

ver), with Jaccard index (RedondAVGjacc, RedondMaxJacc) and with cosine similarity (RedondAVGcos, RedondMaxCos).

The content features cited in Table 1 and the SIMetrix scores have proved their usefulness in the field of text summary evaluation (Lin, 2004), (Hovy et al., 2006), (Giannakopoulos et al., 2008) and (Louis and Nenkova, 2013). In addition, most linguistic quality features cited previously have shown their utility in the assessment of the content (Ellouze, Jaoua, and Hadrich Belguith, 2013) and the linguistic quality (Ellouze, Jaoua, and Hadrich Belguith, 2016), (Pitler, Louis, and Nenkova, 2010), etc. While for some other features we have tried to test their performance (non-redundancy, coreference, etc).

4.2 Selection of relevant features

This step allows us to select the most relevant features that must be kept for the training step. In general, the selection of relevant features is as important as the choice of the learning algorithm. To select the relevant features, we use the "wrapper"method (Kohavi

Feature	Description
Dens_V_N, Rat_N_V and AVG_N_V	Density, Ratio and AVG of verbs and nouns
Rat_NV_AdjAdv	Ratio between the NB of nouns and verbs and the NB of ADJs and ADVs
Rat_InfV_V, Rat_ImpV_V, Rat_PartV_V and Rat_ModV_V	Ratio between the NB of respectively infinitive, imperative, participle and model verbs, and the total NB of verbs

Table 4: List of Added POS Features

and John, 1997) which is based on the evaluation of subsets of features which allows to detect the possible interactions between features. After training models using each subset, the best subset of features is retained. Using the "wrapper"method, we have obtained the relevant features for the best predictive model, in each evaluation task.

4.3 Training and Validation of the Predictive Model

This step helps to build and validate the predictive model of the PYRAMID score. To build the predictive model, we have used several basic algorithms (single algorithms), implemented by the Weka environment (Witten, Frank, and Hall, 2011), using a regression method such as "GaussianProcesses". Moreover, we tried to produce models by using the "ensemble learning" which usually produces more accurate solutions than a basic learning algorithm. In our experiment, we use three "ensemble learning" algorithms which are implemented in the Weka environment:

- "Bagging" (Breiman, 1996) divides the training data into separate samples. Then it creates a model for each sample with the same algorithm. Next, it aggregates the generated models using averaging or majority voting
- "Vote"(Kuncheva, 2004) allows the combination of several predictive models trained on the same dataset using a combination rule like "Majority Voting".
- "Stacking"(Wolpert, 1992) combines several models (made from different basic learning algorithm) that are learned from a classification or a regression task using the same dataset. The combination of the constructed models is made using a machine learning algorithm.

After testing the algorithms, we adopt the one that produces the best predictive model. The validation of each model is performed by cross-validation method with 10 folds.

5 Experimentations

We experimented our method for summary level evaluation on initial summary task (task A) and update summary task (task B) by trying to predict PYRAMID scores. On the system level, we will just average the predicted scores of all the candidate summaries produced by the same summarization system.

5.1 Data Set

The Data Set used in the study consists of the source documents, the manual summaries (reference summaries) and the system summaries presented in the TAC 2008 conference on the update summarization task. This task includes two subtasks, initial summary task and update summary task. In initial summary task, each summarization system had to summarize a set of documents (A) which deals with a particular event. Then, in update summary task, it should summarize a set (B) of documents which addresses the evolution of the same event and considers the knowledge of the set (A). This corpus includes 48 collections, each collection contains a set (A) and a set (B) of documents. Moreover, it includes 2784 (58*48) system summaries that are automatically generated from the set (A) of the 48 collections and by the 58 participating systems, in initial summary task. And 2784 system summaries in update summary task. The corpus also includes reference summaries produced manually by 8 human summarizers. For each collection, 4 reference summaries are produced for set (A) and 4 reference summaries are produced for set (B). In total, 384 (96 * 4) reference summaries. Thus, each system summary can be assessed by comparing the four reference summaries. Similarly, a reference summary can be evaluated by comparing it with the other three reference summaries. Furthermore, the corpus contains the PYRAMID and the linguistic quality of each reference and system summary. The linguistic quality score is an integer between 1 and 5 which reflects five linguistic qualities. In our experiments in summary level evaluation, each model is pro-

Features	Initial Summary
Content	ROUGE-1, ROUGE-2, ROUGE-3 ROUGE-4, ROUGE-SU4, ROUGE-W, AutoSummENG_w333, AutoSummENG_w123, AutoSummENG_w253 KLInputSummary, KLSummaryInput, unsmoothedJSD, smoothedJSD, cosineAllWords, percentTopicTokens, fractionTopicWords, TopicWordOverlap, unigramProb, multinomialProb
Linguistic quality	NbDET, NbPSC, Dens_DET, Dens_N, Dens_V_N, Uber_Index, AvgSBARsent, AvgPPsent, NB_SBAR,AVG_Height_PT, AVG_NB_dep_sent, logProbUnigram, logProbBigram, NbEnt, AvgKLdiv, AntPrevSent, RatWithoutAntNbWord, RedondAVGdice

Table 5: List of Selected Features to Predict Content Score for initial summary task

duced using 2976 CSs where 2784 are system summaries and 192 are reference summaries.

5.2 Evaluation

5.2.1 Summary level

In this subsection, we begin by citing in Table 5 the selected features for the prediction of content score in initial summary task. From Table 5, we remark the selection of most content scores in addition to many linguistic quality features. We have observed the presence of features related to reference clarity and redundancy (AntPrevSent, RedondAVGdice). This means that when evaluating the content, we need to have a candidate summary with clear reference resolution and without redundancy. In addition, we remark the presence of Language modeling features which can be indicators of the fluency and the grammaticality (logProBigram) of a text summary. Now, we give in Table 6 the list of used features in update summary task. From this table, we remark that also in update summary task (task B), many linguistic quality features are selected as relevant ones. Besides, the importance and the necessity of including linguistic quality features is clearly shown in update summary by the use of features related to diverse aspects of linguistic quality like referential clarity (RatWithoutAntNbSent), non-redundancy (RedondMaxDice, RedondMaxOver), etc. We examine the usefulness of the selected features in the prediction of the content score by training them using single and "ensemble learning". The Pearson's correlation (Pearson, 1895) and the RMSE generated by each classifier are presented in Table 7. In fact, the RMSE (Root Mean Square Error) is a measure that determines the differences between score values predicted by a model and the actual score values (in our case PYRAMID manual score). Table 7 shows the performance of the selected features in building models using several single and ensemble of classifiers in the initial and update summary tasks. In the initial summary task, the results show that the model built from the classifier "GaussianProcesses" produced the best correlation(0.769) and the lowest RMSE(0.1189). In the update summary level, Table 7 indicates that the best "ensemble learning" classifier is the "Vote" which provides a model having a correlation of 0.8063 and an RMSE of 0.1128. Another notable observation is that the correlation in the update summary task is more important than the one in the initial summary task.

We pass now to the comparison between the performance of the best obtained model and the baseline metrics that were adopted by the TAC conference as baseline metrics

Features	Update Summary
Content	ROUGE-1 ROUGE-2, ROUGE-4, ROUGE-SU4, ROUGE-L, ROUGE-W, ROUGE-BE, ROUGE-3_{Ad}, ROUGE-4_{Ad}, ROUGE-5_{Ad}, ROUGE-$S4_{Ad}$, AutoSummENG_W_{123}, KLInputSummary, KLSummaryInput, unsmoothedJSD, smoothedJSD, cosineAllWords, percentTopicTokens, topicWordOverlap
Linguistic quality	NbCC, NbV, AVgPSC, AVgV, Dens_CC, Dens_PRP, Dens_V, Dens_V, Dens_ADV, Dens_V_N, Rat_InfV_V, Rat_ImpV_V, Rat_PartV_V, Rat_ModV_V, AVGSyllWord, AVGCarWord, AVGSentParag, RatWordMaxWord, Dens_DistWord, Rac_Dens_DistW, Bilog_Dens_DistW, logSent, logCar, logWord, AvgNPsent, AvgPPsent, AVGCosSim, AVGJSDiver, AVGdiceInd, RatWithoutAntNbSent, RedondMaxDice, RedondMaxOver, RedondMaxCos, RedondAVGcos

Table 6: List of Selected Features to Predict Content Score for Update Summary

Classifiers	Task A	Task B
Single classifiers		
GaussianProcesses	**0.7690(0.1185)**	0.7965(0.1156)
LinearRegression	0.7421(0.1241)	0.7416(0.1270)
SMOReg	0.7391(0.1250)	0.7972(0.1155)
MultiPerceptron	0.7079(0.1311)	0.7111(0.1370)
"Ensemble learning" classifiers		
Vote	0.7470(0.1231)	**0.8063(0.1128)**
Bagging	0.7424(0.1240)	0.8009(0.1142)
Stacking	0.7453(0.1234)	0.8052(0.1130)

Table 7: Pearson Correlation with PYRAMID and RMSE (between brackets) for Various Single and Ensemble learning Classifiers

Scores	Task A	Task B
Baselines		
ROUGE-2	0.5990(0.1482)	0.5830(0.1548)
ROUGE-SU4	0.5090 (0.1399)	0.6205(0.1495)
BE	0.4493(0.1653)	0.5540(0.1587)
AutoSummENG_W_{123}	0.5405(0.1557)	
AutoSummENG_C_{333}		0.6487(0.1451)
SIMetrix_fTW	0.3382(0.1742)	0.3389(0.1793)
Our experimentations		
Combining ROUGE Scores	0.6075(0.147)	0.6440(0.1458)
Combining AutoSummENG scores	0.6841(0.135)	0.6134(0.1505)
Combining SIMetrix scores	0.4648(0.1639)	0.3594(0.1779)
Combining content scores	0.7330(0.1260)	0.7570(0.1248)
Combining selected features (CSF)	**0.7690(0.1185)**	**0.8063(0.1128)**
CSF without ROUGE/BE	0.759(0.1207)	0.7797(0.1194)
CSF without AutoSummENG	0.7631(0.1198)	0.7997(0.1145)
CSF without SIMetrix	0.7414(0.1243)	0.7919(0.1164)
CSF without new features	0.7532(0.1219)	0.7510(0.1260)

Table 8: Pearson Correlation with PYRAMID Score and RMSE (between brackets)for Summary Level

such as R-2, R-SU4 and BE and also we add the best variante of each of the two others famous metrics AutoSummENG and SIMetrix. Table 8 details the different correlations and RMSEs of baseline metrics and our experiments. It should be noted that, in the initial summary task, the models built in our experiments use all the "GaussianProcesses" classifier. In addition, we note that in the update summary task, the models built in our experiments use all the "vote" ensemble learning. From Table 8 and in both tasks, we see the gap between baseline metrics and our experiments, regardless of whether we used the selected features or just content scores. Moreover, we noticed that the inclusion of linguistic quality features in the best model produced improves the performance of this model compared to the model containing just content scores. We note also that the elimination of the new added features in this article, decreases the correlation between the predictive score and PYRAMID score. Furthermore, we find that the elimination of one of the content score classes, reduces the correlation of the predictive score with PYRAMID score.

5.2.2 System Level

Remember that the system level evaluation allows us to estimate the quality of a summarization system; in other words, the system assessment is done by taking into account the quality of all the summaries that are produced by this system. In this article, we tried to calculate the quality of a system $Score_{system}$ by determining the average of the predicted score for summaries produced by the same system. To evaluate this method of calculating the content score for a system, we study the correlation of Pearson"P", Spearman"S" (Spearman, 1910) and Kendall"K" (Kendall, 1938) between the PYRAMID score and the $Score_{system}$ score. Indeed, those correlation measures have been used in the DUC and the TAC conferences to determine the correlation between automatic and manual evaluation metrics. Table 9 details the different correlations between the PYRAMID score and the $Score_{system}$ score or the baseline metrics.In this evalution level, we use as baselines, ROUGE-2, ROUGE-SU4, BE, AutoSummENG_W_{123} and SIMetrix_fractionTopic. As can be seen in Table 9, the best correlation is obtained by our $Score_{system}$. It has the best correlation with the PYRAMID score in both tasks and with the three types of correlation measures.

6 Conclusions and Future Works

In this paper, we presented a method of content evaluation for text summaries. Our work has been motivated by the lack of efficient and accurate automatic tools that evaluate the content of a summary. The proposed method is based on the construction of models that combine selected features which come from multiple feature classes such as ROUGE scores, SIMetrix scores, modeling language features, Syntactic features, etc. The combination of features is performed by testing many single and "ensemble learning" classifiers. Then, we have selected the best algorithm for the prediction of the PYRAMID score. At the initial summary level and in order to evaluate the predictive power of the model constructed using the selected features to predict content score, we have compared the correlation of this model with baselines and with a model containing only content scores. In both tasks, the obtained results show that there is an important gap between baselines and the model combining selected features. We also note that adding linguistic quality features to a model predicting PYRAMID, improves the results.

In system level evaluation, for a specific task and a predicted content score "$Score_{system}$", we have calculated the average of the predicted score values of all the summaries that were built from the same summarization system. In both tasks, the average

	P	S	K	P	S	K
Scores	**Initial Summary**			**Update Summary**		
ROUGE-2	0.8718	0.9364	0.8050	0.9009	0.9588	0.8322
ROUGE-SU4	0.8741	0.9007	0.7477	0.8458	0.9323	0.7796
BE	0.9188	0.9329	0.7889	0.9188	0.9560	0.8297
AutoSummENG_W_123	0.9051	0.9336	0.7946	0.8955	0.9626	0.8384
SIMetrix_fTW	0.5523	0.7764	0.5922	0.4160	0.6298	0.4570
$Score_{system}$	**0.9950**	**0.9761**	**0.8901**	**0.9964**	**0.9866**	**0.9204**

Table 9: Pearson, Spearman and Kendall Correlation with PYRAMID Score on System Level

of the predicted content scores of each system "$Score_{system}$" correlates the best with the PYRAMID score.

As futur work, we project to apply this method of building models to other manual scores like the overall responsiveness score. In addition, we aim to add same features related to semantic level.

References

Breiman, L. 1996. Bagging predictors. *Machine learning*, 24:123–140.

Cohan, A, G. N. 2016. Revisiting summarization evaluation for scientific articles. In *Proceedings of LREC conference*, pages 806–813.

Ellouze, S., M. Jaoua, and L. Hadrich Belguith. 2013. An evaluation summary method based on a combination of content and linguistic metrics. In *Proceedings of RANLP conference*, pages 245–251.

Ellouze, S., M. Jaoua, and L. Hadrich Belguith. 2016. Automatic evaluation of a summary's linguistic quality. In *Proceedings of NLDB 2016 conference*, pages 392–400.

Finkel, J. R., T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the Annual Meeting on ACL*, pages 363–370.

Giannakopoulos, G., V. Karkaletsis, G. Vouros, and P. Stamatopoulos. 2008. Summarization system evaluation revisited: N-gram graphs. *ACM Trans. Speech Lang. Process.*, 5(3):1–39.

Hovy, E., C. Lin, L. Zhou, and J. Fukumoto. 2006. Automated summarization evaluation with basic elements. In *Proceedings of the LREC conference.*

Kendall, M. G. 1938. A new measure of rank correlation. *Biometrika*, 30:81–89.

Klein, D. and C. D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the Annual Meeting on ACL*, pages 423–430.

Kohavi, R. and G. H. John. 1997. Wrappers for feature subset selection. *Artif. Intell.*, 97(1-2):273–324.

Kullback, S. and R. A. Leibler. 1951. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86.

Kuncheva, L. I. 2004. *Combining pattern classifiers: Methods and Algorithms.* Wiley-Interscience.

Lee, H., Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of CoNLL conference*, pages 28–34.

Lin, C. 2004. Rouge: a package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL Workshop*, pages 25–26.

Lin, J. 1991. Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theory*, 37:145–151.

Louis, A. and A. Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

Nenkova, A. and R. J. Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *HLT-NAACL*, pages 145–152.

Pearson, K. 1895. Mathematical contributions to the theory of evolution, ii: Skew variation in homogeneous material. *Philosophical Transactions of Royal Society London (A)*, 186:343–414.

Pitler, E., A. Louis, and A. Nenkova. 2010. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the Annual Meeting of the ACL*, pages 544–554.

Spearman, C. E. 1910. Correlation calculated from faulty data. *British Journal of Psychology*, 3:271–295.

Stolcke, A. 2002. Srilm-an extensible language modeling toolkit. In *Proceedings of ICSLP*, pages 257–286.

Toutanova, K., D. Klein, C. D. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings OF HLT-NAACL*, pages 252–259.

Witten, I. H., E. Frank, and M. A. Hall. 2011. *Data mining: Practical machine learning tools and techniques.* Morgan Kaufmann Publishers Inc.

Wolpert, D. H. 1992. Stacked generalization. *Neural Networks*, 5:241–259.

Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 77-84 recibido 29-11-2016 revisado 10-01-2017 aceptado 07-02-2017

ANALHITZA: a tool to extract linguistic information from large corpora in Humanities research

ANALHITZA: herramienta para extraer información lingüística de corpus extensos para su uso en investigaciones de ciencias humanas

Arantxa Otegi[1], **Oier Imaz**[2], **Arantza Díaz de Ilarraza**[1], **Mikel Iruskieta**[1] y **Larraitz Uria**[1]
[1]IXA Group. University of the Basque Country, UPV/EHU
[2]PRAXIS Research Group. University of the Basque Country, UPV/EHU
arantza.otegi@ehu.eus, oier.imaz@ehu.eus, a.diazdeilarraza@ehu.eus, mikel.iruskieta@ehu.eus, larraitz.uria@ehu.eus

Resumen: El tamaño reducido de los corpus en ciertos campos de investigación se debe a la falta de herramientas para procesar el lenguage de forma masiva y sencilla. En este artículo presentamos ANALHITZA, una herramienta que estamos desarrollando dentro del proyecto Clarin-k que tiene como objetivo principal la creación de tecnologías lingüísticas útiles para la investigación en Ciencias Sociales y Humanidades. ANALHITZA ha sido diseñada para extraer información lingüística online de textos extensos de una forma sencilla. Además, es una herramienta multilingüe que permite analizar textos escritos en tres lenguas: euskera, castellano e inglés. En este artículo, a modo de ejemplo, presentamos tres estudios en los que se ha usado esta herramienta, que puede ser rediseñada para cubrir las necesidades de investigación de muchas de las ramas de Humanidades.
Palabras clave: Herramienta, tecnologías del lenguaje, corpus, análisis de texto, PoS

Abstract: The reduced size of corpora in some areas of research is due to the lack of tools to process massively and easily the language under study. In this article, we present ANALHITZA, a tool which is being developed within the Clarin-k project, whose aim is the creation of linguistic technologies that are useful for research on Social Sciences and Humanities. ANALHITZA has been designed to extract linguistic information online from large corpora in an easy way. Besides, it is a multilingual tool which can process texts written in three languages: Basque, Spanish and English. Moreover, we present three real examples of study where ANALHITZA has been used. The tool can be redesigned or changed, according to the needs of the scientific community in the field of Humanities.
Keywords: Tool, language technologies, corpora, text analysis, PoS

1 Introduction

How can Language Technology (LT) tools be applied in the Humanities research? How can these technologies help in, for example, getting accessible the needed corpora for such researches? Humanities projects are grounded in a dataset that, from a quantitative point of view, is typically used in some kind of statistical analysis to confirm or not a particular hypothesis which will be developed in the process of exploring the dataset. Usually, the size of the dataset used is reduced, because the analysis of bigger amounts of texts is not manually affordable. Having these aspects in mind, some key questions arise: are the researchers in Humanities aware of the possibilities offered by LT tools? Are the researchers in Natural Language Processing (NLP) ready to tackle the problems researchers have in the Humanities field?

There are several reasons why the researcher in Humanities avoid the use of LT tools: *i*) there are not available many tools which can analyze the linguistic phenomena in the language under study; *ii*) in case there is a tool, it may require economic costs or technical expertise to use it; *iii*) the output quality of the tools available cannot be compared to the results obtained by human annotation, or *iv*) the tool is unknown to the community.

Therefore, it is important to make avail-

ISSN 1135-5948

able to researchers in the Humanities and Social Sciences digital multilingual tools that can be easily chained to perform complex operations in order to support them in their work.

In this article, we present three preliminary studies we have been working on in the fields of Humanities and Social Sciences. These studies have been developed on the results produced by ANALHITZA, an application which provides users with linguistic information concerning written texts. Such information is based on an automatic morphosyntactic analysis, which is carried out using NLP tools. The application is still under development.

The article follows the subsequent structure: Section 2 presents some related work. Section 3 describes the system ANALHITZA. In Section 4 we present the results of three studies carried out using the tool. Finally, Section 5 sets out the conclusions and future work.

2 Related work

In the Virtual Language Observatory[1], created in the framework of CLARIN (Common Language Resources and Technology), we can find several tools for the automatic processing of language oriented to eHumanities as well as some interesting resources for different languages. General projects such as MetaShare,[2] developed in the context of MetaNet[3] and ELRA catalog,[4] offer an interesting and useful overview of collections of tools and linguistic resources for general purposes. AntConc[5] and LancsBox[6] are, for example, two interesting tools that provide an easy access to the results, but with the inconvenience that cannot be used online. Another useful tool is CONTAWORDS, an application presented in Villegas et al. (2012), who show that Language Resources and NLP can help in different researches in the Humanities.

In this way, content analysis is accessible with LT, because one can find some related or hidden semantic structures in a text body or check if the semantic structures of the language or knowledge fits with our predictions. Content analysis is aimed at data reduction (Alonso and Volkens, 2012), since texts are very complex and entail high degrees of variability in terms of linguistic expressions (Krippendorff, 2004). Thus, analysis begins with the application of several preprocessing techniques to reduce the complexity of 'texts as data' (Grimmer and Stewart, 2013). Depending on the method and the aimed results, one can use different approaches, to cite some: *a*) Topic Models (TM) erase any information about ordering (bag-of-words) reducing texts to lists of unique words (Blei, 2012) or *b*) Network Text Analysis (NTA), on the contrary, retains ordering to maintain the pattern of textual linkage between concepts in terms of their proximity (Carley, 1997).

The tool we present in this paper, ANALHITZA, aims to be helpful at least in the directions shown here with three different studies: *i*) a specific task of linguistic textual analysis in literature, *ii*) content analysis of transcripts of a deliberative exercise and *iii*) data manipulation to analyze the best indicators of the main topic in a multilingual corpus.

3 System description

ANALHITZA is a tool that, in a nutshell, processes text automatically and extracts some linguistic information concerning the analyzed text.

The in-house version of the system has a simple command-line interface that offers the possibility to pass a single document or a directory which could contain many documents to analyze. The online version, which is publicly available,[7] does not offer the option of analyzing more than one document at once. But using its simple interface (cf. Figure 1), the user can submit a text to be analyzed in one of the following three ways: *i*) uploading a plain text file, *ii*) writing the text in a text box or *iii*) specifying the URL of the website that contains the text. Both versions of the system are able to process texts in three different languages (Basque, English and Spanish). The user has to specify the language, submit the texts to be analyzed and the system will provide the results to the user in a spreadsheet (Excel file).

[1] https://www.clarin.eu/content/virtual-language-observatory.

[2] http://www.meta-share.org/.

[3] http://www.meta-net.eu/.

[4] http://catalog.elra.info/.

[5] http://www.laurenceanthony.net/software.html.

[6] http://corpora.lancs.ac.uk/lancsbox/.

[7] http://ixa2.si.ehu.es/clarink/analhitza.php?lang=en.

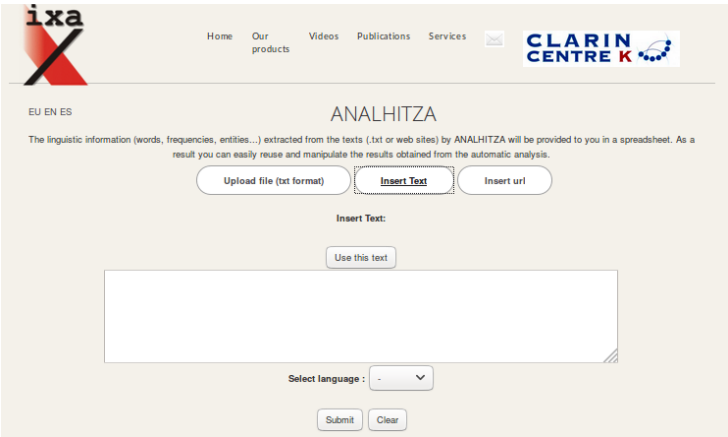

Figure 1: The interface of ANALHITZA

3.1 Automatic text processing

The input of the system is the text itself, which is analyzed and processed with some NLP tools such as tokenizer, lemmatizer, Part of Speech (PoS) tagger and Named Entity Recognizer and Classifier (NERC).

For that purpose, IXA pipes[8] and ixaKat[9] tools are used. IXA pipes is a modular set of NLP tools (or pipes) which provide easy access to NLP technology for several languages (Agerri, Bermudez, and Rigau, 2014). Similarly, ixaKat is a modular chain of NLP tools especifically created for Basque by means of hybridation techniques that combine knowledge and statistical approaches (Otegi et al., 2016). One of the main features of both set of tools is their modularity. That is, the tools in the pipe or in the chain can be picked and changed, as long as they read and write the required data format via the standard streams. All the tools in both sets read and write NAF format,[10] a linguistic annotation format designed for complex NLP pipelines (Fokkens et al., 2014). This way, it is possible the interaction between ixaKat and IXA pipes modules. In that way, Basque texts are analyzed using first a ixaKat tool, and afterwards a IXA pipe tool. Namely, the robust and wide-coverage morphological analyzer and PoS tagger from ixaKat (*ixa-pipe-pos-eu*) is linked to the NERC tool provided by the IXA pipes (*ixa-pipe-nerc*) (Agerri and Rigau, 2016).

Regarding the performance of these tools, all the tools used obtain state-of-the-art results. The Basque morphological analyzer which ixa-pipe-pos-eu is based on obtains 95.17% in accuracy on PoS tagging. The ixa-pipe-pos tool for English and Spanish lemmatization and PoS tagging obtains, respectively, 96.88% and 98.88% in accuracy. The NERC tool obtains a performance of 0.7672 (F1), 0.8621 (F1) and 0.8016 (F1) for Basque, English and Spanish, respectively.

Once the linguistic processing is carried out and based on the information in the output NAF document, some basic maths (such as counting different words, cases or entities, for examples) and filtering are applied.

3.2 Output

All the resulting data are compiled and presented in a spreadsheet to the user. The information is presented in several worksheets or sheet tabs (18 sheet tabs for Basque and 17 for Spanish and English).

In the first sheet tab, some general information is shown, including number of letters, words, lemmas, nouns, adjectives, verbs, adverbs, determiners, conjunctions, prepositions, named entities, sentences as well as the average sentence length and the number of words in the shortest and longest sentences.

In tab sheets from the second to eighth the lemmas of different nouns, adjectives, verbs, adverbs, determiners, conjunctions and prepositions found in the text are listed, with their respective frequency counts. The ninth sheet tab is only available for Basque texts and it shows the different declension cases found in the text. In fact, lemmatization is necessary to recognize the lemmas and the attached determiners (the indefinite singular $-a$ 'one' or the indefinite plural $-ak$

[8] http://ixa2.si.ehu.es/ixa-pipes/.
[9] http://ixa2.si.ehu.es/ixakat/.
[10] http://wordpress.let.vupr.nl/naf/.

'some') and/or the declension cases (ergative *−k*, absolutive *−a/ak*, dative *−(r)i* 'to', ablative *−tik* 'from', destinative *−tzat* 'for', inessive *−n* 'in' and genitive *−(r)en* ''s', among others).[11] In the tenth sheet tab, named entities are listed specifying their frequencies and also their classification-type (person, location or organization). In the following two sheet tabs, different lemmas and word forms (including all PoS tags) with their frequency counts are listed, respectively. Next, different alphabetic letters are listed in the thirteenth sheet tab. Tab sheets from fourteenth to sixteenth show 2-grams, 3-grams and 4-grams extracted from the text. The last two sheet tabs show the lemmatized (with PoS) and unformatted text, respectively.

Based on all that information, users can easily analyze the results and make conclusions in regard to the linguistic aspects of the text.

4 ANALHITZA in Humanities and Social Sciences

ANALHITZA offers users the possibility to extract linguistic information from large corpora in a very easy way, and it can be used to analyze any type of text in most of the disciplines related to Humanities and Social Sciences. For example, it is very useful for analyzing the linguistic characteristics of any text type, for comparing literary texts, news or students' essays written in same or different languages, for studying the language acquisition process of children or second language learners, for creating specialized dictionaries based on real corpora, for analyzing or even reducing the complexity of the texts, etc.

In this section, we briefly explain some experiments carried out using ANALHITZA to show, as example, how it can be exploited in different tasks: *i*) a comparative analysis of two Basque literary books, *ii*) a preprocessing task for content analysis on a bilingual oral corpus and *iii*) an experiment based on n-grams to identify expressions to detect the main topic of each text in a multilingual corpus.

[11] Basque is an ergative and an agglutinative language that constructs phrases by attaching free and bound morphemes (Hualde and de Urbina, 2003).

	Arrieta (2012)		**Alberdi (2013)**	
No of pages	159		139	
No of tales	8		9	
Av. words per tale	3,210		1,974	
No of words in all	25,677		17,765	
No of diff. words	7,793	30.35%	5,739	32.30%
No of diff. lemmas	4,150	16.16%	3,041	17.11%
No of verbs	8,856	34.49%	7,291	41.04%
No of nouns	9,229	35.94%	6,284	35.37%
No of adjectives	1,914	7.45%	1,055	5.93%
No of NE	622	2.42%	382	2.15%
No of decl. words	9,212	35.87%	6,725	
Av. words per sent.	9		7	
Words in longest sent.	97		52	

Table 1: Statistics concerning both books

4.1 Comparative linguistic analysis of two literary books

Because of the linguistic information this tool offers, we consider ANALHITZA a very suitable LT for text analysis. As example, we present a pilot comparative study of two literary books in Basque: *Alter Ero* (Arrieta, 2012) and *Euli Giro* (Alberdi, 2013).[12] Both books are composed of several tales and as they have very similar external characteristics (date of publication, genre, age and place of birth of the authors), we wanted to see whether they are also linguistically similar or not (because books having similar external features can be linguistically very different).

Analyzing the resulting data (cf. Table 1), we have been able to extract some conclusions in quite a fast and easy way (Table 1). For example, Arrieta's stories are a bit longer than Alberdi's (average of 3,210 vs. 1,974 words per tale). In Alberdi's book there are a bit more different lemmas than in Arrieta's, which shows that the lexicon in Alberdi's book is more varied than in Arrieta's.

We have seen that some of the most common lemmas are not content words, which has awaken our curiosity to verify whether in Basque prose there are, in general, more function words than content words (a larger corpora must be analyzed for that). Paying attention to the nouns (which are content words), the most common nouns do not coincide, and are, in general, varied in both books. This shows that each individual tale in the two books relates a different story. However, and although a deeper study is necessary to get more precise conclusions, the most common nouns in Alberdi's book (*ama* 'mum', *esku* 'hand', *andere* 'woman', *etxe*

[12] We want to thank the Susa publishing house for making available many literary works in digital support.

'house', *gizon* 'man') give quite a clear clue about what her stories are related to.

As regards the categories of the words, we have seen that Arrieta uses less verbs and more adjectives than Alberdi, which means that his stories are more descriptive and include less actions than Alberdi's tales, where occur more actions but things are described in less detail. We have found more different NEs in Arrieta's book whereas Alberdi has repeated the same NEs more times. This can be useful to analyze, for example, whether the stories happen to same characters and in same places or not. In this case, the scenarios change from tale to tale.

ANALHITZA also extracts information about letters and declensions cases. Vowels are more frequent than consonants in both books. The average of the declined words and the most common declension cases are very similar in both books (absolutive, inessive and genitive cases are the most common ones and the ergative is a little bit higher in Arrieta's book than in Alberdi's). But are these two facts also some intrinsic characteristics of Basque or just a coincidence?

The main aim of this first analysis was to see what kind of conclusions can be obtained using ANALHITZA. The clearest differences between both books are that one is a bit more descriptive than the other one, and that one contains longer tales than the other. Both conclusions can be useful, for example, for readers or teachers when selecting or recommending a book (the most descriptive one for those who prefer less action and vice versa; and the shortest tales for those who have more difficulties on reading).

In addition, the data obtained with ANALHITZA in this task have raised new questions about the linguistic characteristics, not only of the analyzed literary works but also of Basque literature and even of our language in general, characteristics that can be additionally compared with the main features of Spanish and/or English literary works. However, we have to continue analyzing larger corpora to obtain information and get to such conclusions.

	Basque		**Spanish**	
	No	**Diff.**	**No**	**Diff.**
Sentences	113		94	
Total of words	1126		1423	
Words/lemmas	680	438	580	475
Nouns	480	229	331	185
Adjectives	86	48	129	86
Verbs	312	103	255	111
Entities	34	17	22	11

Table 2: Statistics concerning both languages

4.2 Preprocessing tasks for content analysis in a bilingual corpus composed by political texts

Our aim in this second experiment is to show whether and to what extent ANALHITZA reduces the complexity of the analyzed corpus. Complexity reduction is a necessary step before other techniques for content analysis can be implemented in a text corpus. But, to our knowledge there is not any other tool to preprocess texts of a multilingual corpus including texts written in the Basque language.

The sample for this trial is composed of 40 short argumentative texts (20 in Basque and 20 in Spanish) written by citizens in a deliberative exercise named '*Konpondu*'[13] (CICIR, 2007; CICIR, 2009). The corpus consists of open-ended responses written by participants for oral presentations. We have randomly selected the sample set among those sharing a similar length (more than 300 characters), written in a similar date (April, 2008) and responding to the same question: *In the current situation which difficulties and opportunities do you see for peace and political normalization?*; although, in different towns.

These 40 texts were analyzed with ANALHITZA in two different clusters for each language. Table 2 shows the linguistic characteristics of texts written in two languages: number of elements (No) and different elements (Diff.).

At first glance, ANALHITZA seems to be very effective in terms of data reduction for both languages, but lemmatization is more efficient in Basque than Spanish as

[13]We want to thank Aitziber Blanco and Paul Rios from *Lokarri*, Igor Ahedo and Asier Blas from *Parte-Hartuz* (UPV/EHU) and Gorka Espiau and the *Agirre Lehendakaria Center* (`http://agirrecenter.eus/`) for helping us recollecting the documentation of the '*Konpondu*' initiative.

Difficulties		Opportunities	
Prepr.	**Post.**	**Prepr.**	**Post.**
que	politico	que	vez
de	**partido**	de	oportunidad
la	**violencia**	la	tener
y	ir	a	politico
a	dificultad	para	cada
el	sociedad	las	creer
en	tener	el	querer
no	existir	y	poder
se	poder	en	política
los	parte	una	decir

Table 3: Spanish most frequent word lists

Difficulties		Opportunities	
Prepr.	**Post.**	**Prepr.**	**Post.**
eta	**alderdi**	eta	bake
ez	politiko	da	aukera
da	eta	euskal	herri
alderdi	bake	aukera	gizarte
ere	lortu	behar	euskal
bakea	arazo	dut	eman
politikoen	herritar	bakea	ikusi
dute	jarrera	ez	nahi
behar	**biolentzia**	gure	bide
beste	euskal	bat	herritar

Table 4: Basque most frequent word lists

expected. If we compare data reduction from word/lemma types to tokens, the list of words drops down until 61.11% for Basque and until 66.62% for Spanish. But reduction from word types to lemma types represents a 21.5% in Basque while 7.37% in Spanish.

Moreover, the results facilitate a more informative approximation. This can be seen in Table 3 and in Table 4, where we present the resulting word-frequency lists for Spanish and Basque respectively. Each table is divided in two main sections according to the answers respondents gave about their thoughts in two different papers reflected in the corpus: Difficulties and Opportunities. Each section of the table contains two columns reporting a list of 10 most repeated words before processing (Prepr.) and after processing (Post.) the corpus with ANALHITZA.

Results show that the most frequent word list is much more informative on the post-processing column than before the processing. We see that two words belonging to the Difficulties list are repeated in Spanish and Basque texts: *i*) *partido* (in Spanish) and *alderdi* (in Basque) meaning 'political party', and *ii*) *violencia* (in Spanish) and *biolentzia* (in Basque) meaning 'violence'. In regards the underlined term *sociedad* (meaning 'society' in Spanish), it is in the Difficulties list whereas *gizarte* (meaning 'society' in Basque) is in the Opportunities list. In case we do not preprocess, the columns of Difficulties (first column of Table 3) and Opportunities (third column of Table 3) remain more or less the same (7 words of 10 are the same and none of them are content words).

In addition, NEs provided by ANALHITZA allowed us both *i*) avoiding word ambiguity in several cases: "elkarri" (reciprocal pronoun) and "elkarri_org" (organization), or "eta" (conjunction) and *ETA* (eta_org) (organization) and *ii*) further reduction by identifying several N-grams (Jurafsky, 2009) using PoS lists: *ley_partidos* 'Law on Political Parties' or *Euskal_Herri* 'Basque Country'. Indeed, PoS lists (nouns, verbs and adjectives) could help further reduction due to the fact that other words tend to be discarded as non-informative for content analysis.

Finally, another interesting feature of ANALHITZA from a NTA perspective is the lemmatized text, since the original ordering is retained and this permits network type data extraction from the corpus.

The network maps below (Figure 2), for example, represent two clusters of words to which the term 'violence' belongs in Basque and Spanish responses to the question over Difficulties for peace and normalization.[14] Departing from lemmatized texts provided by ANALHITZA, we have extracted word co-occurrence maps. The size of each word represents the degree of connectivity in the network while links between words show the strength of ties between words in terms of number of co-occurrences. In this example, we can see that the cluster of words to which the term 'violence' belongs differs considerably between both sets. While in the Basque set 'violence' is linked to words like *politika* and *politiko* 'political', *eus_ta_ask* 'eta_org' or *gatazka* 'conflict' and *epaiketa* 'trial', in the Spanish set the cluster is formed by words like *nunca* 'never', *existir* 'exist' or *asesinato* 'killing'.

[14] ConText (`http://context.ischool.illinois.edu/`) was used to extract the co-ocurrence network and Gephi (`https://gephi.org/`) to identify clusters (implementing the modularity algorithm) and network visualization.

Figure 2: A network map of the term "violence" in Basque and Spanish datasets

4.3 Indicators of the main discourse topic in a multilingual parallel corpus

ANALHITZA has been also used to analyze the Multilingual RST Treebank (Iruskieta, Da Cunha, and Taboada, 2015) which contains 15 abstracts for each language (Basque, English and Spanish). These abstracts were published in the proceedings of the International Conference about Terminology celebrated in 1997. The corpus consists of 16,830 words and is available at `http://ixa2.si.ehu.es/rst/`.

As the corpus is annotated with rhetorical structure trees (RS-trees), we extracted the central unit (CU) of each language subcorpus and built new corpora: *a*) a corpus for each language containing only CUs and *b*) a corpus for each language, containing all the text that is not a CU. The aim of this experiment is to know how a linguist can study some word combinations (or n-grams) which indicate the CU of a text in a parallel corpus. Indeed, the detection of the CU of a text can be very useful for different NLP tasks such as question answering, summarization and sentiment analysis. To do so, we analyze in Table 5 whether the combination of the pronoun 'this' ('este' in Spanish and 'hau' in Basque) with a noun is a good indicator of CU and whether it could be used in a CU detector (Iruskieta, Labaka, and Antonio, 2016), filtering the information of n-grams in all the three languages.

Moreover, we see that in the corpus built with CUs, a noun 'N' after the pronoun 'this' is significant because we find nouns that indicate the CU in the three languages (there are other nouns with the pronoun 'this' that are not indicative of CU 17.2%), such as paper

This/este	**Lemma**_Noun	***Hau***	**Freq.**
this	paper_N		6
	artikulu_N	hau	1
this	Study_N		1
este	trabajo_N		1
	lan(txo)_N	hau	2
este	ponencia_N, comunicación_N, presentación_N		6
	hitzaldi_N, komunikazio_N	hau	7

Table 5: Pronoun and noun combinations

– *artikulu*, study – *trabajo* – *lan(txo)*, *ponencia* – *hitzaldi* ('talk' in English). We can see also that Spanish and Basque use similar words *ponencia* and *hitzaldi* 'talk', *comunicación* and *komunikazio* 'communication', *presentación* 'presentation', while in English the most used term is 'paper' in this small corpus.

Additional comparisons based on n-grams could be done in this multilingual corpus to find some collocations or to describe how definitions or examples are indicated.

5 Conclusions

In this paper we have presented ANALHITZA, an application for language processing to extract linguistic information from large corpora in Basque, English and Spanish.

The tool is being developed under the Clarin-k project, whose aim is to offer useful LT tools for Humanities and Social Sciences. As starting point in the creation of NLP based LT tools, we have carried out three experiments which have shown us that ANALHITZA is indeed a very useful and interesting tool to be applied in Humanities as well as in Social Sciences. In fact, it offers

many possibilities for research, such as the comparison of texts written by same or different authors, the comparison of texts written in different periods, genres or languages, the analysis of language acquisition process, the creation of lexicons, the reduction of text complexity, the detection of CUs, etc.

Apart from the three studies presented here, and based on all those opportunities ANALHITZA offers for text analysis, our aim is to continue working on the improvement of the tools as well as on its application and dissemination in different real scenarios such as in class assignments at the university, secondary schools and Basque language academies.

Meanwhile, we continue working in the following improvements: *i*) to extract more information from the analyzed text, such as multi-words, *ii*) to use another external tool for data visualization, *iii*) to allow analyzing other file formats (PDF files), multiple files or a ZIP file in the online application.

In addition, the tool could be improved or redesigned in the foreseeable future, in case we detect that there is some need in a specific branch of study, or if researchers ask us for that.

References

Agerri, R., J. Bermudez, and G. Rigau. 2014. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In *Proceedings of LREC 2014*, pages 3823–3828.

Agerri, R. and G. Rigau. 2016. Robust multilingual Named Entity Recognition with shallow semi-supervised features. *Artificial Intelligence*, 238:63 – 82.

Alberdi, U. 2013. *Euli giro*. Susa.

Alonso, S. and A. Volkens. 2012. *Content-analyzing political texts. A quantitative approach*, volume 47. CIS.

Arrieta, B. 2012. *Alter ero*. Susa.

Blei, D.M. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Carley, K.M. 1997. Network text analysis: The network position of concepts. In Carl W. Roberts, editor, *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*, Routledge Communication Series. pages 79–100.

CICIR. 2007. *Building Peace: the Challenge of Moving from Desire to Implementation*. Columbia University.

CICIR. 2009. *The Challenge of Moving from Desire to Implementation*. Columbia University.

Fokkens, A., A. Soroa, Z. Beloki, N. Ockeloen, G. Rigau, W.R. van Hage, and P. Vossen. 2014. NAF and GAF: Linking linguistic annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*.

Grimmer, J. and B.M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*.

Hualde, J.I. and J. Ortiz de Urbina. 2003. *A grammar of Basque*, volume 26. Walter de Gruyter.

Iruskieta, M., I. Da Cunha, and M. Taboada. 2015. A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. *Language resources and evaluation*, 49(2):263–309.

Iruskieta, M., G. Labaka, and J.D. Antonio. 2016. Detecting the central units in two different genres and languages: a preliminary study of Brazilian Portuguese and Basque texts. *PLN*, 55(4):77–84.

Jurafsky, D. 2009. *Speech & language processing*. Pearson Education. India.

Krippendorff, K. 2004. *Content analysis: An introduction to its methodology*. Sage.

Otegi, A., N. Ezeiza, I. Goenaga, and G. Labaka. 2016. A Modular Chain of NLP Tools for Basque. In *Proceedings of the 19th International Conference on Text, Speech and Dialogue*, pages 93–100.

Villegas, M., N. Bel, C. Gonzalo, A. Moreno, and N. Simelio. 2012. Using Language Resources in Humanities research. In *LREC 2012*, pages 3284–3288.

Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 85-92 recibido 30-11-2016 revisado 12-01-2017 aceptado 09-02-2017

El impacto de las emociones en el análisis de la polaridad en textos con lenguaje figurado en Twitter

The Impact of Emotions on Polarity Analysis using Figurative Language in Twitter

Mª Amparo Escortell Pérez
Universitat Politècnica de València
Camino de Vera, s/n, València, Spain
mapees3@fiv.upv.es

Maite Giménez Fayos, Paolo Rosso
PRHLT Research Center
Universitat Politècnica de València
Camino de Vera, s/n, València, Spain
{mgimenez, prosso}@dsic.upv.es

Resumen: Uno de los retos más complejos a los que se enfrenta el Procesamiento de Lenguaje Natural es el de determinar la polaridad de un tweet (positiva, negativa o neutra) cuando en éste aparece lenguaje figurado, es particularmente complejo en los textos cortos y agramaticales que podemos encontrar en las redes sociales. Este trabajo presenta un estudio exhaustivo sobre la capacidad de distintos recursos léxicos de emociones para analizar la polaridad de un conjunto de datos extraidos de Twitter, detallando el impacto de cada uno de los recursos sobre distintas formas de lenguaje figurado como pueden ser la ironía y el sarcasmo que encontramos profusamente en este corpus. Los resultados obtenidos muestran indicios que apuntan a que la inclusión de información relativa a las emociones ayuda a clasificar correctamente la polaridad tanto a nivel global como a nivel del lenguaje figurado o literal.
Palabras clave: Análisis de sentimientos, emociones, lenguaje figurado, twitter, ironía, sarcasmo, semeval, polaridad.

Abstract: One of the most challenging tasks in Natural Language Processing is to determine the polarity of a tweet (positive, negative or neutral) when figurative language is present, especially in the short and ungrammatical texts that can be found in social media. In this paper we present a comprehensive study of the capacity of several emotional lexicons for Sentiment Analysis of Figurative Language in Twitter, detailing how each resource impacts on different figurative language devices such as sarcasm and irony. There are indications in our results that suggest that using emotional information improves the performance of a Sentiment Analysis model regardless of the presence or not of figurative language in the texts analyzed.
Keywords: Sentiment analysis, emotions, figurative language, twitter, irony, sarcasm, semeval, polarity.

1 Introducción

El análisis de sentimientos es una tarea propia del Procesamiento de Lenguaje Natural (PLN) que trata de determinar la polaridad que un texto pretende transmitir. Generalmente, esta tarea se ha enfocado desde el PLN como una tarea de clasificación automática de un texto en tres clases de polaridad: positiva, negativa o neutra. En el caso del lenguaje literal, las técnicas existentes logran resultados aceptables (Kiritchenko, Mohammad, y Salameh, 2016). Sin embargo, esta tarea es especialmente compleja cuando en el texto encontramos lenguaje figurado, puesto que nos enfrentamos con distintos significados debido al uso de la ironía, la metáfora o el sarcasmo, por lo tanto la polaridad del significado literal puede contrastar fuertemente con el sentimiento que pretende transmitir el sentido figurado. Incluso los seres humanos tenemos dificultad en decidir si una texto es irónico o metafórico. La ironía puede ser muy sutil, mientras que la metáfora se puede representar de muchas formas. Sin embargo, el sarcasmo es más fácil de detectar

ISSN 1135-5948

para los seres humanos. El lenguaje figurado es especialmente común en los textos que podemos encontrar en la web y en las redes sociales, especialmente en Twitter o Facebook. La limitación en la longitud en los textos de la red social Twitter, así como el uso de expresiones plagadas de argot y errores gramaticales, dificulta la comprensión del mensaje. En definitiva, el lenguaje figurado presenta un desafío para el rendimiento de los sistemas de análisis de sentimientos convencionales basados en la semántica léxica de las palabras ya que a menudo resultan insuficientes para detectar los significados indirectos. En el trabajo de Hernández et al., (2016) puede encontrarse un estudio detallado del impacto de la ironía y el sarcasmo en el análisis de sentimientos.

En este trabajo se presenta un estudio del impacto de las emociones en la detección de la polaridad de un tweet. Partimos de la hipótesis de que no todos los recursos disponibles favorecen la detección de la polaridad en igual medida, por lo tanto llevamos a cabo una serie de experimentos para evaluar cómo afectan diferentes recursos de emociones tanto en el lenguaje figurado como en el lenguaje literal. Nuestra metodología está compuesta por dos fases: en la primera de ellas estudiaremos el impacto de los recursos léxicos sobre el entrenamiento de clasificadores que predigan la polaridad del conjunto completo de tweets de la tarea 11 de SemEval2015[1] y a continuación evaluaremos en detalle el impacto de cada uno de los recursos para las diferentes tipologías del lenguaje figurado presentes en este corpus.

El resto del artículo está organizado de la siguiente manera: en primer lugar describiremos el estado de la cuestión en el apartado 2; la metodología se describe en detalle en el apartado 3; a continuación en el apartado 4 se presenta el conjunto de datos sobre el cual validaremos nuestra metodología, que como ya hemos introducido se trata de la tarea 11 de SemEval2015; en el apartado 5 se presentan los experimentos que se han llevado a cabo; a continuación en el apartado 6 evaluaremos los resultados obtenidos y, por último, extraeremos las debidas conclusiones en el apartado 7.

[1] http://alt.qcri.org/semeval2015/task11/

2 Estado de la cuestión

Como ya hemos introducido, la definición más extendida de la tarea de análisis de sentimientos se centra en clasificar los textos en tres categorías: textos positivos, negativos y neutros. Los trabajos pioneros (Pang, Lee, y Vaithyanathan, 2002) abordaron esta tarea como un problema de clasificación supervisada aunque en la literatura también podemos encontrar aproximaciones no supervisadas (Turney, 2002). En el trabajo de Pang y Lee (2008) se recoge un amplio estudio de las distintas técnicas que se han empleado para tratar de resolver la tarea del análisis de sentimientos sobre textos extraídos de Internet.

La detección del lenguaje figurado es una tarea en si misma, y distintas aproximaciones han intentado abordarla. Cuando se trata de analizar los textos, la información disponible en la web se puede utilizar como una fuente de conocimiento para generar características auxiliares. En el trabajo de Veale y Hao (2007) se describe una forma semiautomática de recopilar el conocimiento y la semántica de los estereotipos de la web atacando directamente a las construcciones del lenguaje. Los autores demostraron que alrededor del 20 % de los símiles de la web eran irónicos. Sin embargo, su trabajo no se puede utilizar para detectar la ironía de forma general ya que utilizaba las propias estructuras del lenguaje. Tradicionalmente, el lenguaje figurado se ha intentado detectar explorado las características superficiales de los textos. Por una parte, existen estudios que intentan detectar el lenguaje figurado teniendo en cuenta el orden sintáctico, las propiedades léxicas o los elementos afectivos que componen el texto (Reyes, Rosso, y Buscaldi, 2012; Reyes, Rosso, y Veale, 2013). Por otra parte, otros trabajos se centran en investigar como los hashtags de Twitter se emplean para remarcar una intención figurativa en el mensaje transmitido, en especial para la expresión de la ironía o sarcasmo (Sulis et al., 2016). El interés que despierta la tarea de la detección de la polaridad así como el impacto que tiene sobre ésta el lenguaje figurado motivó en 2015 una tarea en la competición internacional para el la evaluación semántica (*Semantic Evaluation - (SemEval)*) (Ghosh et al., 2015).

Quince equipos participaron en la tarea 11 de SemEval 2015 que fue abordada siguiendo múltiples perspectivas. La mayor parte de los participantes plantearon soluciones super-

visadas para intentar resolver la tarea, predominando dos modelos de aprendizaje automático: las Máquinas de Soporte Vectorial (MSV) y los Modelos Regresión. Dichos modelos se entrenaron utilizando un conjuntos de características cuidadosamente seleccionadas para esta tarea como pueden ser: n-gramas de caracteres, n-gramas de palabras, valores extraídos de distintos léxicos, etc.[2].

Nuestro trabajo pretende extender la aproximación presentada por Hernández et al., (2015), en la cual se abordó la tarea incorporando recursos externos adicionales. Los autores proponen representar un tweet mediante un conjunto de valores de características extraídas de recursos léxicos externos que modelan tanto las emociones como la información psicolingüística contenida en un tweet. Asimismo, el trabajo de Sulis et al., (2016) presenta un análisis de la distribución y correlación de un conjunto de características psicolingüísticas y emocionales extraídas de recursos léxicos para realizar la clasificación de tweets irónicos y sarcásticos.

Sin embargo, a diferencia de los citados trabajos sobre el estudio de las emociones en el lenguaje figurado, en este artículo presentamos un estudio exhaustivo sobre la capacidad de diferentes recursos léxicos de emociones para predecir la polaridad del conjunto de datos de Twitter de la tarea 11 de SemEval2015 detallando cómo afectan estos recursos a los tweets que contienen lenguaje figurado y lenguaje literal.

3 Metodología

En este apartado describiremos la metodología que empleamos para el estudio del impacto de ciertos recursos léxicos sobre la detección de la polaridad.

Se ha trabajado con diferentes recursos: LIWC, EmoLex y Smilies, que se detallarán en el siguiente apartado. Los recursos que hemos estudiado almacenan distintos niveles de información respecto a las palabras. El nivel más básico es la información sobre si una palabra es "positiva" o "negativa" aunque también incluyen otras categorías que indican que emociones están vinculadas a las palabras. Primeramente, se ha realizado una serie de experimentos para determinar la polaridad de los tweets utilizando únicamente las categorías de "positivo" y "negativo" de dichos recursos por separado y a continuación para todas las categorías de los recursos.

El procedimiento que hemos llevado a cabo para evaluar el impacto de cada recurso sobre la detección de la polaridad consistió en una vez tokenizados los datos de entrenamiento y test, desarrollar un estudio ablativo que evalúa cómo el uso de diferentes técnicas, como la bolsa de palabras (*Bag of words (BOW)*) o *Term frequency – Inverse document frequency (Tf-Idf)*, así como los recursos para representar un tweet, afectan a la calidad de la clasificación. Este proceso se explicará en detalle en el apartado 5.

Independientemente de la representación elegida, se ha entrenado un sistema de clasificación automático para inferir la polaridad utilizando la librería scikit-learn (Pedregosa et al., 2011). Teniendo en cuenta que el sistema debía predecir un valor de polaridad continuo, se ha empleado una Máquina de Soporte Vectorial adaptada para regresión (MSVR).

Este estudio se ha realizado tanto a nivel de todo el corpus, como a nivel de los distintos tipos de lenguaje figurado presentes en este corpus. Para dividir los tweets entre aquellos que contienen lenguaje figurado o lenguaje literal utilizamos los hashtags, asumiendo que el usuario etiqueta su propio tweet con el tipo de lenguaje empleado facilitando su comprensión, siguiendo la aproximación presentada por Sulis et al., (2016).

3.1 Recursos

Se han utilizado varios recursos para obtener los sentimientos y emociones de los tweets. Estos recursos son los siguientes:

NRC Word-Emotion Association Lexicon (EmoLex)(Mohammad y Turney, 2010): El recurso NRC Emotion Lexicon es una lista de palabras en inglés con sus correspondientes asociaciones con las ocho emociones básicas de Plutchik (Plutchik, 1980): ira, miedo, anticipación, confianza, sorpresa, tristeza, alegría y el disgusto (*anger, fear, anticipation, trust, surprise, sadness, joy y disgust*) y dos sentimientos: positivo y negativo (*negative y positive*). Este recurso fue manualmente anotado. Si la palabra pertenece a la categoría se indica con un 1, en caso contrario con un 0. En este recurso podemos encontrar 14,182 palabras etiquetadas. En la Tabla 1

[2]Para más información ver el artículo de Gosh et al., (2011).

se muestra un ejemplo de como se codifica la información en este recurso.

Palabra	Categoría	Asociación
dark	anger	0
dark	anticipation	0
dark	disgust	0
dark	fear	0
dark	joy	0
dark	negative	0
dark	positive	0
dark	sadness	1
dark	surprise	0
dark	trust	0

Tabla 1: EmoLex: representación de la palabra *dark*

Linguistic inquiry and word count (LIWC) (Pennebaker, Francis, y Booth, 2001): Este recurso le asocia a cada palabra una serie de categorías. En total hay un conjunto de 64 categorías diferentes y se muestra la asociación para un total de 4485 palabras.

Smilies (Suttles y Ide, 2013): También se utilizó este recurso de *smilies* que clasifica 176 *smilies* diferentes según la emoción asociada a los mismos. En este trabajo, en lugar de asociar un smilie a una de las seis emociones básicas definidas en la teoría de Ekman (alegría, ira, miedo, asco, sorpresa, tristeza) (Ekman, 1972), los autores utilizan los ocho tipos de emociones avanzadas definidas en la teoría de Plutchik. A partir de estos ocho tipos de emociones y utilizando una lista con los hashtags emocionales más frecuentes, los autores de este recurso seleccionaron quince categorías para etiquetar los diferentes *smilies*: feliz, risueño, amoroso, enfadado, triste, llanto, disgustado, sorpresa, beso, guiño, lengua, escéptico, indeciso, avergonzado y maligno (*happy, laugh, love, annoyed, sad, cry, disgust, surprise, kiss, wink, tongue, skeptical, indecision, embarassed y evil*). Se puede apreciar un ejemplo en la Tabla 2.

Emoticono	Emoción
:D	LAUGH
:@	SAD
;-)	WINK
3:-)	EVIL

Tabla 2: Smilies: Clasificación de *smilies*

4 Descripción de la tarea

En la tarea 11 de SemEval 2015 se utilizó un corpus de lenguaje figurado extraído de la red social Twitter el cual presenta un gran número de ironías, sarcasmos o metáforas, sin embargo, no se puede garantizar que se manifieste cualquiera de estos fenómenos en cada uno de los tweets.

La ironía y el sarcasmo normalmente se utilizan para criticar o burlarse y, por lo tanto, sesgar la percepción del sentimiento hacia un valor negativo, por lo que no es suficiente para un sistema determinar simplemente si el sentimiento de un tweet dado es positivo o negativo atendiendo únicamente al lenguaje literal presente en el mismo.

Los organizadores de la tarea proporcionaron el corpus etiquetado siguiendo una escala de 11 puntos que oscilaban desde -5 (muy negativo, para tweets con significados muy críticos) a 5 (muy positivo, tweets con significados muy optimistas). El punto cero de esta escala se utiliza para determinar los tweets neutros.

Los sistemas se evaluaron utilizando dos métricas la distancia coseno y el Error Cuadrático Medio (ECM), ambas métricas apropiadas para problemas de regresión. Por simplicidad computacional se ha decidido evaluar los sistemas aquí presentados únicamente el Error Cuadrático Medio, que define el error cometido por el vector de predicciones $\hat{Y} \in \mathbb{R}^n$ con respecto al vector con los valores correctos para esas n muestras $Y \in \mathbb{R}^n$ siguiendo la siguiente fórmula:

$$ECM = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \qquad (1)$$

Para más información acerca de los detalles de la tarea se pueden consultar las actas de la tarea (Ghosh et al., 2015).

El reto de nuestro sistema es determinar cómo un recurso o la combinación de estos, pueden influir en la capacidad del sistema de clasificación que desarrollemos para predecir el sentimiento presente en un tweet. Validaremos esta metodología propuesta sobre el conjunto de datos que describiremos en detalle en el siguiente apartado.

4.1 Corpus

El conjunto de datos empleado en la tarea 11 de SemEval fueron recolectados a través de la API Twitter4J, que soporta la recolección de tweets en tiempo real mediante la búsqueda consultas. Se utilizaron consultas de hashtags como #sarcasm, #sarcastic y #irony para obtenerlo.

Este conjunto de datos fue recogido durante 4 semanas, del 1 de junio al 30 de junio de 2014. Se eliminaron aquellos tweets que no cumplieran una serie de condiciones como por ejemplo, no contener al menos 30 caracteres sin incluir el hashtag. Asimismo, se filtró también únicamente aquellos tweets que estuvieran escritos en inglés, por lo tanto se trata de una tarea monolingüe.

Cada tweet fue etiquetado por siete anotadores, tres de los cuales eran hablantes nativos de inglés y el resto de los anotadores eran competentes en el idioma. A todos ellos se les pidió asignar una puntuación que oscilaba desde -5 a 5, donde 0 es el valor neutro para aquellos tweets que tienen el mismo valor negativo que positivo.

El sentimiento general de cada tweet se calculó como una media ponderada de las siete puntuaciones donde las puntuaciones de los nativos del inglés valían el doble.

El conjunto de tweets de entrenamiento y test está compuesto por 8000 y 4000 tweets respectivamente. Un ejemplo de varios tweets se muestra en la siguiente tabla:

Tweet	Polaridad
"@erikaekengren: From 50 to 100 degrees in less than a week #kansas" #cantwait **#sarcasm**	-3
Updated my router and it froze. Now I can't access the internet to google a solution. **#irony** #thankfulforsmartphones	-3.48
I've had a lot of wake up calls in my day, but I've always been good at hitting the snooze **#metaphor** #nailedit	0.22

Tabla 3: Ejemplos extraídos del corpus de la tarea 11 de SemEval 2015. En la columna "Polaridad" se especifica la polaridad con la que se puntuó de media el tweet mostrado como ejemplo

5 Experimentación

Nuestra experimentación se ha llevado a cabo en varias fases. En la primera de ellas se ha estudiado el impacto de los recursos sobre el conjunto completo de tweets y a continuación se ha evaluado el grado de impacto para cada uno de los diferentes conjuntos de tweets con lenguaje figurado en el corpus.

Se ha tokenizado el corpus utilizando la librería NLTK (Bird, Klein, y Loper, 2009), se han eliminado las palabras que no aportan información discursiva (*stopwords*) y se ha convertido todo el texto que no fueran *smilies* a minúsculas. A continuación, se han obtenido las representaciones BOW y Tf-Idf de los tweets utilizando la librería scikit-learn (Pedregosa et al., 2011).

Para los recursos de EmoLex y LIWC se han creado diccionarios que representan eficientemente la información. Cada entrada del diccionario corresponde con una categoría emocional y en ella se almacenan las palabras que forman parte de dicha categoría. Para utilizar el recurso de *smilies* se ha tenido que crear un tokenizador ad hoc con cada una de las expresiones regulares necesarias para identificar todos los *smilies*.

Una vez se han obtenido los diccionarios de cada recurso, se han elaborado representaciones vectoriales de las muestras de entrenamiento y test. Cada uno de estos vectores indican para cada tweet, el número de veces que aparece una palabra de las categorías que se tienen en el diccionario. De esta manera se tiene un vector diferente para cada recurso que posteriormente se combinaran para realizar la experimentación. La combinación de estos vectores consiste simplemente en agregar al final del vector del primer recurso el vector del segundo.

Utilizando estas estructuras, se han realizado varios experimentos sobre todo el conjunto del corpus. Primero utilizando únicamente las categorías "positivo" y "negativo", y, a continuación, utilizando todas las categorías disponibles de cada uno de los recursos.

En la última columna de la Tabla 4 se muestran los resultados alcanzados tras combinar diferentes recursos para entrenar una SVR y calculando el resultado utilizando el ECM. Además, para poder comparar los re-

Recurso	**#irony**	**#sarcasm**	**#not**	**otros**	**total**
BOW + EmoLex p/n	0,8466	0,5790	5,8184	6,8359	4,6025
TF-IDF + EmoLex p/n	0,8781	0,5794	5,9199	6,9498	4,6825
BOW + EmoLex Todas	0,8459	0,5787	5,8136	6,8269	4,5972
TF-IDF + EmoLex Todas	0,8770	0,5800	5,9101	6,9374	4,6744
BOW + LIWC p/n	0,8471	0,5817	5,8257	6,8421	4,6074
TF-IDF + LIWC p/n	0,8754	0,5821	5,9148	6,9451	4,6789
BOW + LIWC Todas	**0,8331**	0,5780	5,8076	6,8187	4,5897
TF-IDF + LIWC Todas	0,8583	0,5782	5,8972	6,9265	4,6628
BOW + EmoLex + LIWC p/n	0,8451	0,5788	5,8203	6,8421	4,6022
TF-IDF + EmoLex + LIWC p/n	0,8756	0,5796	5,9173	6,9460	4,6796
BOW + EmoLex + LIWC Todas	0,8338	0,5759	**5,7883**	**6,7972**	**4,5756**
TF-IDF + EmoLex + LIWC Todas	0,8586	**0,5755**	5,8778	6,9054	4,6487
BOW + *Smilies*	0,8484	0,5817	5,8227	6,8407	4,6360
TF-IDF + *Smilies*	0,8748	0,5813	5,9041	6,9355	4,6719
Baseline: Naïve Bayes	-	-	-	-	5,6720

Tabla 4: Resultados obtenidos empleando la métrica ECM evaluando cada uno de subconjuntos de tipos de lenguaje figurado que podemos encontrar en este corpus. Señalamos con la abreviatura "p/n" aquellas representaciones en las que únicamente se emplean las características "positiva" y "negativa" del recurso en cuestión, mientras que aquellos recursos en los que empleamos todas las categorías se señalan como "Todas"

sultados se incluye el ECM de un sistema de control (*Baseline*) Naïve Bayes entrenado con una aproximación de bolsa de palabras facilitado por la organización de la tarea.

Dada esta experimentación preliminar, la segunda fase ha consistido en determinar el grado de impacto de cada uno de los recursos sobre las diferentes expresiones de lenguaje figurado más habituales en este corpus. Para ello se han diferenciado un total de cuatro grupos de tweets, en función de la aparición de los siguientes hashtags: *#irony* (765 tweets), *#sarcasm* (536 tweets), *#not* (981 tweets) y *otros* (1718 tweets). Si un tweet tiene varios hashtags pertenecerá a ambos conjuntos.

En este último conjunto,*otros* se agrupan aquellos los tweets que no forman parte de los tres primeros grupos y que por lo tanto asumimos que se trata de lenguaje literal. Esta separación se ha hecho sobre los datos del test del corpus, aceptando que el usuario ha empleado el hashtag para auto-etiquetar el tipo de lenguaje que su tweet contenía. La organización de SemEval reportó resultados sobre un conjunto de test con metáforas, pero no aparece el hashtag #metaphor en el test y no hemos podido llevar a cabo una separación automática de este conjunto de tweets. Por lo que en el conjunto *otros* aparecerán metáforas, que según las actas de las tarea (Ghosh et al., 2015) es una de las formas de lenguaje figurado más difíciles de clasificar.

Una vez separados los tweets, se ha llevado a cabo la segunda fase de la experimentación en la que se han utilizado como datos de entrenamiento todos los tweets del conjunto de entrenamiento de la tarea, pero como datos de test se ha utilizado cada uno de los grupos acabamos de describir. Al igual que en la primera fase de experimentos, se ha probado cada recurso por separado así como la combinación de ellos. En la Tabla 4 se muestran para cada uno de los grupos los resultados obtenidos con los diferentes recursos.

6 Análisis de los Resultados

Como ha podido comprobarse en los resultados de la experimentación que hemos presentado en el apartado anterior, la inclusión de nuevos recursos nos conduce a mejorar significativamente el comportamiento de los mo-

delos que entrenemos. Sin embargo, el ECM varía considerablemente en función del subconjunto de lenguaje que estemos considerando. Todos los modelos que hemos presentados consiguen mejorar el modelo de control o *baseline*, lo cual nos indica que, efectivamente, los recursos léxicos que aportan información acerca de las emociones ayudan a mejorar la predicción del sentimiento comunicado en un tweet. Cabe destacar que, cuando incluimos la información respecto a todas las emociones disponibles en un recurso léxico, y no únicamente las categorías "positivas" y "negativas", conseguimos mejorar el comportamiento del modelo entrenado independientemente de si estamos ante lenguaje figurado o literal. Además, a pesar de que estamos ante un corpus con una baja frecuencia de *smilies* y por lo tanto la cobertura del léxico smilies es escasa, este recurso también consigue mejorar el sistema. La aproximación que incluye la bolsa de palabras y los recursos afectivos EmoLex y LIWC con todas las emociones obtiene resultados satisfactorios aunque en el caso de los tweets en los que aparece el *hashtag #irony* la aproximación no utiliza el recurso EmoLex es la que presenta un mejor comportamiento mientras que en el caso de los tweets con el *hashtag #sarcasm* el mejor sistema utiliza una representación de palabras basada en Tf-Idf. Sin embargo, la diferencia entre estos sistemas no es significativa y podemos concluir que el mejor sistema es el que emplea las características "BOW + EmoLex + LIWC Todas". El mejor sistema participante en la tarea, ClaC (Ozdemir y Bergler, 2015), obtuvo un EMC 2.117 para lo cual se desarrolló un complejo proceso para la extracción de la polaridad de las palabras en función del contexto en el que aparezcan. Nuestro sistema comparte características con el sistema denominado ValenTo (Farıas et al., 2015), aunque en este trabajo se incluyen más recursos que pretendemos explorar en trabajos futuros.

El estudio de cómo se distribuye el error entre los distintos subgrupos de tweets ha arrojado resultados sorprendentes: la mayor parte del error se concentra en el subgrupo que no contenía ningún tipo de hashtags del conjunto de hashtags estudiando (#irony, #sarcarm, #not), lo cual puede explicarse porque no se puede analizar de modo independiente el impacto de la metáfora sobre el conjunto *otros*. No obstante, se requiere un estudio pormenorizado de los tweets y la polaridad para explicar este fenómeno.

7 Conclusiones y Trabajo Futuro

En este trabajo, se ha presentado un estudio sobre la capacidad de distintos recursos léxicos de emociones para predecir la polaridad de un conjunto de datos extraídos de Twitter. Se ha visto el impacto de cada uno de ellos sobre las distintas formas de lenguaje figurado como la ironía y el sarcasmo y la importancia de desarrollar técnicas capaces de representar esa información para clasificar el sentimiento que el autor emitió en un texto. Se han obtenido unos resultados que apuntan a que la inclusión de información relativa a las emociones ayuda a clasificar correctamente la polaridad tanto a nivel global como a nivel del lenguaje figurado o literal .

Como trabajo futuro, se pretende extender el estudio a distintos algoritmos de aprendizaje automático para comprobar cómo afecta a su comportamiento la inclusión de información emocional recurrente o ruidosa, puesto que las MSV son capaces de descartar aquellas muestras no significativas para la clasificación y son más robustas respecto al ruido. Asimismo, se evaluará de forma sistemática la contribución de la representación de palabras y de nuevos recursos léxicos en la detección de la polaridad de un tweet. Igualmente, se estudiará cómo se puede aumentar la cobertura de los recursos léxicos, es decir el número de palabras que encontramos en el diccionario, utilizando técnicas como por ejemplo la corrección automática del texto, para eliminar, en la medida de lo posible los errores gramaticales presentes en Twitter.

Agradecimientos

Este trabajo se ha desarrollado en el marco del proyecto de investigación SomEMBED (TIN2015-71147-C2-1-P) del Ministerio de Economía y Sostenibilidad (MINECO). Asimismo, el trabajo de la segunda autora ha sido financiado a través del Programa de Ayudas de Investigación y Desarrollo de la Universitat Politècnica de València (PAID 2015).

Bibliografía

Bird, S., E. Klein, y E. Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc.

Ekman, P. 1972. Universals and cultural differences in facial expressions of emo-

tions. *Nebraska Symposium on Motivation*, 19:207–283.

Farıas, D. I. H., E. Sulis, V. Patti, G. Ruffo, y C. Bosco. 2015. Valento: Sentiment analysis of figurative language tweets with irony and sarcasm. *SemEval-2015*, página 694.

Ghosh, A., L. G., T. Veale, P. Rosso, E. Shutova, J. Barnden, y A. Reyes. 2015. SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter. *Proc. 9th Int. Workshop on Semantic Evaluation (SemEval 2015)*, páginas 470–478.

Hernández, I. y P. Rosso. 2016. Irony, sarcasm, and sentiment analysis. En F. Pozzi E. Fersini E. Messina, y B. Liu, editores, *Sentiment Analysis in Social Networks*. Morgan Kaufmann, capítulo 7, páginas 113–128.

Kiritchenko, S., S. Mohammad, y M. Salameh. 2016. Semeval-2016 task 7: Determining sentiment intensity of english and arabic phrases. En *Proceedings of the International Workshop on Semantic Evaluation (SemEval), San Diego, California, June*.

Mohammad, S. M. y P. D. Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. En *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, páginas 26–34. Association for Computational Linguistics.

Ozdemir, C. y S. Bergler. 2015. Clacsentipipe: Semeval2015 subtasks 10 b, e, and task 11. En *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, páginas 479–485.

Pang, B. y L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Pang, B., L. Lee, y S. 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*, páginas 79–86. Association for Computational Linguistics.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, y others. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Pennebaker, J. W., M. E. Francis, y R. J. Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001.

Plutchik, R. 1980. Emotion: Theory, research, and experience. *Theories of Emotion*, 1.

Reyes, A., P. Rosso, y D. Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12.

Reyes, A., P. Rosso, y T. Veale. 2013. A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*, 47(1):239–268.

Rosenthal, S., P. Nakov, S. Kiritchenko, S. Mohammad, A. Ritter, y V. Stoyanov. 2016. SemEval-2015 Task 10: Sentiment Analysis in Twitter. *Proc. 9th Int. Workshop on Semantic Evaluation (SemEval 2015)*.

Sulis, E., I. Hernández, P. Rosso, V. Patti, y G. Ruffo. 2016. Figurative Messages and Affect in Twitter: Differences Between #irony, #sarcasm and #not. *Knowledge-Based Systems*, 108:132–143.

Suttles, J. y N. Ide. 2013. Distant Supervision for Emotion Classification with Discrete Binary Values. *Computational Linguistics and Intelligent Text Processing*, páginas 121–136.

Turney, P. D. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. En *Proceedings of the 40th annual meeting on association for computational linguistics*, páginas 417–424. Association for Computational Linguistics.

Veale, T. y Y. Hao. 2007. Learning to understand figurative language: from similes to metaphors to irony. *Proceedings of CogSci 2007*.

Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 93-100 recibido 30-10-2016 revisado 13-01-2017 aceptado 07-02-2017

On the Feasibility of External Factual Support as Wikipedia's Quality Metric

Sobre la Factibilidad del Soporte Factual Externo como Métrica de Calidad para Wikipedia

Carlos G. Velázquez[1], **Leticia C. Cagnina**[1,2], **Marcelo L. Errecalde**[1]
[1]LIDIC - Universidad Nacional de San Luis. San Luis, Argentina.
[2]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET).
e-mails: carvear20@yahoo.com.ar, {lcagnina,merreca}@unsl.edu.ar

Abstract: Developing metrics to estimate the information quality of Wikipedia articles is an interesting and important research area. In this article, we propose and analyse the feasibility, of a new quality metric based on the "external factual support" of an article. The rationale behind this metric is identified, a formal definition of the metric is presented and some implementation aspects are introduced. Preliminary results show the feasibility of our proposal and its potential to discriminate high quality versus low quality Wikipedia's articles.
Keywords: Quality metrics, Wikipedia, featured articles, external support

Resumen: El desarrollo de métricas para estimar la calidad de información de los artículos de Wikipedia es un área de investigación intersante e importante. En este artículo, se propone una nueva métrica de calidad basada en el "soporte factual externo" de un artículo y se analiza su viabilidad. Los motivos que dan sustento a esta métrica son identificados, se presenta una definición formal de la misma y también se dan detalles de su implementación. Los resultados preliminares obtenidos, muestran la viabilidad de nuestra propuesta y su potencial para discriminar entre artículos de alta y baja calidad en Wikipedia.
Palabras clave: Métricas de calidad, Wikipedia, artículos destacados, soporte externo

1 Introduction

Automatic assessment of Information Quality (IQ) is a topic of growing interest, mainly due to the increasing popularity of user-generated Web content and the unavoidable divergence of the delivered content's quality (Baeza-Yates, 2009). In this context, Wikipedia, the largest and most popular user generated knowledge source on the Web, presents different challenges related to quality assurance. Its size and its dynamic nature render a manual quality assurance completely infeasible. This has resulted in an increasing number of articles related to automatic IQ assessment in Wikipedia that can be categorized into three research lines: a) Featured articles identification (Blumenstock, 2008; Lipka and Stein, 2010); b) Development of quality metrics (Lih, 2004; Stvilia et al., 2005); and c) Quality flaws detection (Anderka, Stein, and Lipka, 2012; Ferretti et al., 2014).

In this paper we focus on the second task, development of quality metrics for Wikipedia, an area where several methods have been recently proposed (Lex et al., 2012; Ingawale et al., 2013). A distinctive characteristic of most of those works is that they exclusively rely on "local" information directly obtained from the Wikitext content of the article or its edition history. However, in many cases, this information alone would seem to be insufficient to capture some IQ aspects which are intuitively related to "external information". Our hypothesis is that the *external support* of the information contained in Wikipedia articles can be useful to identify quality aspects of those articles. In order to start working on this hypothesis, we propose a quality metric named "external factual support". To this end, we first introduce in Section 2 some general concepts on quality metrics for Wikipedia. Then, in Section 3, motivations for the proposed metric and its

ISSN 1135-5948

formal definition are presented. Section 4 gives implementation details of the metric, a description of the data sets and experimental results validating our proposal. Finally, in Section 5 some conclusions are drawn and possible future work is discussed.

2 Quality metrics for Wikipedia

In a nutshell, a quality metric is a quantitative *estimation* of *to what extent* a textual resource (a Wikipedia article in this case) satisfies a specific property, such as *informativeness*, *reputation*, *generality*, *completeness*, etc. As we can see, quality metrics are *subjective*, in the sense that different people could define them in different ways. That contrasts with other "objective" properties such as article's *length* or *number of pictures* in the article, which are usually termed *quality measures*. Quality measures are directly *measured* with a suitable computer program while quality metrics are *estimated* by using some arbitrary formula. As an example, assume d is an arbitrary Wikipedia article, $len(d)$ the measure representing the length of d and $nuin(d)$ another measure that gives the number of images in d. One could represent the (abstract) property "*informativeness*" by means of a metric $info$ defined as: $info(d) = len(d) + 4 \times nuin(d)$. Obviously, another person might use another criteria to define the same quality metric. In Stvilia et al. (2005), for instance, 7 arbitrary quality metrics which are based on 19 quality measures were presented. The proposed IQ metrics showed to be successful in discriminating high quality Wikipedia articles.

Quality metrics can be used for ranking (and visualizing) documents according to the property represented by the metric. For instance, Wikipedia articles could be shown in decreasing order according to their estimated informativeness. On the other hand, they can also be integrated as part of other more general processing systems, such as text categorization or text clustering systems. In those cases, quality metrics can be used alone as features for representing the documents or combined with other arbitrary features.

As far as we know, the first works that specifically addressed the definition of quality metrics in Wikipedia date back to 2004 (Lih, 2004; Viégas, Wattenberg, and Dave, 2004), where concepts likes "*reputation*" of an article are defined by using the article edition history. In contrast, in Emigh and Herring (2005) different features are proposed to identify "formal language", which are directly derived from the article *content* (*POS* tags, for instance). An aspect recurrently used in definitions of quality metrics for Wikipedia is the social/collaborative structure generated between article *editors* and the *articles* been edited. Results obtained in Wilkinson and Huberman (2007) agree with those presented in Anthony, Smith, and Williamson (2009; Lih (2004) about the influence of qualified and occasional collaborators in the quality of the articles. Hu et al. (2007) also analyse collaborative models for measuring quality aspects based on relations between "good collaborators" and "good articles". Finally, in Ingawale et al. (2013) the interaction among editors and articles is visualized as a *network* (or *graph*) and graph theory is used to infer *structural properties* associated to quality of articles.

In Lex et al. (2012) is recognized that to assess factual accuracy of Web content, more complex, semantic features are needed. A common approach is employing Open Information Extraction (Etzioni et al., 2008) or methods that use background knowledge on semantic relations available in ontological resources. These methods extract relational information about entities, i.e. facts like $f = (Mozart, was\_born\_in, Salzburg)$. Besides, they exploit semantic relationships such as meronymy and hypernymy to infer relational information between entities not explicitly given in the text. In order to measure information quality based on factual information, different approaches are identified. Afterwards, they propose very simple metrics, named *fact frequency-based features*, which attempt to determine the informativeness level of a document. These features are the closest antecedent and the basis for the proposal presented in the paper in hand. Therefore, they will be described in this section with more details in order to make easier the understanding of the "external factual support" concept presented in Section 3.

Fact frequency-based features only require information about the number of facts obtained by an information extraction process from a textual resource. For instance, if t is an arbitrary textual resource (e.g. a paragraph, a document, a corpus), and F_t is the collection of facts extracted from t by an arbi-

trary information extraction method IE, it is direct computing the *fact count* of t, denoted $fc(t)$. It is simply defined as the total number of facts obtained from t by IE, $fc(t) = |F_t|$. Obviously the fact count directly depends on the size of the textual resource t, so it is usually normalized according to the size of t. This quantity is referred in Lex et al. (2012) as the *factual density of* t, and denoted $fd(t)$. In that case, if $size(t)$ is a measure intended to quantify the size of t,[1] the factual density of t, is defined as $fd(t) = \frac{fc(t)}{size(t)}$. As it will be seen in Section 3, facts from the F_t collection will be used to compute the external factual support of t, where t corresponds to a Wikipedia article.

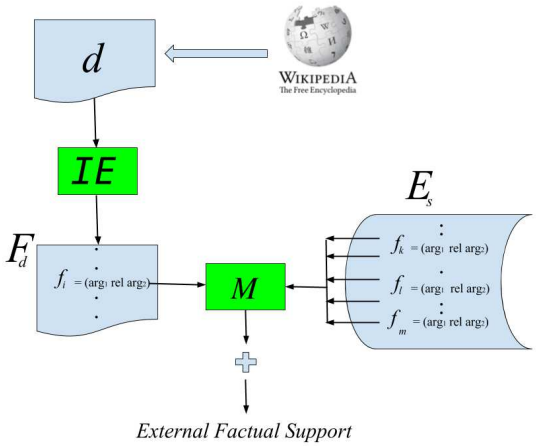

Figure 1: Computation of the EFS metric

3 External Factual Support

Most of the above-mentioned approaches assume that all the relevant information to determine the Wikipedia articles' quality is present in the content of an article or in its edition history. However, that is not always the case. For instance, let consider the *original research* (OR) aspect, one of three core content policies that, along with "Neutral point of view" and "Verifiability", determines the type and quality of material acceptable in Wikipedia articles.[2] OR refers to a problem (flaw) exhibited by material such as facts, allegations, and ideas for which no reliable, published sources exist. To demonstrate that you are not adding OR, you must be able to cite reliable, published sources that are directly related to the topic of the article, and directly support the material being presented. However, checking for the absence of inline citations of sources does not guarantee that OR will be detected because all the statements might involve well known information. For example: the statement "Paris is the capital of France" needs no source, because no one is likely to object to it and we know that sources exist for it. The statement is *attributable*, even if *not attributed*. As it can be seen, a Wikipedia article that violates the "No Original Research" principle will directly affect its chances of being a "featured article". However, the necessary information to determine this aspect cannot be realistically obtained if only the article's content is considered and some kind of extra "external information" is required.

Our main aim in this paper is defining a measure that estimates the *external support* of a document d, i.e., how much information in an external source E_s contributes to show that the content in d is either true, important, well known or all of them together. To do this, we take as basis (the same as in Lex et al. (2012)) the set of facts F_d, that is, the collection of facts extracted from d by an arbitrary information extraction method IE (for instance, the ReVerb Open Information Extraction framework[3]). Our idea in the present work is taking a closer look to the information available about each fact $f_i \in F_d$ and estimating the external support $s_e(f_i)$ that this fact has in the external source E_s. The computation of the external support of f_i will be based on a *matching* mechanism $\mathcal{M}$ in charge of deciding if a fact in E_s "matches"[4] f_i (or not). Then, the external support $\mathcal{S}_e(d)$ of the whole document d will be a weighted sum of the support $s_e(f_i)$ of each fact $f_i \in F_d$. That intuitive idea of the *external factual support* (EFS) of a document is illustrated in Figure 1 and more formally defined below.

Definition 1. (External Factual Support) Let d be a document and $F_d = \{f_1, \ldots, f_n\}$ the collection of facts extracted from d by an arbitrary information extraction method IE. The **external factual support of** d, denoted $\mathcal{S}_e(d)$, is defined as

[1] For instance, it could be the number of words or sentences in t or the number of characters of t.

[2] http://en.wikipedia.org/wiki/Wikipedia:No_original_research

[3] http://reverb.cs.washington.edu/

[4] Initially, that means that both facts are the same. Then, we will see that other (more relaxed) types of pairings between facts will be allowed.

$$\mathcal{S}_e(d) = \sum_{i=1}^{n} w_i s_e(f_i) \quad (1)$$

where w_i is the *weight* that fact f_i is given in document d and $s_e(f_i)$ is the *external factual support* of f_i.

The idea of using weights w_i's to give different "importance" to the facts f_i's (and their associated external factual supports $s_e(f_i)$) is intuitively simple. It is motivated by the idea that in specific situations some information is available about which facts could be more relevant than others in a document d. In Magdy and Wanas (2010), for instance, facts obtained from sentences appearing earlier in the document are given a higher weight.

We use a different approach that consists in using information directly provided by the information extraction method IE. For instance, fact-extraction systems like Reverb associate with each extracted fact f_i a *trust* or *confidence* value c_i. Typically, c_i indicates how confident is the extractor about the accuracy of the extracted fact f_i. In that way, a direct method to determine the weight w_i is simply taking the confidence value of f_i, $w_i = c_i$. However, other alternatives to set w_i are also valid like, for instance, considering some type of "threshold" t, such that $w_i = c_i$ only in those cases where c_i is greater than t. Thus, for example, if a threshold $t = 0.8$ were considered, the w_i formula in that case would be:

$$w_i = \begin{cases} c_i & \text{if } c_i \geq 0.8 \\ 0 & \text{si } c_i < 0.8 \end{cases} \quad (2)$$

It is also clear here, that a trivial setting for w_i is giving the same uniform value to all the extracted facts (for instance $w_i = 1$).

From Equation 1 we can see that another key component to compute $\mathcal{S}_e(d)$ is the EFS of f_i, $s_e(f_i)$. Intuitively, this quantity should give some information about how many times the fact f_i was found in the external source E_s. Thus, if f_i appears N_i times in E_s, a direct option is using $s_e(f_i) = N_i$ as external factual support of f_i. However, we also could be interested in the *boolean case*, that is, only evaluating if f_i was found in E_s or not. In that case, $s_e(f_i)$ might be defined as:

$$s_e(f_i) = \begin{cases} 1 & \text{if } f_i \in E_s \\ 0 & \text{in other case.} \end{cases} \quad (3)$$

Another aspect that must be taken into account in the support computation is the *size* of a document d. Intuitively, we can speculate that a greater size of d will result in a higher value of $\mathcal{S}_e(d)$. Thus, some kind of "normalization" in our metric definition could be desirable. Therefore, instead of directly considering the $\mathcal{S}_e(d)$ formula shown in Equation 1, we use a more general equation that allows to specify that no normalization is required, or different normalization units when the results need to be normalized. Thus, our EFS formula for a document now is defined as:

$$\widehat{\mathcal{S}}_e(d) = \frac{\mathcal{S}_e(d)}{nor} \quad (4)$$

with the *normalization factor* nor taking one of the following values: a) $nor = 1$ (no normalization), b) $nor = NL_d$ (number of lines in d, c) $nor = NW_d$ (number of words in d), $nor = |F_d|$ (number of facts extracted from d).

In summary, if the different options for w_i are identified as: C when $w_i = c_i$, T when Equation 2 is used and U when $w_i = 1$; we identify the alternatives for $s_e(f_i)$ as: N when $s_e(f_i) = N_i$ and B for the "boolean case" (Equation 3), and the normalization alternatives are denoted as: N (no normalization), L (lines-based normalization), W (words-based normalization) and F (facts-based normalization), we can see that different methods for computing the external factual support are obtained by simply considering different combinations of the weight w_i, the external support of the facts ($s_e(f_i)$) and the used normalization (if any). Following the above specified naming convention, each of those components will be assigned a character in a "code" that will identify the used support. Thus, for instance, an EFS identified as "CNW" corresponds to the case in which w_i is the confidence level assigned by the fact-extraction system (Reverb in our case) to f_i, the external support of f_i is the number of occurrences of f_i in E_s and the results are normalized taking into account the number of words in each document d. Table 1 summarizes different support codifications that result from using different alternatives for w_i, $s_e(f_i)$ and nor.

There is an aspect that has not been analysed yet but, as it will be seen in the next section, deserves a lot of attention: the process

Codification	w_i	$s_e(f_i)$	nor
CNN	c_i	N_i	1
CNL	c_i	N_i	NL_d
CNW	c_i	N_i	NW_d
CNF	c_i	N_i	$\|F_d\|$
CBN	c_i	Equation 3	1
CBL	c_i	Equation 3	NL_d
CBW	c_i	Equation 3	NW_d
CBF	c_i	Equation 3	$\|F_d\|$
TNN	Equation 2	N_i	1
TNL	Equation 2	N_i	NL_d
TNW	Equation 2	N_i	NW_d
TNF	Equation 2	N_i	$\|F_d\|$
TBN	Equation 2	Equation 3	1
TBL	Equation 2	Equation 3	NL_d
TBW	Equation 2	Equation 3	NW_d
TBF	Equation 2	Equation 3	$\|F_d\|$
UNN	1	N_i	1
UNL	1	N_i	NL_d
UNW	1	N_i	NW_d
UNF	1	N_i	$\|F_d\|$
UBN	1	Equation 3	1
UBL	1	Equation 3	NL_d
UBW	1	Equation 3	NW_d
UBF	1	Equation 3	$\|F_d\|$

Table 1: EFS codifications

used to "match" a fact f_i with the facts in the external source E_s when the $s_e(f_i)$ value needs to be computed. Up to now, we have assumed that a fact f_i is "found" in E_s when there is a "perfect" matching with the external fact, that is to say, they are the same fact. However, we will see later that this "strict matching" approach produces low recall values and the matching process needs to be relaxed.

4 Implementation aspects and experimental results

To test the feasibility of the proposed quality metric it is necessary to generate adequate data sets with high and low quality Wikipedia's articles. Intuitively, the EFS metric should help discriminating in these data sets between both types of articles. Wikipedia has a definite concept of information quality standard represented by the concepts of "Featured articles" and "Good articles". Its editors annotate articles with respect to these information quality criteria which makes them perfectly suited as positive examples of the highest quality articles that one would expect to find in Wikipedia. Featured/Good articles were identified by searching for files in a Wikipedia dump that contains the featured article or good article template in the Wikitext. As low quality examples, we used non-featured articles that were randomly selected from the remaining articles in the dump or taken from a set of articles that have a specific "flaw", as it will be explained below.

Our dataset consists of 2445 Wikipedia articles, 1000 featured/good and 1445 non-featured articles. They will be referred from now on as the "featured article" (FA) set and the "non-featured article"(NF) set respectively. In fact, we can differentiate in the NF set two subsets: one, the subset formed by 939 "regular" non-featured articles randomly selected from the snapshot of the English Wikipedia from October 2011; the other one consists of 506 articles that have been tagged as having the "original research" flaw in the corpus used in the PAN'12 competition on "Quality Flaw Prediction in Wikipedia" (Anderka and Stein, 2012). Original research is one of the aspects explicitly prohibited in the politics for featured articles of Wikipedia so, in some way, articles with this flaw would represent "reliable" non-featured (low quality) examples.

The whole dataset was processed in order to obtain 24 EFS measures that correspond to the 24 codifications described in Table 1. We used as external source E_s the ReVerb ClueWeb Extractions data set (Fader, Soderland, and Etzioni, 2011). This data set contains approximately 15 million binary assertions from the Web. It is a subset of ReVerb's output run on the English portion of the ClueWeb09 corpus.[5]

As it was pointed out above, the "strict matching" approach used to determine the EFS of each fact produced very low recall values. In fact, for many arbitrary Wikipedia articles d_j, all the extracted facts $F_{d_j} = \{f_{j_1}, \ldots, f_{j_n}\}$ will have a EFS $s_e(f_{j_i}) = 0$, for $i = 1 \ldots n$ and, in consequence, the external factual support of d_j, $\widehat{\mathcal{S}}_e(d_j)$ will be 0 for all the codifications shown in Table 1. Thus, for instance, if only the articles d that have $\widehat{\mathcal{S}}_e(d) \neq 0$ are considered, a reduction in the number of articles is observed in the (sub)-sets of our dataset: from $|FA| = 1000$ to 346 and from $|NF_R| = 1445$ to 153. We denote $FA^\star$ and $NF^\star$ those "reduced" (non-zero external factual support) sets of articles (see Table 2).

It is interesting to notice that, despite the low recall problem that introduces the

[5]More information on the ReVerb homepage at: http://reverb.cs.washington.edu/

"strict matching" approach, we can already see some "discriminative" capabilities of the external factual support. The percentage of FA documents with external support $\neq 0$: $346/1000 = 34.6$ %, is considerably higher than the percentage of non-featured articles with external support $\neq 0$ in the NF set: $153/1445 = 10.59\%$.

This is an encouraging reason for keep working on the external support measures and also poses a challenging scenario to be addressed in the experimental work. That is to say, $FA^\star$ and $NF^\star$ constitute by themselves a difficult dataset to test our EFS measures. It represents a sub-collection of the original dataset where the "negative" class ($NF^\star$) includes those examples that are the nearest to the positive examples because they have at least some minimum EFS (with respect to "strict matching" approach).

Obviously, to obtain a metric that gives more information on all the considered documents, it is necessary to define alternative (more relaxed) matching criteria than the exact matching of facts. We have a lot of possibilities to do this and, in fact, they will be considered in future works. However, in the present work we decided to start with two very simple matching approaches that we called the *local* and *global* matching approaches.

The local matching approach simply measures the component-by-component degree overlapping of each part of a fact and the (external) fact we are comparing to. More formally, let $f_i = (s_{i_1}, s_{i_2}, s_{i_3})$ be the fact we are computing the external factual support, and let $f_e = (s_{e_1}, s_{e_2}, s_{e_3})$ be a fact in the external source, $f_e \in E_s$. The *local matching* of f_i with respect to f_e will be:

$$\mathcal{M}_l(f_i, f_e) = J(s_{i_1}, s_{e_1}) \times J(s_{i_2}, s_{e_2}) \times J(s_{i_3}, s_{e_3})$$

where $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$ is the *Jaccard similarity coefficient* of sets A and B. That is to say, $\mathcal{M}_l(f_i, f_e)$ computes the product of the component-by-component overlapping degree of both facts, considering that each part of a fact is a set of terms.

The global matching approach only differs from the local one in that it considers all the parts in a fact as a single set, that is

$$\mathcal{M}_g(f_i, f_e) = J(s_{i_1} \cup s_{i_2} \cup s_{i_3}, s_{e_1} \cup s_{e_2} \cup s_{e_3})$$

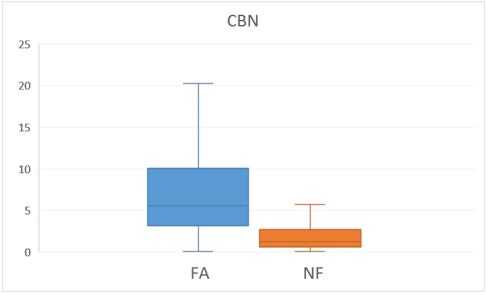

Figure 2: FA's vs NF's CBN values

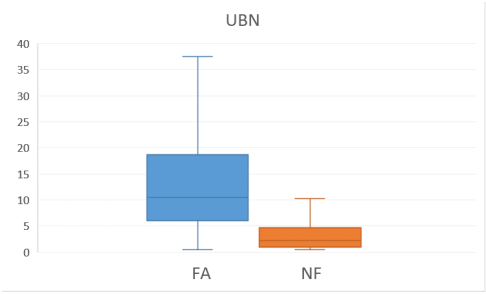

Figure 3: FA's vs NF's UBN values

In that way, $\mathcal{M}_g$ provides a "more relaxed" notion of matching than $\mathcal{M}_l$. For instance, facts $(Bach, moved\_to, Weimar)$ and $(Bach, lived\_in, Weimar)$ would produce a zero $\mathcal{M}_l$ value (second components of facts do not overlap) but that value would be greater than zero under the $\mathcal{M}_g$ matching criterion.

The only aspect to decide in each case is which would be an appropriate threshold value t, such that when $\mathcal{M}(f_i, f_e) \geq t$ it will be considered that f_i and f_e "match". In the experimental work we empirically determine two different thresholds $t_l = 0.3$ and $t_g = 0.4$ for $\mathcal{M}_l$ and $\mathcal{M}_g$ respectively, that produced fairly reasonable matches between facts.

In Table 3 a summary of the number of documents of each sub-group of the data set is shown and also of the reduced version ($DS^\star$) that results of considering non-zero EFS measures when different matching approaches are used. There, we can see that an almost perfect recall is obtained for featured articles for the "relaxed" matches, with $\frac{|FA^\star|}{|FA|} = 960/1000 = 96\%$ of external support $\neq 0$ for the local matching, and $\frac{|FA^\star|}{|FA|} = 999/1000 = 99.9\%$ of external support $\neq 0$ for the global matching approach. Following a similar analysis to the one carried out with strict matching, we can see that these recall values are higher for FA articles than for the ones obtained with NF articles in both, lo-

cal and global matching: $96\% > 890/1445 = 61.6\%$ and $99.9\% > 1234/1445 = 85.4\%$ respectively.

Results showed in Table 3 are indicative of the "coarse grained" capabilities of EFS to discriminate Wikipedia's articles according to quality criteria. Other more detailed analysis, as the one presented in Ingawale et al. (2013), consists in comparing how a metric evaluates in terms of min, max and average values when applied to featured versus non-featured articles. Here, we constrain this analysis to the two codifications of the metric (UBN and CBN) that obtained the highest information gain in relation to both categories (featured and non-featured) when the global matching is used. For both codifications, we present this information by showing their box plots in Figures 2 and 3. In those graphs, it can be seen that values for both codifications of the metric are consistently higher in FA than in NF articles.

For a comprehensive analysis of the metric it also would be interesting to analyse its performance as feature for Wikipedia's article representation in standard text categorization tasks. Space constraints prevent us from doing an exhaustive study of this type but, in a similar way to the analysis performed in Stvilia et al. (2005), we present some preliminary results of its performance in a simple binary (FA versus NF) supervised task. With documents represented with the 24 EFS codifications (24 features), global matching, standard 10-fold cross validation tests with decision trees (DTs)[6] and (backpropagation) multi-layer neural networks (NNs) produced the following results: DT obtained an accuracy of 85.9382 with 1919 correctly classified instances out of 2233 Wikipedia articles; NN, on the other hand, correctly classified 1970 instances with an accuracy of 88.2221. Those results were better than the best ones obtained with the *word count* measure (Blumenstock, 2008) with an accuracy of 82.8221 and the *factual density* quality metric (Lex et al., 2012) that achieved an accuracy of 79.2229. We also tested the capabilities of our metric in non-supervised (clustering) categorization tasks (Pinto Avendaño, 2009), with a simple k-means algorithm with $k = 2$ and the five codifications with the highest information gain (UBN, CBN, TBL, TBF and TBW). In this case, these codifications were able of generating good groupings of FA and NF articles with only 26.562% of instances incorrectly classified. That result is similar to the one reported in Stvilia et al. (2005) where an article representation with seven different quality metrics obtained 25% of instances incorrectly classified with k-means algorithm.

5 Conclusions and Future Works

Using "external" information to assess the IQ of a document seems to be an interesting idea already posed by Juffinger, Granitzer, and Lex (2009) in the context of a blog credibility ranking task. Magdy and Wanas (2010) also measure the support of textual documents by using very basic facts derived from Noun-to-Noun phrases of a document. These facts are compared to those obtained from the information retrieved by a well known search engine (Bing). The procedure used to obtain facts, how the match between facts is determined and the used external resource differ from the ones used in this article. However, it could be considered as the previous work closest to our idea of "external factual support".

In the present article, the motivations behind our EFS metric, its formal definition and the main implementation aspects were introduced. Different data sets for research in quality metrics for Wikipedia were generated, described and made available for other researchers. They include plain texts of high and low quality Wikipedia articles and data sets with the values of the proposed metric in its 24 variants (see Table 1).[7] In this context, preliminary statistics obtained with the EFS metric show its capability to (coarse-grained) filtering and more fine numeric analysis of featured versus non-featured articles. This good performance, was also observed in some basics experiments using EFS codifications as representation features for categorization tasks of Wikipedia's articles. As future work, we want to do an exhaustive study on this point comparing the performance of EFS codifications with the obtained with other state of the art proposals in the area (Lipka and Stein, 2010) and other approaches based on factual information (Lex et al., 2012). In

[6] Weka's J4.8 implementation of the decision tree learning algorithm C4.5.

[7] Those interested readers can contact the first author of the article to have access to these collections.

	Featured Articles (FA)	Non-featured Articles (NF)
Data Set (DS)	$\|FA\| = 1000$	$\|NF\| = 1445$
Reduced Data Set ($DS^\star$)	$\|FA^\star\| = 346$	$\|NF^\star\| = 153$

Table 2: Data sets description - Strict matching

	DS	$DS^\star$ - Strict Matching	$DS^\star$ - Local Matching	$DS^\star$ - Global Matching
FA	$\|FA\| = 1000$	$\|FA^\star\| = 346$	$\|FA^\star\| = 960$	$\|FA^\star\| = 999$
NF	$\|NF\| = 1445$	$\|NF^\star\| = 153$	$\|NF^\star\| = 890$	$\|NF^\star\| = 1234$

Table 3: Data sets description - Strict, Local and Global matching

the last case, the focus will be on determining to what extent EFS information extend/complement "internal" factual information present in the analysed Wikipedia article. Besides, other more elaborated matching mechanisms and external sources will be considered.

Finally, the feasibility of using the proposed EFS metric in other domains beyond the Wikipedia encyclopedia will also be considered. We believe that our EFS metric can be easily applied to domains such as blogs content and news. To do that, we have to analyse which would be adequate external sources for the computation of the external support of the facts.

References

Anderka, M. and B. Stein. 2012. Overview of the 1st int, competition on quality flaw prediction in wikipedia (CLEF 2012).

Anderka, M., B. Stein, and N. Lipka. 2012. Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. In *35rd Annual Int. ACM SIGIR Conf. on Research and Development in Inf. Retrieval*. ACM.

Anthony, D., S. Smith, and T. Williamson. 2009. Reputation and reliability in collective goods: The case of the online encyclopedia wikipedia. *Rationality & Society*, 21(3):283–306.

Baeza-Yates, R. 2009. User generated content: how good is it? In *3rd Workshop on Information Credibility on the Web (WICOW'09)*, pages 1–2. ACM.

Blumenstock, J. E. 2008. Size matters: word count as a measure of quality on Wikipedia. In *17th Int'l Conference on World Wide Web*. ACM.

Emigh, W. and S. Herring. 2005. Collaborative authoring on the Web: a genre analysis of online encyclopedias. In *Proc. of the 38th annual Hawaii int. conference on system sciences*. IEEE CS.

Etzioni, O., M. Banko, S. Soderland, and D. Weld. 2008. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.

Fader, A., S. Soderland, and O. Etzioni. 2011. Identifying relations for open information extraction. In *Proc. of the Conf. of Empirical Methods in Natural Language Processing*, Scotland.

Ferretti, E., M. L. Errecalde, M. Anderka, and B. Stein. 2014. On the use of reliable-negatives selection strategies in the PU learning approach for quality flaws prediction in wikipedia. In *25th Int. Workshop on Database and Expert Systems Applications*, pages 211–215.

Hu, M., E. Lim, A. Sun, H. Lauw, and B. Vuong. 2007. Measuring article quality in Wikipedia: models and evaluation. In *16th ACM International CIKM'07*, pages 243–252. ACM.

Ingawale, M., A. Dutta, R. Roy, and P. Seetharaman. 2013. *Network analysis of user generated content quality in Wikipedia. Online Information Review*, 37(4):602–619.

Juffinger, A., M. Granitzer, and E. Lex. 2009. Blog credibility ranking by exploiting verified content. In *Proc. of WICOW 2009*, pages 51–58. ACM.

Lex, E., M. Völske, M. Errecalde, E. Ferretti, L. Cagnina, C. Horn, B. Stein, and M. Granitzer. 2012. Measuring the quality of web content using factual information. In *2nd joint WICOW/AIRWeb Workshop on Web quality*. ACM.

Lih, A. 2004. Wikipedia as participatory journalism: reliable sources? Metrics for evaluating collaborative media as a news resource. In *5th Int. Symposium on Online Journalism*, pages 16–17.

Lipka, N. and B. Stein. 2010. Identifying featured articles in wikipedia: writing style matters. In *Proc. of the 19th int. conference on World wide web*, WWW '10, pages 1147–1148, NY, USA. ACM.

Magdy, A. and N. Wanas. 2010. Web-based statistical fact checking of textual documents. In *Proc. of SMUC '10*, pages 103–110, NY, USA. ACM.

Pinto Avendaño, D. 2009. On Clustering and Evaluation of Narrow Domain Short-Text Corpora. *Procesamiento del Lenguaje Natural*, 42:129–130.

Stvilia, B., M. Twidale, L. Smith, and L. Gasser. 2005. Assessing information quality of a community based encyclopedia. In *10th Int. Conf. on Information Quality*, pages 442–454. MIT.

Viégas, F., M. Wattenberg, and K. Dave. 2004. *Studying Cooperation and Conflict Between Authors with History Flow Visualizations*. In *Proc. of the SIGCHI Conf.*, pages 575–582. ACM.

Wilkinson, D. and B. Huberman. 2007. *Cooperation and Quality in Wikipedia*. In *Proc. of the 2007 Int. Symposium on Wikis*, pages 157–164. ACM.

Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 101-108 recibido 30-11-2016 revisado 13-01-2017 aceptado 07-02-2017

Analysis of patient satisfaction in Dutch and Spanish online reviews

Análisis de la satisfacción del paciente a partir de comentarios online escritos en holandés y en español

Salud María Jiménez-Zafra,
M. Teresa Martín-Valdivia,
Sinai Group
Universidad de Jaén
Campus Las Lagunillas s/n. E-23071
{sjzafra, maite}@ujaen.es

Isa Maks,
Rubén Izquierdo
Computational Linguistics
VU University Amsterdam
De Boelelaan 1105, 1081 HV
{isa.maks,ruben.izquierdobevia}@vu.nl

Abstract: Sentiment Analysis is a well-known task of Natural Language Processing that has been studied in different domains such as movies, phones or hotels. However, other areas like medical domain remain yet unexplored. In this paper we study different polarity classification techniques applied on health domain. We present a corpus of patient reviews composed by a Dutch part (COPOD: Corpus of Patient Opinions in Dutch) and a Spanish part (COPOS: Corpus of Patient Opinions in Spanish). Experiments have been carried out using a supervised method (SVM), a cross-domain method (OpeNER) and a dictionary lookup method for both languages. Obtained results overcome the baseline in almost all the cases and are higher than other polarity classifiers in patient domain. Regarding the bilingualism, the developed systems for Dutch and Spanish have a similar performance for F1-measure and Accuracy.
Keywords: Polarity classification, medical domain, patient opinion corpus, opener

Resumen: El Análisis de Sentimientos es una tarea del Procesamiento del Lenguaje Natural que ha sido estudiada en diferentes dominios como el de películas, teléfonos móviles u hoteles. Sin embargo, otras áreas como el dominio médico no han sido exploradas todavía. En este trabajo presentamos un corpus de opiniones de pacientes formado por una parte en holandés (COPOD: Corpus of Patient Opinions in Dutch) y por otra parte en español (COPOS: Corpus of Patient Opinions in Spanish). Además, se han realizado diferentes experimentos en ambas lenguas utilizando un método supervisado (SVM), una aproximación basada en *cross-domain* y un método basado en diccionario. Los resultados obtenidos superan el método base en casi todos los casos e incluso los resultados de otros clasificadores de polaridad en el dominio del paciente. Con respecto al bilingüismo, los sistemas desarrollados para holandés y español proporcionan resultados similares para las medidas F1 y Accuracy.
Palabras clave: Clasificación de la polaridad, dominio médico, corpus de opiniones de pacientes, opener

1 Introduction

The examination of patients conducted by specialists when they suffer from some disease can mainly generate two types of information: i) medical reports with the personal and professional observations of physicians and ii) patient experiences. The experiences of these patients are sometimes published on the Internet generating a valuable information source that may contain not only facts but also opinions.

The field of study that analyses people's opinions, sentiments, evaluations, attitudes, and emotions from written language is known as Sentiment Analysis (SA) (Liu, 2012). In last years, the development and study of techniques for SA has been increased due to the vast amount of opinionated documents written on the Internet. Most of studies to date have focused on extracting opinions from user

ISSN 1135-5948

generated reviews in non-medical domains such as movies, phones or hotels. However, the study conducted by Fox and Duggan (2013) states that more than 85% of U.S. Internet users search online for health information. In addition, some platforms such as PatientsLikeMe[1] or Patient Opinion[2] expressing opinions related to health issues are becoming very popular. However, most of research on medical SA has been focused on English although interest in health-related information published in languages other than English is worldwide growing. For example, Van de Belt et al. (2013) present a study related to the preferences of the Dutch population revealing that 83% of people use the Internet as the main source for health-related information. In addition, 42.3% of Dutch population indicates that they sometimes search online for health-related information before visiting a physician. According to a study in 2015[3], 62% of the Spanish people consult the Internet to be informed about topics related to the health.

In this work we present a corpus of Dutch and Spanish patient reviews and apply sentiment analyses techniques in order to classify the reviews as positive or negative. Different methods are applied and compared: a supervised method (SVM), a cross-domain method (OpeNER) and a dictionary lookup method.

The rest of the paper is organized as follows: First we present an overview of related works, section 3 introduces the corpus, and section 4 describes the different methods. In section 5 the different experiments and their results are given. In section 6 the results are analyzed and discussed and we finish with section 7 that refers to conclusions and future works.

2 Related works

Despite of research in medical SA is scarce, there are some works dealing with opinions and sentiments in medical documents. A good review can be found in (Denecke and Deng, 2015). They consider 3 main areas of research in medical context according to the textual source: biomedical literature, clinical notes and medical web content. In this work we deal with patient opinions posted on the Web and we focus on polarity classification in order to identify whether the opinion expressed in a review is positive or negative. Thus, we are going to make a revision of the main studies on binary polarity classification in medical web content.

Qiu et al. (2011) present an interesting social science work in order to study how cancer survivors and caregivers benefit from participation in an online health community. The authors also apply Machine Learning (ML) techniques to analyse sentiment of 298 posts randomly selected from the Cancer Survivors Network[4]. AdaBoost with lexical and style features is the classifier that provides the best accuracy (79.2%).

Na et al. (2012) propose a rule-based linguistic approach for clause-level sentiment classification using existing resources such as UMLS, MPQA, SentiWordNet and Meta-Map. They test the approach over a set of 1,000 clauses of drug reviews manually labelled achieving an F1-measure of 0.70. Bobicev et al. (2012) build a corpus of tweets containing Personal Health Information and apply different ML algorithms to classify the sentiment expressed on the tweets with strong agreement between the annotators (669 tweets). The best result is obtained with a Naïve Bayes classifier (F1-measure=0.77).

Greaves et al. (2013) apply ML techniques to classify 6,412 online comments of patient experiences in hospitals of the English National Health Service. They also conduct binary classification experiments but using SA to capturing patient experience from texts. Their goal is to automatically predict whether a patient would recommend a hospital or not, whether the hospital was clean or not and whether the patient was treated with dignity or not. The best F1-measure values obtained in these experiments were 0.89 for hospital recommendation, 0.87 for cleanliness and 0.85 for respect. The algorithms that provided better results were multinomial Naïve Bayes and Bagging. Biyani et al. (2013) perform sentiment classification of user posts in an online health community (Cancer Survivor Network[5]) by exploiting domain-specific and general information features about sentiment expression and combining them in a semi-supervised setting using a co-training algorithm. The approach is tested on a set of

[1] https://www.patientslikeme.com

[2] https://www.patientopinion.org.uk/

[3] http://insights.doctoralia.es/informe-doctoralia-sobre-salud-e-internet-2015/

[4] http://csn.cancer.org

[5] http://www.csn.cancer.org

293 posts getting an F1-value of 0.84. Later, this work was extended in (Ofek et al., 2013). The authors show that classifiers trained using abstract features extracted from a dynamic sentiment lexicon outperform those trained using features extracted from a general sentiment lexicon. The number of features is reduced from thirteen to six and they obtain an F1-value of 0.81 with a Random Forest classifier.

Sharif et al. (2014) propose a representational richness framework that they evaluate on the AskaPatient dataset, a collection of 114,000 forum posts. The framework leverages novel feature representations that extract underlying sentiments in medical social media content. The feature set can be categorized into four categories: baseline features, semantic features, emotion related features and domain specific features. For evaluation, 24,000 posts are used for training and 90,000 for test with an SVM classifier getting an accuracy of 78.2%. Melzi et al. (2014) also apply an SVM classifier on a set of 3,000 sentences related to messages collected on the English-language Spine Health website. The best result is obtained with unigrams, bigrams, emotion words and patterns (F1 = 0,66).

All these studies are focused on English. However, recently, a corpus of 743 Spanish patient opinions extracted from the medical web Masquemedicos[6] has been presented (Plaza-del Arco et al., 2016). In order to demonstrate the usefulness of the resource, the authors conduct experiments using a general lexicon and a machine learning approach for the polarity classification of the reviews obtaining an F1 value of 0.72 and 0.71 respectively. This corpus is used in this paper in order to compare patient experiences in Dutch and Spanish, two languages of growing interest in health-related issues on the Web 2.0.

3 Resources

In this section, a detailed description is provided of the main resources employed in the experimental framework. We built a corpus of patient reviews that consists of a Dutch part (COPOD: Corpus of Patient Opinions in Dutch) and a Spanish part (COPOS : Corpus of Patient Opinions in Spanish).

COPOD, the Dutch part of the corpus, has been built by crawling the well-known medical forum Zorgkaart Nederland[7] on June 28, 2016. It is composed of 156,975 patient reviews about their experiences with physicians of 60 specialties. Each review contains information about the medical entity and the patient's opinion. In relation to the medical entity, the following elements have been extracted: the name, the profession, the specialty of the doctor and the city where the consultation was performed. With respect to the patient's opinion the following information has been included: the review text, the date, the disease treated, a rating for different aspects and an overall rating. The overall rating refers to a scale from 1 to 10 stars and corresponds to the average of the ratings of the different aspects (appointment, therapy, staff attention, information, listening, and accommodation). The number of reviews per rating is shown in Table 1. In addition, statistics of the corpus are shown in Table 3.

Rating (rt)	Reviews
$1 < rt <= 2$	1,781
$2 < rt <= 3$	2,302
$3 < rt <= 4$	3,690
$4 < rt <= 5$	4,290
$5 < rt <= 6$	3,459
$6 < rt <= 7$	2,870
$7 < rt <= 8$	12,465
$8 < rt <= 9$	49,577
$9 < rt <= 10$	76,541
Total	156,975

Table 1: COPOD - Reviews per rating

COPOS, the Spanish part of the corpus, is the first corpus of patient opinions in Spanish. It consists of 743 reviews about medical entities of 34 specialties that were extracted by crawling the medical forum Masquemedicos[8] on December 3, 2015. Each review contains information about the name and specialty of the medical entity, the city where the consultation was performed, the textual opinion, the date when the opinion was written and an overall evaluation with stars (from 0 to 5 stars). The number of reviews per rating can be seen in Table 2. Furthermore, some interesting features of the corpus are shown in Table 3.

[6] http://masquemedicos.com/

[7] https://www.zorgkaartnederland.nl/

[8] http://masquemedicos.com/

Rating (rt)	Reviews
0	3
1	88
2	18
3	35
4	51
5	548
Total	743

Table 2: COPOS - Reviews per rating

	COPOD	COPOS
#Sentences	534,317	2,009
#Words	7,341,779	32,365
Avg. sentences per review	3.4	2.7
Avg. words per sentence	13.7	16.1
Avg. words per review	46.7	43.6
#Adjectives	916,046 (12%)	3,002 (9%)
#Adverbs	540,355 (7%)	2,282 (7%)
#Nouns	1,173,732 (15%)	7,393 (22%)
#Verbs	1,111,525 (15%)	5,593 (17%)

Table 3: COPOD statistics

4 Methods

In this section we present the methods that have been used in the current research. We have carried out a set of experiments using an SVM classifier, a dictionary lookup method and OpeNER tool aiming at a binary classification of reviews.

4.1 Method I: SVM classifier

SVM classifier is based on the principle of Structural Risk Minimization of the computational learning theory (Vapnik, 2013). The theory is founded on the seeking of the hyper plane that maximizes the margin of separation between the objects belonging to two different classes. For the experimentation related to this study we applied 10-fold cross-validation and we used TF-IDF as weighting scheme and the libSVM implementation (Chang and Lin, 2011) with the following parameters: linear kernel, C=0.0 and epsilon=0.001.

4.2 Method II: Dictionary lookup method

The dictionary lookup method is a rule-based approach which starts from the review text and uses a sentiment lexicon to find positive and negative words in the review. The approach is a vote-algorithm: for each review the number of matched positive and negative words from the sentiment lexicon are counted. We then assign the majority polarity to the review. In the case of a tie positive polarity is assigned. The sentiment lexicon used to analyze the Dutch reviews is an automatically derived language sentiment lexicon obtained by WordNet propagation (cf. (Maks et al., 2014)). As this lexicon includes only lemma and part-of-speech we first lemmatized the text with the Dutch Alpino-parser[9]. The analysis of the Spanish reviews uses iSOL (Molina-González et al., 2013), that has been generated by translating into Spanish the Bing Liu English Lexicon. Both lexicons are general language lexicons and have not been adapted for the medical domain.

4.3 Method III: OPeNER cross-domain method

The cross-domain method makes use of the classifier that has been developed in the framework of the OpeNER project[10]. This project strived for the development of different opinion mining and sentiment analysis tools for several European languages including Dutch and Spanish. The set of tools includes classifiers that use Conditional Random Fields and are designed for finding opinion expressions in text. The tools have been trained on hotel reviews and our experiments aim at finding out how well these models can be applied in other domains. As the task at hand aims at classification at document level, instead of expression level, we calculate an overall opinion score by subtracting the number of negative expressions from the number of positive ones. If the result is smaller than zero the review is considered negative otherwise it is considered positive.

5 Experiments

As we have mentioned in Section 3, COPOS and COPOD were built in a similar way. Actually, both web sites (Zorgkaart Nederland

[9]http://www.let.rug.nl/ vannoord/alp/Alpino/
[10]http://www.opener-project.eu

	#Negative reviews	#Positive reviews	Total
COPOD	12,063	144,912	156,975
COPOS	109	634	743
COPOD-743	109	634	743

Table 4: Reviews per Corpus

and Masquemedicos) have analogous content about comments related to hospitals and doctors. Perhaps, the main difference lies in the rating scale since opinions in COPOD are ranking from 1 to 10 stars whereas in COPOS the scale is from 0 to 5 stars. On the other hand, COPOS is much smaller than the Dutch corpus. Due to this fact, we have also created a selection of COPOD that consists of 743 reviews with a similar distribution across rating categories as the Spanish corpus (COPOD-743). In this way, we can better compare the results using two comparable corpus in two different languages: Dutch and Spanish.

We have carried out a set of experiments using the SVM classifier, the dictionary lookup method and the OPeNER classifier. In our experiments we focus on a binary classification of the reviews. In order to select the positive and negative examples, we consider COPOD reviews with more than 5 stars as positive and the remaining reviews as negative. On the other hand, we consider positive reviews in COPOS if they have 3, 4 or 5 stars, and negative ones if their rating is 0, 1 or 2 stars. A summary of the number of reviews that composed each set is shown in Table 4.

In order to evaluate the different methods we calculated the usual measures per class: Precision (P) and Recall (R).

$$P_c = \frac{TP_c}{TP_c + FP_c} \quad (1)$$

$$R_c = \frac{TP_c}{TP_c + FN_c} \quad (2)$$

where TP (True Positives) are those assessments where the system and human experts agree on a label assignment, FP (False Positives) are those labels assigned by the system that do not agree with the expert assignment, FN (False Negatives) are those labels that the system failed to assign as they were given by the human expert, and TN (True Negatives) are those non-assigned labels that were also discarded by the expert. The Precision tells us how well the labels are assigned by our system (the fraction of assigned labels that are correct) whereas the Recall measures the fraction of the expert's labels found by the system. Finally, Precision and Recall are combined using the Macro-averaged F1 and Accuracy is measured in order to take into account all the correct results including TN (Sebastiani, 2002):

$$F1 - measure = \frac{1}{\mid c \mid} \sum_{i=1}^{|c|} \frac{2P_iR_i}{P_i + R_i} \quad (3)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

We created baseline measures assigning the most frequent class, i.e. positive to all reviews (cf. Table 5). Following, the experimentation results obtained with the different approaches over the three data sets are shown in Table 6, Table 7 and Table 8. It can be seen that libSVM provides the best accuracy results in all the datasets whereas the dictionary-based approach is around baseline and OpeNER is a bit below it.

6 Result analysis

According the corpus statistics, both COPOD and COPOS include a higher number of positive than negative opinions. It seems that patients are in general quite satisfied with the doctor's visit or they tend to write rather about their good experiences than about the bad ones.

One of the most salient results is that negative reviews have low performance across all methods when compared to positive reviews. With respect to the SVM method the reason can be found in the relative low number of negative training data but scarcity cannot affect the performance of the other methods. When comparing positive and negative reviews we found some characteristics that might explain it better. First of all, negative reviews are longer than positive ones: in COPOD the average length of positive reviews is 43.8 words

	COPOD	COPOS	COPOD-743
Majority baseline (positive class)	0.92	0.85	0.85

Table 5: Majority baselines

	COPOD	COPOS	COPOD-743
Precision negative class	0.76	0.90	0.91
Recall negative class	0.70	0.17	0.41
Precision positive class	0.97	0.88	0.89
Recall positive class	0.98	0.99	0.99
F1 measure	0.86	0.71	0.78
Accuracy	0.96	0.88	0.90

Table 6: Results for SVM

	COPOD	COPOS	COPOD-743
Precision negative class	0.52	0.60	0.66
Recall negative class	0.68	0.46	0.73
Precision positive class	0.97	0.91	0.94
Recall positive class	0.94	0.95	0.92
F1 measure	0.77	0.73	0.81
Accuracy	0.92	0.87	0.89

Table 7: Results for Dictionary based approach

	COPOD	COPOS	COPOD-743
Precision negative class	0.56	0.60	0.70
Recall negative class	0.10	0.08	0.12
Precision positive class	0.94	0.86	0.85
Recall positive class	0.98	0.99	0.99
F1 measure	0.56	0.53	0.56
Accuracy	0.91	0.86	0.85

Table 8: Results for OpeNER

whereas the average length of negative reviews is 71.8, and in COPOS these values are 38.2 and 74.5, respectively. A closer look at the texts reveals that negative reviews tend to describe events with a lot of detail and relatively often contain contextual opinions that require a wider context for correct interpretation. For example, one of the reviews contains the -in this case negative- expression *and after that an extra operation was needed*. It is only the broader context that explains that this extra operation was needed after an earlier surgery that went wrong for unnecessary reasons. This kind of expressions are hard to automatically identify and classify. Secondly, we noted that relatively many negative reviews are in the middle rating categories (3,4 and 5 stars for COPOD and 2 and 3 for COPOS) whereas most positive reviews are in the extreme rating categories (9 and 10 stars for COPOD and 5 for COPOS). That may also explain low performance on the negative class as earlier research (cf. (Maks and Vossen, 2013)) already showed that reviews of the middle rating categories are hard to classify because they often include a mixture of positive and negative opinions. For example, in the specific case of COPOD we have realized that comments rating with 6, 7 or even 8 stars could be considered semantically negative when you read the textual information, although we have taken as negative reviews until 5 stars. Thus, a better partition of the corpus considering for instance negative re-

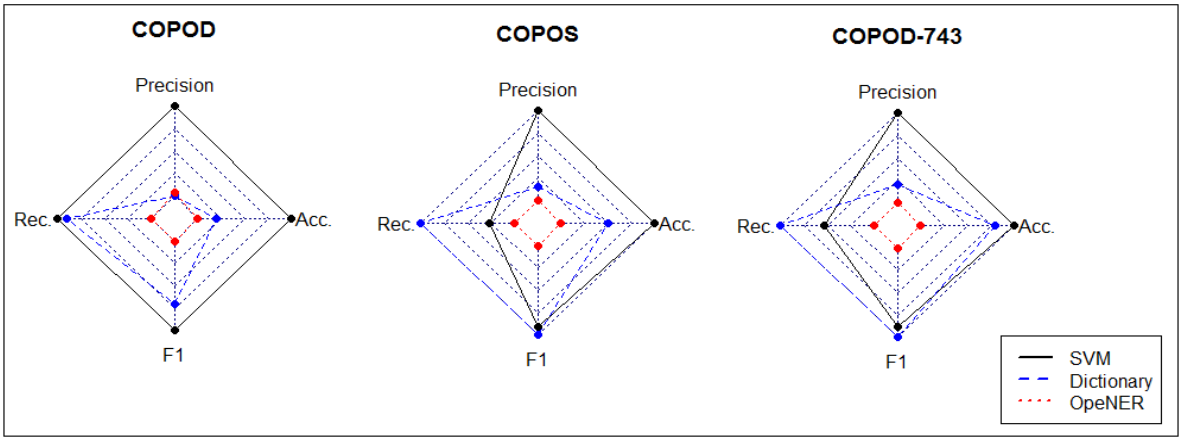

Figure 1: Analysis of classification methods using radar charts (Larger area implies better overall performance of the method)

views between 0 and 7 stars and positive ones whose between 8 and 10 could improve the final results and help to correctly classify negative examples.

On the other hand, it is interesting to note that the behavior of the systems are very similar for both languages, always presenting a better performance for positive class than for negative one. As expected the dictionary lookup method and the OpeNER method have lower overall performance than SVM (Figure 1) as both are methods not adapted for the medical domain. Another difference is that SVM's bag of words approach works at document level whereas the other 2 methods work at expression level identifying and classifying each separate opinion expression. Although expression-level classification may not be the best approach for the current task we think that is needed for more fine-grained tasks such as, for example, aspect-based sentiment analysis.

7 Conclusion

In this paper we have presented a corpus with patient reviews written in Dutch and Spanish. We have conducted different experiments using a supervised method, a cross-domain method and a dictionary lookup method.

Research in medical domain for SA is very scarce and this paper present a background with the main works of the area. On the other hand, most of research is focused on English although interest in subjective medical information is growing in other languages. For this reason, we have centered our work on Dutch and Spanish and we have presented several approaches to tackle the problem. The results show low differences between languages and, although the SVM method has a better performance, the dictionary approach also reaches good accuracy. Perhaps the worst result is obtained with the cross-domain approach, but we must take into account that the OpeNER tool has been trained over the tourism domain and it has been directly applied to the medical domain.

Finally, we consider this paper as a preliminary research and our future work will be focused on other issues related to SA for health such as the study of aspect-based SA in medical domain using the generated corpus or the generation of resources adapted to the medical domain.

Acknowledgments

This work has been partially supported by a grant from the Ministerio de Educación Cultura y Deporte (MECD - scholarship FPU014/00983), Fondo Europeo de Desarrollo Regional (FEDER) and REDES project (TIN2015-65136-C2-1-R) from the Spanish Government.

References

Biyani, P., C. Caragea, P. Mitra, C. Zhou, J. Yen, G. E. Greer, and K. Portier. 2013. Co-training over domain-independent and domain-dependent features for sentiment analysis of an online cancer support community. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2013*, pages 413–417. IEEE.

Bobicev, V., M. Sokolova, Y. Jafer, and

D. Schramm. 2012. Learning sentiments from tweets with personal health information. In *Canadian Conference on Artificial Intelligence*, pages 37–48. Springer.

Chang, C.-C. and C.-J. Lin. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Denecke, K. and Y. Deng. 2015. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial intelligence in medicine*, 64(1):17–27.

Fox, S. and M. Duggan. 2013. Health online 2013. *Health*, pages 1–55.

Greaves, F., D. Ramirez-Cano, C. Millett, A. Darzi, and L. Donaldson. 2013. Use of sentiment analysis for capturing patient experience from free-text comments posted online. *Journal of medical Internet research*, 15(11):e239.

Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Maks, I., R. Izquierdo, F. Frontini, R. Agerri, and P. Vossen. 2014. Generating Polarity Lexicons with Wordnet propagation in five languages. In *Proceedings of LREC2014*, Reykjavik.

Maks, I. and P. Vossen. 2013. Sentiment Analysis of Reviews: Should we analyze writer intentions or reader perceptions? In *Proceedings of RANLP 2003*, pages 415–419, Hissar, Bulgaria.

Melzi, S., A. Abdaoui, J. Azé, S. Bringay, P. Poncelet, and F. Galtier. 2014. Patient's rationale: Patient Knowledge retrieval from health forums. In *eTELEMED: eHealth, Telemedicine, and Social Medicine*.

Molina-González, M. D., E. Martínez-Cámara, M.-T. Martín-Valdivia, and J. M. Perea-Ortega. 2013. Semantic orientation for polarity classification in spanish reviews. *Expert Systems with Applications*, 40(18):7250–7257.

Na, J.-C., W. Y. M. Kyaing, C. S. Khoo, S. Foo, Y.-K. Chang, and Y.-L. Theng. 2012. Sentiment classification of drug reviews using a rule-based linguistic approach. In *International Conference on Asian Digital Libraries*, pages 189–198. Springer.

Ofek, N., C. Caragea, L. Rokach, P. Biyani, P. Mitra, J. Yen, K. Portier, and G. Greer. 2013. Improving sentiment analysis in an online cancer survivor community using dynamic sentiment lexicon. In *Social Intelligence and Technology (SOCIETY), 2013 International Conference on*, pages 109–113. IEEE.

Plaza-del Arco, F. M., M. T. Martín-Valdivia, S. M. Jiménez-Zafra, M. D. Molina-González, and E. Martínez-Cámara. 2016. COPOS: Corpus Of Patient Opinions in Spanish. Application of Sentiment Analysis Techniques. *Procesamiento del Lenguaje Natural*, 57:83–90.

Qiu, B., K. Zhao, P. Mitra, D. Wu, C. Caragea, J. Yen, G. E. Greer, and K. Portier. 2011. Get online support, feel better–sentiment analysis and dynamics in an online cancer survivor community. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 274–281. IEEE.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, March.

Sharif, H., F. Zaffar, A. Abbasi, and D. Zimbra. 2014. Detecting adverse drug reactions using a sentiment classification framework.

Van de Belt, T. H., L. J. Engelen, S. A. Berben, S. Teerenstra, M. Samsom, and L. Schoonhoven. 2013. Internet and social media for health-related information and communication in health care: preferences of the Dutch general population. *Journal of medical Internet research*, 15(10):e220.

Vapnik, V. 2013. *The nature of statistical learning theory*. Springer Science & Business Media.

Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 109-116 recibido 30-11-2016 revisado 13-01-2017 aceptado 13-02-2017

Spanish Morphological Generation with Wide-Coverage Lexicons and Decision Trees

Generación Morfológica del Español con Lexicones de Amplia Cobertura y Árboles de Decisión

Daniel Ferrés, Ahmed AbuRa'ed and Horacio Saggion
Large Scale Text Understanding Systems Lab
TALN - DTIC
Universitat Pompeu Fabra
C/ Tànger, 122-140
08018 Barcelona, Spain
{*daniel.ferres,ahmed.aburaed,horacio.saggion,*}*@upf.edu*

Abstract: Morphological Generation is the task of producing the appropiate inflected form of a lemma in a given textual context and according to some morphological features. This paper describes and evaluates wide-coverage morphological lexicons and a Decision Tree algorithm that perform Morphological Generation in Spanish at state-of-the art level. The Freeling, Leffe and Apertium Spanish lexicons, the J48 Decision Tree algorithm and the combination of J48 with Freeling and Leffe lexicons have been evaluated with the following datasets for Spanish: i) CoNLL2009 Shared Task dataset, ii) Durrett and DeNero dataset of Spanish Verbs (DDN), and iii) SIGMORPHON 2016 Shared Task (task-1) dataset. The results show that: i) the Freeling and Leffe lexicons achieve high coverage and precision over the DDN and SIGMORPHON 2016 datasets, ii) the J48 algorithm achieves state-of-the-art results in all of the three datasets, and iii) the combination of Freeling, Leffe and the J48 algorithm outperformed the results of our other approaches in the three evaluation datasets, improved slightly the results of the CoNLL2009 and SIGMORPHON 2016 reported in the state-of-the-art literature, and achieved results comparable to the ones reported in the state-of-the-art literature on the DDN dataset evaluation.
Keywords: Morphological generation, morphological lexicons, decision trees, natural language generation

Resumen: La Generación Morfológica es la tarea de producir la forma flexionada apropiada de un lemma en un determinado contexto textual y en concordancia con algunas características morfológicas. En este artículo se presentan y se evaluan algunos lexicones morfológicos de amplia cobertura y un algoritmo de árboles de decisión para la Generación Morfólogica en español. Los lexicones para el español Freeling, Leffe y Apertium, el algoritmo de árboles de decisión J48 y la combinación de los lexicones Freeling y Leffe con el J48 han sido evaluados con los siguientes conjuntos de datos para el español: i) conjunto de datos de la CoNLL2009 Shared Task, ii) el conjunto de datos de verbos para el español de Durrett y DeNero (DDN), y iii) el conjunto de datos para el español de la evaluación SIGMORPHON 2016 Shared Task (task-1). Los resultados muestran que: i) los lexicones morfológicos consiguen alta cobertura y precisión en los conjuntos de datos DDN y SIGMORPHON 2016, ii) el algoritmo J48 por si sólo alcanza resultados en el estado del arte en los tres conjuntos de evaluación, y iii) que la combinación de predicciones de Freeling, Leffe y el algoritmo J48 mejora los resultados de nuestras otras implementaciones en los tres conjuntos de datos evaluados, que además mejoran ligeramente los resultados reportados en el estado del arte en los conjuntos de datos del CoNLL2009 y del SIGMORPHON 2016, y que consiguen resultados comparables con los reportados en el estado del arte de la evaluación del conjunto de datos DDN.
Palabras clave: Generador morfológico, lexicones morfológicos, árboles de decisión, generación de lenguaje natural

ISSN 1135-5948

1 Introduction

Morphological Generation is the task of producing the appropiate inflected form of a lemma in a given textual context and according to some morphological features. An example of morphological inflection in Spanish language is shown in Figure 1: the lemma *cantar* (sing) inflected with the verbal morphological features of *number* (plural), *person* (1st), *mode* (indicative), *tense* (imperfect) generates the inflected form *cantábamos* (sang). Morphological Generation is a crucial part of the Surface Realization phase of Natural Language Generation (NLG) systems. NLG for Spanish has been applied in complex applications such as Dialogue Systems (Amores, Pérez, and Portillo, 2006), Machine Translation (Forcada et al., 2011), and Textual Simplification (Bott et al., 2012) among others. Morphological Generation can be performed with the following resources: i) morphological lexicons (Molinero, Sagot, and Nicolas, 2009; Forcada et al., 2011; Padró and Stanilovsky, 2012), ii) hand-made or learned inflected rules or decision trees (Durrett and DeNero, 2013; Nicolai, Cherry, and Kondrak, 2015), and iii) other supervised learning systems (Bohnet et al., 2010; Dušek and Jurcicek, 2013; Ahlberg, Forsberg, and Hulden, 2015; Faruqui et al., 2016; Cotterell et al., 2016; Kann and Schütze, 2016). Morphological lexicons are hand-made or semi-automatically generated dictionaries with inflected forms stored in the following way <*inflected form, lemma, morphological features*>. These lexicons can have wide coverage and achieve high precision but are not able to inflect new unseen lemmas.

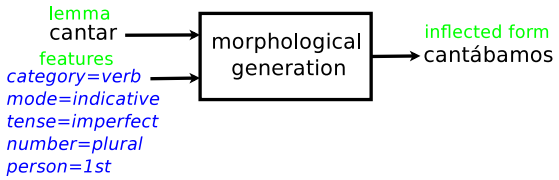

Figure 1: Example of Morphological Generation

Inflection rules can be generated manually or automatically by Rule Induction or Decision Tree learning algorithms. The advantage of rule induction systems over other supervised machine learning technologies is that the models (rules or trees) are human readable and interpretable. Thus these kind of models can be modified and extended with human supervision.

The main contribution of this paper is to present and evaluate a novel approach to Morphological Generation that combines predictions from the free-available and wide-coverage Spanish lexicons Freeling and Leffe with the ones from a human-interpretable J48 Decision Tree model. We will show that our approach achieves state-of-the-art results in coverage, precision and accuracy in Spanish Morphological Generation over several benchmarking datasets. The resource is available online.[1]

2 Related Work

This section describes morphological lexicons and supervised learning approaches to Morphological Generation for Spanish. COES is a morphological tool for Spanish (Rodríguez and Carretero, 1996) which is composed by a lexicon and a set of about 3,500 derivative rules that cover most of the morphological rules of the Spanish language. Freeling[2] (Padró and Stanilovsky, 2012) is an open source language analysis software that has a Spanish dictionary of about 650,000 inflected forms corresponding to 76,000 lemma to Part-of-Speech (PoS) combinations. This dictionary was obtained from the Spanish Resource Grammar (SRG) (Marimon, Seghezzi, and Bel, 2007). *Leffe*[3] (Molinero, Sagot, and Nicolas, 2009) is a wide coverage morphological and syntactic lexicon for Spanish that merged the following high quality existing lexicons for Spanish: Multext, USC, ADESSE, and SRG . Leffe has about 165,000 unique (lemma,PoS) pairs, which correspond to approximately 1,590,000 entries that associate a form with both morphological and syntactic information (approximately 680,000 unique (form,PoS) pairs). Apertium is a free/open-source platform for rule-based machine translation (Forcada et al., 2011) that has a morphological analyzer and generator based on finite state transducers. It also has a dictionary for Spanish[4] with over 46,000 lemmas and morphological inflectional rules that can cover more than 7.9 million inflected forms (of which most are verbal forms with enclitics).

[1] https://www.upf.edu/web/taln/resources
[2] http://nlp.lsi.upc.edu/freeling/
[3] https://gforge.inria.fr/projects/alexina/
[4] https://sourceforge.net/p/apertium/svn/HEAD/tree/languages/apertium-spa/apertium-spa.spa.dix

Unimorph[5] (Kirov et al., 2016) is a multilingual morphological resource extracted from Wiktionary that includes data in Spanish. Bohnet et al. (2010) presented a Support Vector Machine (SVM) based multilingual dependency oriented stochastic deep sentence realizer which has a morphological generator. They used the Levenshtein edit distance to map lemmas to word forms. The input to the classifier is the lemmata of a sentence, its dependency tree and the already ordered sentence. Dušek and Jurcicek, (2013) presented a morphological realizer that uses also edit scripts based on the Levenshtein distance and multi-class logistic regression classifiers. They used some generic features across all languages: lemma, PoS tag, morphological features (e.g. case, gender,...) and suffixes of the lemma up to 4 characters.

Durrett and DeNero, (2013) presented a supervised approach to learn and predict morphological paradigms that automatically acquires the orthographic transformation rules of morphological paradigms from labeled examples, and then learns the contexts in which those transformations apply using a discriminative sequence model. Nicolai, Cherry, and Kondrak, (2015) used supervised inflection generation with discriminative string transduction and reranking. They transform character alignments into inflection rules and select them with a discriminative semi-Markov Model. Ahlberg, Forsberg, and Hulden, (2015) presented a system that learns morphological paradigms and is able to predict inflection tables from unseen lemmas. Their system is based on the longest common subsequence and SVMs. Faruqui et al. (2016) used character sequence to sequence learning for morphological inflection with encoder-decoder neural networks. The SIGMORPHON 2016 Shared Task (task-1)[6] on Morphological Reinflection (Cotterell et al., 2016) covered Morphological Inflection for 10 languages (including Spanish). The LMU system (Kann and Schütze, 2016) achieved the best accuracies for Spanish with 98.84% and 99.05%. They used a character-based sequence-to-sequence attention model called MED (Morphological EncoderDecoder) with an RNN encoder-decoder architecture (Bahdanau, Cho, and Bengio, 2014).

[5] http://ckirov.github.io/UniMorph/

[6] http://ryancotterell.github.io/sigmorphon2016/

3 System Description

Our Morphological Generation system can be executed in three ways: i) lexicon-based morphological inflection, ii) Decision Trees based predictions, and iii) a combination of lexicon-based generation and Decision Trees based predictions.

3.1 Lexicon-Based Inflection

The Lexicon-Based inflection simply uses the information existing in a morphological lexicon to generate a new inflected form. This methodology offers high quality predictions but does not perform predictions on unseen forms in the lexicon. The Morphological Lexicons used in this paper are: Freeling, Apertium (derived forms without enclitics), and Leffe (derived forms) (see in Table 1 some statistics about these lexicons).

3.2 Decision Trees

The Decision Trees algorithm used to predict is the J48 algorithm,[7] an open source Java implementation of the C4.5 algorithm (Quinlan, 1993) in the WEKA[8] data mining tool. The algorithm takes a lemma, PoS and PoS features as input and then generates the proper form according to the morphological features derived from the PoS and the features extracted from the lemma (see the description of these features in Tables 2 and 3).

Feature	Lexical categories
Case	Pronouns (i.e. ordinal, qualificative and possessive in case of Adjectives)
Gender	Adjectives, Determiners Nouns, Pronouns, Verbs (i.e. masculine, feminine, or common)
Mood	Verbs (i.e. indicative, infinitive, subjunctive, gerund, imperative and participle).
Number	Nouns, Verbs, Adjectives (i.e. singular, plural, and invariable)
Person	Determiners, Pronouns, Verbs (i.e. first, second and third person).
Possessornum	Determiners (i.e. singular, plural and invariable)
Polite	Pronouns (i.e. yes, no)
Tense	Verbs (i.e. past, present, future, imperfect and conditional).
Type	Adjectives, Adverbs, Determiners, Numerals (i.e. ordinal, qualificative and possessive in case of Adjectives).

Table 2: Morphological features associated to lexical categories

[7] J48 has been chosen because in our initial experiments it achieved better performance than other human-interpretable algorithms available in WEKA.

[8] http://www.cs.waikato.ac.nz/~ml/weka/

dataset	size (#tokens)								
	nouns		verbs		adjectives		total (all lexical categories)		
	lemmas	forms	lemmas	forms	lemmas	forms	lemmas	forms	CRAE top10k(%)
Apertium	16,668	36,013	3,927	255,253	6,279	23,093	36,253	323,846	92.02
Freeling	49,528	107,638	7,660	497,801	18,618	62,979	76,335	669,216	91.48
Leffe	70,944	154,106	8,359	530,309	28,509	97,136	162,394	852,347	93.51

Table 1: Statistics of the Morphological Lexicons used in the evaluation

Features Set	Description	example
Length	number of characters	6
Last characters	last, penultimate, and antepenultimate characters	r,a,t
Last n-grams	last bigram last trigram	ar,tar
1-grams	all lemma unigrams	c,a,n,t,a,r
1-grams order	all lemma unigrams with associated order position	c1,a2,n3,t4,a5,r6
1-grams reverse order	all lemma unigrams with associated reverse order	c6,a5,n4,t3,a2,r1
2-grams	all lemma bigrams	ca,an,nt,ta,ar
Last 3-gram with a skip	the last three chars skipping the penultimate one	t_r
Penultimate 3-gram with a skip	last penultimate three chars skipping the antepenultimate one	n_a
Phonetics	consonant and vowels in the same way of the character and n-grams previous features	c,v,c, vc,cvc, c_c,c_v

Table 3: Features associated to the lemma. Includes the example of features extracted for the lemma *cantar* (sing)

The system is based on the Levenshtein edit distance algorithm between the lemma and the target word form. The edit distance algorithm calculates how many character operations are needed to transform and edit one string (e.g. lemma) to another (e.g. target word form). The possible operations are: Insert(index, character), implemented by inserting the presented character into the index position of the lemma; Replace(index, character), implemented by replacing the character in the index position of the lemma with the presented character; and Delete(index), implemented by deleting the character at the index position of the lemma (see an example of some edit scripts in Figure 2).

lemma	inflected form	edit script
rehusar	rehusemos	R(0,s) R(1,o) I(2,m) I(2,e)
enviar	envié	R(0,é) D(1)
millón	millones	I(0,s) I(0,e) R(1,o)
joven	jóvenes	I(0,s) I(0,e) R(3,ó)
hermoso	hermosísimo	I(1,m) I(1,i) I(2,í) I(2,s)
averiar	averiado	R(0,o) I(1,d)

Figure 2: Examples of edit scripts

The index starts from the last character of the lemma to the first (i.e. 0 indicates the last character, 1 indicates the last-1, etc...). For example: in order to obtain the verb form *envié* (sent) the operations *R(0,é)* and *D(1)* are supposed to be applied on the lemma *enviar* (send). The implementation starts with D(1) deleting the letter *'a'* then R(0,é) will replace the last letter with *'é'*. The J48 algorithm constructs a decision tree for each lexical category. The lexical categories that can be used by the system are common nouns, verbs, adjectives, adverbs, pronouns, determiners, and numerals. The decision tree built will learn a mapping index which refers to a sequence of operations to transform a lemma to an inflected form according to the lemma based and PoS based features.[9] After training, a J48 model is learnt in which the system takes a lemma, PoS and PoS features as input and then generates a test instance based on the training features from the input data, the model will predict a mapping index which refers to a sequence of operations forming the edit script as shown at Figure 2. The operations are implemented from right to left order in the sequence of edit scripts after applying the operations sequence on the lemma the final form will be obtained.

3.3 Lexicon and Decision Trees

This configuration combines lexicon-based and Decision Trees based predictions. This configuration gives priority to predictions that can be obtained with the lexicon; thus trying to ensure a high precision because of the wide-coverage of frequent forms by the lexicon. This wide-coverage is shown in Table 1: a significant number of inflected forms in the lexicons (more than 852,000 in the case Leffe lexicon) and the high coverage (about 92%-93%) of the *Corpus de Referencia del Español Actual (CREA)* corpus top 10,000

[9]Some specific features for verbs contained in the SIGMORPHON dataset are not described here. These features are *Alt*, *Aspect*, *Polar*, and *Polite*.

frequent words[10] in Spanish. The system works in the following way: i) firstly the system seeks if there are inflected forms associated to the lemma and the morphological features in the lexicon(s) and selects them[11], ii) otherwise, if no inflected forms are found in the lexicon, then the system will execute the J48 Decision Tree classifier associated to the lexical category of the current lemma to be inflected with the morphological features and the lemma based features as input data to predict the edit operations necessary to generate the new inflected form. The combinations of lexicons and J48 used are the following ones: i) Freeling executed in combination with the J48 algorithm (Freeling + J48), ii) Leffe executed in combination with the J48 algorithm (Leffe+J48), and iii) Freeling executed in combination with Leffe and the J48 algorithm (F+L+J48). The last combination uses firstly the Freeling lexicon to find the inflected form, otherwise uses the Leffe (if Freeling fails to retrieve an inflected form), and otherwise uses J48 (if both Freeling and Leffe fail).

4 Evaluation

The evaluation of the systems presented was performed using separately the following datasets for Spanish (described below): i) CoNLL2009 Shared task Dataset for Spanish, ii) Durrett and DeNero datasets for Spanish Verbs, iii) SIGMORPHON 2016 Shared Task task-1 dataset for Spanish. The size in tokens of the training, development and evaluation splits of the datasets is reported in Table 4. The training split is used to train the Decision Trees models, the evaluation split is used for evaluation, and the development split is not used.

The experiments presented in this evaluation are designed to evaluate the following sets of measures: i) the *coverage*, *precision* and *accuracy* of the morphological lexicons over the datasets, and ii) the *accuracy* of the J48 algorithm applied over the datasets predicting alone or in combination with the morphological lexicons. The *coverage* measure of the morphological lexicons tell us about which percentage of the evaluation dataset can be automatically predicted with the lexicon and without performing prediction based on supervised learning. On the other hand, the *precision* measure indicates the percentage of this coverage that is correctly inflected. It is supposed that because lexicons have been produced by human experts, the inflections derived from a lexicon will have more confidence compared with the ones predicted by a supervised learning algorithm. The coverage and precision measures will be measured only over the DDN and SIGMORPHON 2016 datasets, because the CoNLL2009 includes the numeric lexical category not present in the lexicons evaluated. The accuracy measures will be calculated in all three datasets and with respect to some specific lexical categories in the case of CoNLL2009 (7 categories) and SIGMORPHON 2016 (3 categories). The CoNLL2009 evaluation will include the accuracy of some token subsets: tokens excluding the punctuation, only inflected forms, and only unseen forms in training set.

4.1 CoNLL2009 Dataset

The CoNLL2009 Shared Task[12] (Hajič et al., 2009) is to predict syntactic and semantic dependencies and their labeling. The Spanish datasets were generated from the AnCoraES[13] corpora (Taulé, Martí, and Recasens, 2008), a multilevel annotated corpora for Spanish (mainly news). It has about 528,000 tokens annotated manually, semi-automatically, or fully automatically. The data size of the training, test and development datasets for Spanish is 427,442, 50,630 and 50,368 tokens respectively. The lexical categories appearing in this dataset are: nouns, verbs, adjectives, adverbs, pronouns, determiners, conjunctions, adpositions, interjections, dates, numerals and punctuation.

4.2 Durrett and DeNero (DDN)

Durrett and DeNero, (2013) evaluated their approach to Morphological Paradigm prediction with full morphology tables extracted from Wiktionary.[14] For Spanish they extracted morphology tables of verbs (231,135 total items). They used 208,335 tokens to train, 11,400 tokens for development, and 11,400 tokens for test.

[10] `http://corpus.rae.es/lfrecuencias.html`

[11] Note that there are some special cases in which two or more inflected forms can be inflected (e.g. the verbal forms *cantara* and *cantase* for a set of morphological features of the verb *cantar* (sang)).

[12] `http://ufal.mff.cuni.cz/conll2009-st/`

[13] `http://clic.ub.edu/corpus/ancora`

[14] `http://cs.utexas.edu/~gdurrett`

	size (#tokens)			eval dataset information					
dataset	train.	dev.	eval.	-pun %	infl. % -	unk.%	#noun	#verb	#adj
ConNLL2009	427,442	50,368	50,630	85.42	29.96	6.16	11,500	5,941	3,431
DDN (ES-V)	208,335	11,400	11,400	100	98.25	100	-	11,400	-
SIGMORPHON	12,575	1,596	23,229	100	90.31	100	2,914	18,739	1,576

Table 4: Statistics of the Spanish evaluation datasets. (-pun) = indicates % excluding punctuation, (infl.) = % of only forms that differ from the lemma, (unk) = % of forms unseen in the training set

4.3 SIGMORPHON 2016 Dataset

The SIGMORPHON 2016 Shared Task (task-1) data came mainly from the English edition of Wiktionary. The data extraction process is described in (Kirov et al., 2016). The Spanish dataset has 1,596 instances for development, 12,575 for training, and 23,229 for testing. The lexical categories present in the dataset are nouns, verbs and adjectives.

5 Results

The coverage and precision measures of the lexicons over the DDN and the SIGMORPHON 2016 datasets are shown in Table 5: Freeling and Leffe achieve high coverage and precision with coverages of more than 88% and up to 92% and precisions over 99.4%.[15]

Lexicon	DDN	SIGMORPHON
Apertium	59.50 (99.95)	58.81 (99.23)
Freeling	91.08 (99.99)	88.87 (99.42)
Leffe	92.41 (99.99)	90.35 (99.41)

Table 5: Lexicon evaluation results: % of coverage and precision (between parentheses)

The accuracy measures of the lexicons and the J48 Decision Tree algorithm over all evaluation datasets are shown in Tables 6, 7, and 8. The J48 algorithm achieves state-of-the-art results in all of the three datasets, outperforming some statistical algorithms but with slightly inferior results with respect to other approaches existing in the literature. The combination of morphological lexicons and the J48 algorithm outperformed the results of our other approaches in the three evaluation datasets tested, improved slightly the results with respect to the CoNLL2009 and SIGMORPHON results reported in the state-of-the-art literature, and achieved results comparable to the ones reported in the state-of-the-art literature on the DDN dataset evaluation. The results of the DDN dataset evaluation experiments (see Table 6) show that the combination of J48 and these two lexicons (Freeling and Leffe) improve the results of Durrett and DeNero, (2013) and Nicolai, Cherry, and Kondrak, (2015) and equals the ones reported by Ahlberg, Forsberg, and Hulden, (2015) but are still slightly inferior to the ones obtained by Faruqui et al. (2016).

Algorithm	Acc.(%)
Apertium	59.47
Freeling	91.07
Leffe	92.40
J48	99.57
Freeling+J48	99.89
Leffe+J48	99.85
Freeling+Leffe+J48	99.92
(Durrett and DeNero, 2013)	99.67
(Nicolai, Cherry, and Kondrak, 2015)	99.90
(Ahlberg, Forsberg, and Hulden, 2015)	99.92
(Faruqui et al., 2016)	**99.94**

Table 6: DDN evaluation results in accuracy

The results of the SIGMORPHON 2016 Shared Task task-1 dataset evaluation (see Table 7) show that the J48 algorithm achieves state-of-the art accuracy but slightly below the accuracies achieved by the best system at SIGMORPHON 2016 (Kann and Schütze, 2016). The combination of J48 with the lexicons outperforms the accuracies achieved by Kann and Schütze, (2016) and the other participants of SIGMORPHON 2016 (Cotterell et al., 2016).

Algorithm	Total	Noun	Verb	Adj.
Apertium	53.46	25.08	58.16	50.06
Freeling	88.36	76.38	91.72	70.49
Leffe (L)	89.82	81.91	91.59	83.37
J48	98.31	98.49	98.27	98.54
Freeling+J48	99.21	98.73	99.30	99.11
L+J48	99.19	98.73	99.26	**99.23**
Freeling+L+J48	**99.23**	**98.76**	**99.31**	**99.23**
Kann et al., 2016	98.94	-	-	-
Kann et al., 2016	99.05	-	-	-

Table 7: SIGMORPHON 2016 Spanish datasets evaluation accuracy (%) results

[15]The precision of the lexicons over the SIGMORPHON 2016 dataset could be increased if a manual revision detects annotation errors in this dataset.

Algorithm	Total	-pun	infl.	unk.	Noun	Verb	Adj	Adv	Pron	Det	Num
Apertium-spa	71.61	67.56	62.10	39.28	71.26	85.18	79.27	0	0.77	47.14	-
Freeling	84.04	81.77	86.34	43.94	72.60	95.03	76.21	72.89	43.29	87.69	-
Leffe	77.56	74.35	75.11	44.64	72.47	82.86	51.47	75.97	0	82.70	-
J48	99.05	98.92	**97.26**	92.81	99.80	94.32	99.24	**98.51**	**99.56**	99.95	**92.16**
Freeling+J48	**99.06**	**98.93**	97.11	**94.74**	**99.84**	**95.77**	**99.44**	96.19	98.00	**99.97**	**92.16**
Leffe+J48	98.13	97.87	94.02	94.70	**99.84**	95.70	99.41	96.19	**99.56**	92.91	**92.16**
Freeling+Leffe+J48	**99.06**	**98.93**	97.11	**94.74**	**99.84**	**95.77**	**99.44**	96.19	98.00	**99.97**	**92.16**
(Bohnet et al., 2010)	98.48	-	-	-	-	-	-	-	-	-	-
(Dušek and Jurcicek, 2013)	99.01	98.86	97.10	91.11	-	-	-	-	-	-	-

Table 8: CoNLL2009 Spanish results in accuracy (%)

Finally, the results of the CoNLL2009 Shared Task evaluation experiments (see Table 8) show that both J48 and J48 combined with the lexicons improved very slightly the results of the statistical learning approaches that evaluated the CoNLL2009 dataset for Spanish: the SVM approach (Bohnet et al., 2010) and the Logistic Regresion one (Dušek and Jurcicek, 2013).

6 Discussion

Morphological Generation is a crucial task in several advanced Language Technology applications that require so high precision that no errors should be passed to the output presented to final users; because a single wrongly inflected word could affect the end user's trustworthyness in the application. The system presented in this paper pretends to minimize inflection errors and improve the precision of Morphological Generation systems by incorporating the benefits of the lexicons to the ones obtained by supervised learning algorithms. The evaluation of the *coverage* and *precision* measures over two of these lexicons indicate that the lexicons can predict with a precision over 99.95% and 99.23% in the DDN and SIGMORPHON 2016 datasets respectively. In addition to the fact that the results of combining lexicons and Decisions Trees compare or outperform most of the state-of-the-art results, it has to be taken into account that the model generated by the J48 Decision Trees algorithm is human-interpretable and can be modified and extended in the same way as decision rules. On the other hand, common accuracy errors obtained in all three datasets were those produced by a wrong prediction of the edit scripts to inflect the form by the J48 model (e.g. given the lemma *endeudar* (indebt) generates *endieudas* instead of the correct form *endeudas*). But the most frequent errors were dataset-specific errors such as: i) verbal forms with enclitics without enough features in the dataset to be learnt by the model or recognized by the lexicon, phrasal verbs, and numerals (not present in the lexicons) for CoNLL2009 dataset, and ii) wrong morphological features present in the SIGMORPHON 2016 dataset that affect the lexicon-based predictions but not the J48 ones.

7 Conclusions and Further Work

This paper describes and evaluates free-available morphological lexicons and a Decision Tree algorithm that can perform Morphological Generation in Spanish at state-of-the art level. The Freeling, Leffe and Apertium Spanish lexicons, the J48 Decision Tree algorithm and the combination of J48 with Freeling and Leffe lexicons have been evaluated with the following datasets for Spanish: i) CoNLL2009 Shared task dataset, ii) Durrett and DeNero dataset of Spanish Verbs, iii) SIGMORPHON 2016 Shared Task (task-1) dataset. The results show that: i) the Freeling and Leffe lexicons achieve high coverage and precision over the DDN and SIGMORPHON 2016 datasets, ii) the J48 algorithm achieves state-of-the-art results in all of the three datasets, and iii) the combination of Freeling, Leffe and the J48 algorithm outperformed the results of our other approaches in the three evaluation datasets, improved slightly the results of the CoNLL2009 and SIGMORPHON 2016 reported in the state-of-the-art literature, and achieved results comparable to the ones reported in the state-of-the-art literature on the DDN dataset evaluation.

Further work includes: i) performing tuning of the J48 algorithm parameters using the development data, ii) the adaptation of the system to other Ibero-Romance languages such as: Catalan, Galician, and Portuguese, and iii) investigate data-driven methods to detect wrong predictions.

Acknowledgements

This work was partly funded by the ABLE-TO-INCLUDE project (European Commission CIP Grant No. 621055), the TUNER project (TIN2015-65308-C5-5-R, MINECO/FEDER, UE), and the Spanish MINECO Ministry (MDM-2015-0502).

References

Ahlberg, M., M. Forsberg, and M. Hulden. 2015. Paradigm Classification in Supervised Learning of Morphology. In *Proceedings of NAACL 2015*.

Amores, G., G. Pérez, and P. M. Portillo. 2006. Reusing MT Components in Natural Language Generation for Dialogue Systems. *Procesamiento del Lenguaje Natural*, (37):215–224.

Bahdanau, D., K. Cho, and Y. Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR*, abs/1409.0473.

Bohnet, B., L. Wanner, S. Mille, and A. Burga. 2010. Broad Coverage Multilingual Deep Sentence Generation with a Stochastic Multilevel Realizer. In *Proceedings of COLING 2010*, pages 98–106.

Bott, S., L. Rello, B. Drndarevic, and H. Saggion. 2012. Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish. In *Proc. of COLING 2012*, pages 357–374.

Cotterell, R., C. Kirov, J. Sylak-Glassman, D. Yarowsky, J. Eisner, and M. Hulden. 2016. The SIGMORPHON 2016 Shared Task—Morphological Reinflection. In *Proceedings of SIGMORPHON 2016*.

Durrett, G. and J. DeNero. 2013. Supervised Learning of Complete Morphological Paradigms. In *Proc. of NAACL 2013*.

Dušek, O. and F. Jurcicek. 2013. Robust Multilingual Statistical Morphological Generation Models. In *Proceedings of ACL 2013 Student Research Workshop*, pages 158–164.

Faruqui, M., Y. Tsvetkov, G. Neubig, and C. Dyer. 2016. Morphological Inflection Generation Using Character Sequence to Sequence Learning. In *Proceedings of NAACL 2016*.

Forcada, M. L., M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. M. Tyers. 2011. Apertium: a Free/Open-Source Platform for Rule-Based Machine Translation. *Machine Translation*, 25(2):127–144.

Hajič, J., M. Ciaramita, R. Johansson, D. Kawahara, M. A. Martí, L. Màrquez, A. Meyers, J. Nivre, S. Padó, J. Štěpánek, et al. 2009. The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages. In *Proceedings of the CoNLL-2009: Shared Task*, pages 1–18.

Kann, K. and H. Schütze. 2016. MED: The LMU System for the SIGMORPHON 2016 Shared Task on Morphological Reinflection. In *Proceedings of SIGMORPHON 2016*.

Kirov, C., J. Sylak-Glassman, R. Que, and D. Yarowsky. 2016. Very-large Scale Parsing and Normalization of Wiktionary Morphological Paradigms. In *Proceedings of LREC 2016*.

Marimon, M., N. Seghezzi, and N. Bel. 2007. An Open-Source Lexicon for Spanish. *Procesamiento del Lenguaje Natural*, 39.

Molinero, M. A., B. Sagot, and L. Nicolas. 2009. Building a Morphological and Syntactic Lexicon by Merging Various Linguistic Resources. In *Proceedings of NODALIDA 2009*.

Nicolai, G., C. Cherry, and G. Kondrak. 2015. Inflection Generation as Discriminative String Transduction. In *Proceedings of NAACL 2015*.

Padró, L. and E. Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of LREC 2012*.

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Rodríguez, S. and J. Carretero. 1996. A Formal Approach to Spanish Morphology: the COES Tools. *Procesamiento del Lenguaje Natural*, 19:119.

Taulé, M., M. A. Martí, and M. Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *LREC 2008*.

Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 117-124 recibido 01-12-2016 revisado 15-01-2017 aceptado 09-02-2017

Extracción de información temporal de la DBpedia: propuesta de integración en un corpus semiestructurado

Extraction of temporal information of the DBpedia: Integration proposal in a semi-structured corpus

Adolfo Merás, Ana García Serrano y **Ángel Castellanos**
ETSI Informática, UNED
C/Juan del Rosal 16
28040 Madrid
adolfo@meras.com.es, {agarcia, acastellanos}@lsi.uned.es

Resumen: En este trabajo, se hace una propuesta para la extracción automática de información temporal en la DBpedia, suficientemente general para ser aplicada a diferentes dominios. Se experimenta en un dominio concreto, para el que se identificarán y gestionarán recursos DBpedia relacionados. Con la información temporal extraída de los recursos, se alimentará una línea de tiempo y se intersecará a su vez con la información temporal extraída del dominio, en este caso del corpus DIMH (textos semiestructurados o fichas). A continuación, se enriquecerán las fichas originales con la información temporal y se visualizarán y accederá a los resultados organizados sobre la base de su dimensión léxica y temporal. Ante la ausencia de un *gold standard* para evaluar intrínsecamente la propuesta, se aplican criterios dependientes del dominio y de los usuarios y se pone a disposición de la comunidad científica (*GitHub*) el corpus anotado temporalmente.
Palabras clave: DBpedia, extracción - recuperación y acceso a la información, datos abiertos, análisis de conceptos formales.

Abstract: The goal of this work is to make a proposal for the automatic extraction of temporal information in the DBpedia, general enough to be applied to different domains. The experiment is performed using a concrete domain by the identification and management of domain related DBpedia resources. With the relevant temporal information extracted from the resources it will be feed a timeline and intersected with the temporal information of the DIMH corpus (semi-structured texts or cards). Thus, we will enrich these cards with related events of the timeline. In order to visualize the results, we are using a graphical interface to facilitate the lexical and the temporal information access. In the absence of a gold standard to intrinsically evaluate the proposal, it will be applied domain and users dependent criteria and the annotated corpus is made available to the scientific community (GitHub).
Keywords: DBpedia, extraction – retrieval and access to information, open data, formal concept analysis.

1 Introducción

Comprender la información de un dominio concreto exige la construcción de un contexto, en el que la dimensión temporal debe hacerse explícita (Tran et al., 2015), tanto sobre entidades (personajes, obras etc.) como sobre conceptos históricos. En un sistema de acceso a la información, aunque la información temporal "perfecta" la aportaría un experto humanista, la contextualización temporal automática de un corpus puede aprovechar recursos de la Web de los Datos, como es la DBpedia (Zang et al., 2015). En este trabajo se propone contextualizar y enriquecer la información de un corpus de textos con información temporal extraída automáticamente de la DBpedia, mediante:

1. La identificación de los recursos de la DBpedia relevantes al dominio, utilizando las etiquetas *rdf:type* (clase), o *dcterms:subject* (categoría);
2. la extracción de información temporal, teniendo en cuenta la consistencia entre recursos "hermanos" (*owl:sameAs*) en distintas DBpedia;

ISSN 1135-5948

3. la integración de la información anterior en el corpus del dominio.

La identificación de fechas en un texto es una tarea relativamente compleja ante la ausencia de uniformidad y normalización para la incorporación de información temporal, en cualquier lengua, como el español (Vicente-Díez et al., 2008), (Vicente-Díez et al., 2010), (Vázquez Méndez y García Serrano, 2015) o el catalán (Llorens et al., 2009).

En este trabajo se identifican fechas tanto en español como en inglés. Se describe brevemente la estructura y la forma en que aparece la información temporal en DBpedia en la sección 1.1. El criterio sobre el que se identifican y extraen los recursos de la DBpedia relacionados con un dominio se basa en dos parámetros, el *tope* y la *profundidad*, y permite evitar la pérdida de precisión, al desechar *categorías* y *clases* no relevantes.

El corpus de dominio utilizado en este trabajo es el formado por las fichas semiestructuradas del corpus DIMH, que se describe con más detalle en la sección 1.2. La extracción de información temporal del corpus conlleva el tratamiento del multilingüismo (en oraciones cortas), el análisis de colisiones entre las fechas extraídas, y la representación de intervalos temporales (como se detalla en la sección 3). La notación inicial se realiza con TimeML (Pustejovsky et al., 2003), (Vázquez Méndez y García Serrano, 2015), referencia en la que se encuentra una descripción de otras herramientas. TimeML también se utiliza en el nuevo GATE-Time (Derczynski, 2016).

La información de cada ficha del corpus ha sido enriquecida con nuevas anotaciones de eventos relacionados con una línea de tiempo (alimentada automáticamente desde la DBpedia, según se describe en la sección 2), y de la información temporal extraída del corpus. Para ello se ha diseñado un formato de anotación temporal, *.moment* (descrito en la sección 3.4). El corpus enriquecido con la información temporal está disponible en un repositorio público (que se describe en la sección 4.1).

A pesar de que el modelo desarrollado no es dependiente del dominio, obviamente sí se aplica a dominios concretos, por lo que a la hora de realizar una interfaz que permita la experimentación de la propuesta y la evaluación futura de los expertos humanistas, se ha organizado la información léxica y temporal extraída siguiendo una aproximación no supervisada (García Serrano y Castellanos, 2016). Así, se hace posible la visualización del corpus anotado DIMH, tanto por su dimensión léxica como temporal (sección 4).

1.1 Organización de la DBpedia

En lo que sigue se describe brevemente la estructura de la información contenida en la DBpedia que es de interés en este trabajo.

La DBpedia (Lehmann et al., 2015) es un repositorio de datos etiquetados según varias ontologías. Su fuente original y principal de recursos es la Wikipedia, utilizándose principalmente métodos automáticos para la estructuración de los datos. Existen varios proyectos de DBpedia activos hoy día, cada uno conteniendo la información en un idioma específico, tal y como sucede con la Wikipedia. Por ejemplo en español es.dbpedia.org (Mihindukulasooriya et al., 2015).

Cada tema en la DBpedia es un "recurso". Como es de esperar, se tratan temas similares en cada DBpedia de idiomas distintos e incluso en otros recursos dentro de la misma DBpedia, creando lo que se denominan recursos "hermanos". Dado un recurso, podemos identificar a sus hermanos con la etiqueta *owl:sameAs*; así, el recurso (a), tiene como hermanos los recursos (b) y (c), siendo:
(a) es.dbpedia.org/page/Alzamiento_de_Varsovia
(b) it.dbpedia.org/resource/Rivolta_di_Varsavia/html
(c) de.dbpedia.org/page/Warschauer_Aufstand

Cada recurso contiene múltiples etiquetas que aportan la semántica a la información que engloban, por ejemplo: *rdf:type* (clase), o *dcterms:subject* (categoría). Además hay etiquetas que contienen fechas (con diferentes formatos), como: es.dbpedia.org/property/fecha o es.dbpedia.org/property/date.

Aunque no se puede garantizar que el valor contenido siempre cumpla con el formato ISO 8601[1], sí ocurre cuando se presenta el tipo *xsd:date*[2]. Sin embargo es difícil establecer cómo, porqué o para qué se ha definido una fecha en la DBpedia, ante la ausencia de unos criterios conocidos por todos al respecto.

[1] http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=40874

[2] https://www.w3.org/TR/xmlschema11-2/#date

1.2 Corpus DIMH

El corpus DIMH (https://dimh.hypotheses.org/) consta de 7792 fichas que documentan con metadatos de material cartográfico imágenes de planos, mapas y dibujos del Archivo General de Simancas[3]. Cada ficha posee un identificador único y varias etiquetas con las que el anotador especificó los campos relativos a cada tipo de información incluida (por ejemplo <Titulo> o <Notas>, que incluyen texto libre). Estas fichas fueron enriquecidas con anotaciones relacionadas con entidades nombradas, sintagmas nominales y lemas (García-Serrano y Castellanos, 2016).

En este trabajo, se puede considerar a este corpus como de carácter general en tanto permite la experimentación en una tarea de extracción de información temporal.

2 Creación de una línea de tiempo

Para diseñar e implementar un método eficiente para alimentar una línea de tiempo de manera automática relacionada con un dominio en concreto, se decide utilizar la DBpedia (entre ellas la del español[4]). El método diseñado se divide en dos tareas: la de identificación y extracción de recursos acordes al tema deseado, y la de extracción de la información temporal contenida en estos recursos, para poder incluir eventos en la línea de tiempo, que tengan fecha de inicio y de fin.

2.1 Extracción automática de recursos de la DBpedia relevantes a un dominio

Se plantearon dos aproximaciones para abordar el problema, seleccionándose la segunda.

2.1.1 Modelo supervisado por un experto

En la DBpedia los recursos tienen asociados las etiquetas *clase*[5] (componente de la ontología que pertenece a la jerarquía, actualmente un grafo acíclico dirigido) y *categoría*[6] (agrupación de páginas que comparten un tema en común), siendo el conjunto de las *categorías* más amplio que el de las *clases*. Por ejemplo *Batalla de Lepanto* pertenece a la clase *Societal Event* y a más de una decena de categorías (*Batallas de España del siglo XVI, Batallas de la Armada de España, Guerras turco-venecianas, Reinado de Felipe II*…). Se seleccionó un conjunto de *clases* que acotan un conjunto de recursos relacionados con un tema del dominio, para utilizarlas como predicado en una consulta que ha de obtener los recursos necesarios para la identificación posterior de los eventos.

El problema es que el conjunto de clases que se obtienen automáticamente al inicio es muy pequeño; por tanto, en una primera fase se plantea expandir los recursos iniciales relacionándolos con nuevos recursos utilizando las *categorías* asociadas; luego se pondría a disposición de un experto humanista el conjunto de clases obtenido con esta expansión, y este elegiría el subconjunto con el cual se diseña el predicado para la búsqueda de recursos en la DBpedia.

A modo de ejemplo, partiendo del tema *Batalla de Lepanto*[7] vemos que posee la clase *MilitaryConflict*. De esta semilla se obtienen los recursos relacionados, con una consulta SPARQL como:

*SELECT * WHERE { ?s rdf:type dbpedia-owl:MilitaryConflict}*

Utilizando la propiedad *subject* de los recursos, se extraerán las categorías asociadas y con estas, más recursos con consultas SPARQL como:

*SELECT *WHERE { ?s dcterms:subject http://es.dbpedia.org/resource/Categoría:Batallas_de_España_del_siglo_XIII }*

Se entrará en un ciclo limitado por una cantidad de iteraciones (*profundidad*) cuya finalidad será obtener automáticamente un conjunto extenso de recursos del cual extraer las *clases*. Estas clases se presentan a un experto humanista para que escoja las adecuadas, por ejemplo *MilitaryConflict* y *MilitaryPerson*.

En una segunda fase se obtienen los recursos con la lista final de *clases* incluida en un predicado, con una consulta SPARQL como:

*SELECT * WHERE { {?s rdf:type dbpedia-owl:MilitaryConflict} UNION {?s rdf:type dbpedia-owl:MilitaryPerson}}*

Se han tenido en cuenta dos parámetros para las consultas a la DBpedia, *tope* (el límite a la lista de resultados que ha de devolver cada consulta de recursos de la DBpedia) y

[3] http://www.mcu.es/ccbae/es/mapas/principal.cmd
[4] http://es.dbpedia.org
[5] www.w3.org/1999/02/22-rdf-syntax-ns#type
[6] http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=terms#subject
[7] http://es.dbpedia.org/page/Batalla_de_Lepanto

profundidad (la cantidad de iteraciones para expandir las categorías-recursos), sabiendo que incrementar su valor provoca que disminuya la calidad del resultado de la expansión de las *categorías*.

Para no perder precisión, se limitaron las *categorías* obtenidas en la segunda iteración y siguientes a aquellas que mostrasen una semejanza mínima (*similitud*) con las categorías obtenidas en la primera iteración. La medida utilizada expresa la relación entre el número de elementos del conjunto de recursos comunes contra el total definido por ambas:

$$S_{AB} = \frac{2 * \#recursos\ comunes}{\#recursos\ categoría\ A + \#recursos\ categoría\ B}$$

Por otra parte, como se observa que para nombrar las categorías en la DBpedia, se intenta utilizar un vocabulario reducido y expresiones sencillas, como por ejemplo Batallas_de_ Suecia_ del_siglo_XVIII, en vez de Acontecimientos_ bélicos_de_Suecia_del_siglo_XVIII, es posible filtrar automáticamente los términos que deben contener las *categorías* para ser incluidas. Para ello se utiliza una lista de términos aportada por los expertos humanistas denominada de conocimiento previo, porque representa la terminología de unos intereses o tarea concreta en un momento dado.

2.1.2 Modelo automático

Esta aproximación se basa en la anterior pero no utiliza la expansión categoría-recurso sólo como un método de obtención de nuevos recursos, sino también para obtener los recursos finales de los que extraer los eventos. Básicamente, se utilizan los recursos "comunes" entre las *categorías* que participan en el proceso de medición de similitudes a partir de la primera iteración, como salida final. Adicionalmente se aplicará el filtrado de términos del conocimiento previo utilizado en la aproximación anterior por los buenos resultados obtenidos.

2.1.3 Comparación de resultados

Debido a la no disponibilidad de un *gold standard*, la cobertura de los resultados plantea las dificultades que se detallan a continuación. Con la aproximación supervisada por expertos, los recursos para el resultado final se obtienen con una consulta, que utiliza el predicado escogido manualmente, obteniendo una cantidad de recursos imposible de evaluar manualmente. Entonces se establece un límite al número de resultados de la respuesta, pero aun así no está claro cómo establecer los parámetros *tope* y *profundidad* adecuados en la aproximación automática, para que la comparación entre ambas fuese efectiva.

Se decidió utilizar sólo la precisión como medida, porque evita la sensación de fracaso que provocaría a un usuario experto (humanista en el caso del corpus DIMH) el encontrar estos fallos en los resultados automáticos.

Se catalogaron como *falsos positivos* aquellos recursos obtenidos no acordes con el tema y *verdaderos positivos* a los acordes. Es interesante indicar que en el criterio para revisar manualmente estos últimos (del tema Guerras y Batallas), no sólo se incluyeron acontecimientos bélicos, sino personajes, cosas y lugares famosos exclusivamente por su participación en este tipo de eventos.

La precisión de la aproximación supervisada por expertos es de 0,9994 y la de la automática de 0,9476, luego ambas aproximaciones tienen una alta precisión. Se observó en las pruebas que la mayor causa de errores en la segunda aproximación está relacionada con personajes históricos no relacionados con el tema.

La primera aproximación es útil si la preselección de clases por un experto es necesaria o sencilla de llevar a cabo (que no lo es para un experto). Por lo tanto, en el experimento que se presenta, se ha decidido utilizar la aproximación automática ante una precisión tan cercana y así potenciar un proceso no supervisado por expertos.

2.2 Extracción de información temporal de recursos DBpedia

Para la detección de fechas que parecen dentro de los recursos DBpedia seleccionados, por ejemplo los identificados y descargados en la fase anterior, se han estudiado y comparado tres estrategias, seleccionándose la tercera.

2.2.1 Estrategia 1

Consiste en utilizar una herramienta de detección de información temporal en texto plano con HeidelTime (HT) (Strötgen y Gertz, 2010). Se desecha porque:

- Extraer fechas en un texto del que ya se ha realizado un etiquetado semántico, sería desaprovechar un trabajo previo para realizarlo de nuevo y con menor precisión.

- Detectar fechas no relevantes para el tema (actualizaciones del equipo de trabajo, etc.), genera un nuevo problema y por lo tanto para resolverlo habría que filtrar ese ruido.

2.2.2 Estrategia 2

Esta estrategia consiste en extraer primero información de ciertas etiquetas de la DBpedia de las que se conoce a priori que denotan fechas relacionadas con el tema. Las dos etiquetas que en la DBpedia en español que se utilizan al efecto son: http://es.dbpedia.org/property/fecha y http://es.dbpedia.org/property/date.

De un total de 23203 recursos que contienen estas etiquetas sólo 1036, un 4%, contienen un valor de tipo *xsd:date*. Una observación manual de un número grande (más de 150) de las propiedades "fecha", muestra que datan aspectos muy variados y sería necesaria la supervisión manual de cada propiedad para establecer las relevantes. Por ello se desecha esta estrategia, pues la hipótesis es potenciar una solución lo más desatendida posible.

2.2.3 Estrategia 3

Esta estrategia consiste en, primero obtener las propiedades de tipo *xsd:date*, que contienen las fechas en los recursos, y a continuación seleccionar las relevantes. Analizando los recursos que contienen esta propiedad se observa que las fechas no relevantes, denominadas "ruido", suelen referirse a meta-información sobre la creación del recurso, actualizaciones, etc. Sobre todo, se constata que las fechas relevantes suelen repetirse entre los recursos hermanos, mientras que las "ruido" no. Por lo tanto, como criterio, se utiliza la presencia de una fecha en varios recursos "hermanos".

Todos los recursos contienen una etiqueta *owl:sameAs* que define otros recursos (mayormente de DBPedia en otros idiomas) denominados recursos "hermanos", cuyo contenido es similar al que hay en español. De esta forma la estrategia seleccionada consiste en:

- Identificar todas las fechas en el recurso y en los recursos "hermanos", fechas definidas por la etiqueta *xsd:date*.
- Realizar un filtrado y dejar sólo las fechas que aparezcan en al menos 2 recursos.
- Escoger una fecha de inicio y de fin del intervalo del evento. De entre las fechas preseleccionadas se escoge la menor y la mayor, respectivamente.

3 Extracción de la información temporal en el corpus DIMH

Las fichas contienen información etiquetada aunque hay algunas zonas etiquetadas que no contienen información temporal relevante a la época histórica del objeto descrito, o de las entidades que se referencian (como las referencias bibliográficas) así que se desecharon.

3.1 Anotación inicial con HeidelTime

De acuerdo con la discusión sobre estándares y herramientas de anotación detallada en (Vázquez y García-Serrano, 2015), se ha escogido la herramienta HeidelTime[8] (HT), para la notación temporal inicial de las fichas del corpus DIMH. HT es un anotador temporal multilingüe para expresiones temporales, usando el estándar TIMEML (Pustejovsky, 2003) y crea marcas <TIMEX3> sobre el texto original que denotan expresiones temporales explícitas. Es configurable con opciones para especificar el dominio del texto y el idioma de procesado. Para este trabajo se creó un script que invoca a HT para cada una de las fichas y que crea un nuevo fichero con la información temporal anotada.

3.2 Representación en intervalos

La propuesta de representación está basada en la lógica temporal de Allen (1983), que define operadores para expresar las relaciones temporales entre los intervalos identificados en el texto (ficha). Por ejemplo X *During* Y, para indicar que el evento X sucede mientras el evento Y está sucediendo. La información temporal se representa en intervalos, porque expresa explícitamente su tiempo de creación/ interés/ pervivencia; la alternativa sería utilizar un punto, pero no sería posible expresar un mes de cierto año (1870-05) mientras que con un intervalo sí, el que tiene de inicio a (1870-05-01) y como fin a (1870-05-31).

Otro argumento a favor del intervalo es que facilita la resolución de relaciones temporales entre fechas y fichas. Si la ficha A tiene un evento con fecha 1850-05-25 y la ficha B con

[8] http://dbs.ifi.uni-heidelberg.de/index.php?id=129

(a) (b)

Figura 1: Interfaz de búsqueda sin resolución de colisiones (a) y con resolución (b)

1850-05 y se quisiese utilizar el sistema de puntos sustituyendo el día de mes faltante con el primer día del mes entonces A tendría 1850-05-25, B 1850-05-01 y no se podría establecer que los eventos se han solapado. Utilizando el sistema de intervalos, A (1850-05-25__1850-05-25) y B (1850-05-01__1850-05-31) sí se muestra el solapado, ambos eventos transcurrieron en mayo de 1850.

Los operadores utilizados para expresar las relaciones entre intervalos que permiten el análisis de colisiones son X *Before/ After/ Same/ MeetsBefore/ MeetsAfter/ Overlaps/ During* Y.

3.3 La extracción con resolución de colisiones

La información temporal en las fichas del corpus DIMH está presente en varios idiomas, predominando el español. HT permite procesar cada ficha en cada uno de los idiomas presentes en la ficha y dar por válida la información marcada si el idioma de procesamiento con HT coincide con el del texto marcado. Para implementar esta estrategia fue necesario el desarrollo de un módulo de detección de idiomas basado en modelos de Markov (Padró y Padró, 2014). Este módulo es independiente de dominio.

Una vez desarrollado el módulo anterior y realizada la anotación temporal del corpus DIMH, se observó que en una ficha se pueden encontrar varios intervalos que se contengan entre sí, además de períodos extremadamente extensos que se deben considerar en este dominio como "ruido". Como el objetivo de la extracción de esta información era anotar y agrupar las fichas en momentos históricos, y construir una interfaz de búsqueda en la dimensión léxica y temporal, se diseña e implementa una estrategia que, ante una colisión, selecciona el intervalo más específico. Por ejemplo, teniendo tres intervalos:

1733-01-01_1733-12-31 (año 1733)
1733-04-09_1733-04-09 (9 de abril de 1733)
1503-01-01_1805-12-31 (ruido)

y aplicando la resolución de colisiones sólo se extraería un intervalo, *1733-04-09_1733-04-09*.

El efecto en la interfaz gráfica es notable. Por ejemplo, cuando no se aplica la resolución de colisiones, hay fichas con grandes intervalos temporales sin relación con la obra referida, que crean una especie de "ventana temporal" muy amplia y que provoca una organización con relaciones entre conceptos con atributos temporales muy alejados en el tiempo. En el ejemplo de la figura 1 (a) se visualiza el concepto SigloXVIII-SigloXVII, que está relacionado con conceptos que contienen a su vez, los atributos SigloXVI, SigloXIX, SigloXX y SigloXXI.

Con la resolución de colisiones, los intervalos temporales suelen ser significativos (comprobación manual) y relacionados con la obra descrita en las fichas (plano, mapa o dibujo) (ver figura 1 (b)).

3.4 Integración

La integración de la información temporal del corpus DIMH y de los eventos extraídos de la DBpedia, se ha organizado alrededor de un conjunto de términos relevantes para los

expertos humanistas, denominado de conocimiento previo (Merás, 2016).

La información temporal del corpus DIMH obtenida, la línea de tiempo alimentada con eventos de la DBpedia y su información temporal asociada, se ha integrado utilizando un formato de anotación (ejemplo al final de esta sección), almacenado en el fichero *.moment* con cuatro secciones claramente definidas, una para la información temporal extraída <*time_temporalInformation*>, otra para los eventos relacionados de la línea de tiempo <*time_relatedEvents*>, otra para los términos del conocimiento previo encontrados <*existing_words*> y el resto para la información original. En un futuro, podrían estudiarse estándares del Linked Open Data (http://linkeddata.org/) y su posible aplicación.

```
<Ficha id="183679">
  <Tipo>Ilustraciones y Fotos</Tipo>
  <Titulo> [ Muestras de tripe común
labrado negro y felpa azul turquesa]
…
  <Publicacion>    <TIMEX3    tid="t1"
type="DATE" value="1776">1776</TIMEX3>
…
<time_temporalInformation>
<temporalInformationItem  start="1776-
01-01" end="1776-12-31" />
</time_temporalInformation>
<time_relatedEvents>
  <event id="0000000018" title= "Siglo
XVIII">
  <class> century</class>
  <moment  start="1701-01-01"  end  =
"1800-12-31" /> </event>
  <event    id="0000001059"    title=
"Reinado de Carlos III de España">
  <class> monarchs_europe </class>
  <class> monarchs_spain </class>
  <moment   start="1759-08-10"   end=
"1788-12-14" /> </event>
  </time_relatedEvents>
  <existing_words>
  <word id="0000000021" name="azul" />
  </existing_words>
</Ficha>
```

4 Interfaz para el análisis exploratorio

Para presentar a los expertos humanistas los resultados obtenidos y para realizar una búsqueda tanto por la dimensión temporal como por la dimensión léxica, se desarrolló una interfaz para la información organizada sobre la base de los términos del conocimiento previo, el contenido original de las fichas del corpus DIMH y la información temporal recabada, utilizando el modelo de visualización propuesto en (Filter, 2015).

Por ejemplo, si un experto comienza a buscar el término *fuerte*, cuando le aparecen los resultados podría filtrar adicionalmente si le interesa los del *siglo XVII* o los construidos con *Felipe II*; de manera contraria podría comenzar buscando a *Felipe II* como período histórico y luego decantarse por el término *fuertes*.

4.1 Recursos disponibles

Se puede acceder a la interfaz a través de la dirección: http://albali.lsi.uned.es:8111/ (ver figura 1 y figura 2). El repositorio https://github.com/meras0704/DBpediaTime contiene a:

(a) *Fichas_moment.zip*, que contiene 7792 ficheros en formato *.moment* con el resultado de la anotación y enriquecimiento de las fichas DIMH.

(b) *Solution_DBpediaTime*, solución implementada en Visual Studio y lenguaje C# con los programas necesarios. A su vez contiene 4 proyectos: *Blatella.Common*, la funcionalidad de uso común, *Blatella.ML.Common*, *TFM.Common*, la funcionalidad utilizada para la identificación de la información temporal, y *TFM.IU.WindowsForm*, la interfaz gráfica (primitiva pero suficiente).

5 Conclusiones y trabajos futuros

En este trabajo se ha propuesto un modelo experimental para la extracción de información temporal de la DBpedia y para enriquecer y contextualizar temporalmente la información de un corpus de textos de un dominio concreto. Por una parte se ha definido un modelo de extracción, basado en intervalos, que permite resolver las colisiones entre información temporal identificada en un corpus. Además se ha diseñado un método para la extracción automática de eventos de la DBpedia, integrándolos en una línea de tiempo, y se ha definido para ello el formato *.moment*.

Se ha experimentado la propuesta en el corpus DIMH y se ha desarrollado una interfaz de acceso a la información léxica y temporal.

Como trabajos futuros se desea confirmar la independencia del dominio del modelo propuesto para anotación temporal desde la DBpedia, plantear un modelo de evaluación no dependiente de los usuarios expertos y detectar patrones de evolución (Neouchi et al., 2001) en los atributos de una clase a lo largo del tiempo.

6 Agradecimientos

Este trabajo ha sido financiado parcialmente por los proyectos DIMH (HAR2012-31117) y Musacces (S2015/HUM3494).

Bibliografía

Allen, J. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11): 832-843.

Derczynski, L., J. Strötgen, D. Maynard, M. A. Greenwood and M. Jung. 2016. GATE-Time: Extraction of Temporal Expressions and Events. In *10th LREC*.

Filter J. 2015. Interactive Visualization of Large Concept Lattice. *Facultad de Ciencias de la Computación*, U. Magdeburgo. Alemania

Ganter B. 2002. Formal Concept Analysis: Methods and Applications. Computer Science. TU Dresden.

García-Serrano, A. y A. Castellanos. 2016. Conceptualización, acceso y visibilidad de la información en el proyecto DIMH. Cap. 16 *El dibujante ingeniero al servicio de la monarquía hispánica (XVI-XVIII)*, páginas 379-400. ISBN: 978-84-942695-6-1.

Lehmann, J., R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer and C. Bizer. 2015. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web J*ournal, 6(2): 167-195.

Llorens H., B. Navarro, E. Saquete. 2009. Detección de expresiones temporales TimeML en Catalán mediante roles semánticos y redes semánticas. *Procesamiento del Lenguaje Natural* (43): 13-21.

Merás A. 2016. Propuesta para extracción, representación y organización de información temporal en textos semiestructurados: aplicación al corpus DIMH. *Tesis del máster "Lenguajes y Sistemas Informáticos" de la UNED*.

Mihindukulasooriya N., M. Rico, R. García-Castro, A. Gómez-Pérez. 2015. An Analysis of the Quality Issues of the Properties Available. *Spanish Dbpedia, LNCS 9422*, páginas 198-209.

Neouchi R., A. Tawfik and R. Frost. 2001. Towards a Temporal Extension of Formal Concept Analysis. *Proceedings of the 14th Canadian Conference on AI*, Ottawa, Ontario.

Padró M., Ll. Padró. 2014. Comparing methods for language identification. *Procesamiento del Lenguaje Natural* (33): 155-161.

Pustejovsky J., J. Castaño, R. Ingria, R. Sauri, R. Gaizauskas, A. Setzer and G. Katz. 2003. TimeML: Robust Specification of Event and Temporal Expressions in *Proceedings of the IWCS International Workshop on Computational Semantics*.

Strötgen, J. and M. Gertz. 2010. HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions. En *Proceedings of the 5th International Workshop on Semantic Evaluation*, páginas 321-324, Uppsala, Sweden, July. ACL.

Tran, N., A. Ceroni, N. Kanhabua, and C. Niederée. 2015. Back to the past: Supporting interpretations of forgotten stories by time-aware re-contextualization. In *Proc. of the ACM International Conference on Web Search and Data Mining*, páginas 339-348.

Vázquez-Méndez, A. y A. García-Serrano. 2015. Anotación y representación temporal de tweets multilingües. *Procesamiento del Lenguaje Natural* (54): 53-60.

Vicente-Díez, M.T., D. Samy and P. Martínez. 2008. An empirical approach to a preliminary successful identification and resolution of temporal expressions in Spanish news corpora. *Proc. of the Sixth Int. Language Resources and Evaluation Conf. (LREC'08)*, Marrakech, Morocco, May, 2008, European Language Resources Association (ELRA), ISBN: 2-9517408-4-0, páginas 2153-2158.

Vicente-Díez M.T., J. Moreno-Schneider, P. Martínez. 2010. Temporal information needs in ResPubliQA: an attempt to improve accuracy. The UC3M Participation at CLEF 2010, *CLEF 2010 LABs and Workshops, Notebook Papers*, Padova, Italy, September.

Zhang, L., W. Chen, T. Tran and A. Rettinger. 2015. Time-Aware Entity Search in DBpedia. In *European Semantic Web Conference, p*áginas175-179.

Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 125-132 recibido 01-12-2016 revisado 12-01-2017 aceptado 09-02-2017

Building the Gold Standard for the Surface Syntax of Basque

Construcción de un Gold Standard para la Sintaxis Superficial del Euskera

Itziar Aduriz[1], María Jesús Aranzabe[2], José María Arriola[2]*
Arantza Díaz de Ilarraza[2], Itziar Gonzalez-Dios[2], Ruben Urizar[2]
IXA Research Group
[1]University of Barcelona
[2]Universtity of the Basque Country UPV/EHU
* josemaria.arriola@ehu.eus

Resumen: En este artículo presentamos el proceso de construcción de SF-EPEC, un corpus de 300.000 palabras, sintácticamente anotado, que pretende ser un Gold Standard para el pocesamiento sintáctico superficial del euskera. En primer lugar, describimos el conjunto de etiquetas diseñado para este propósito; siendo el euskera una lengua aglutinante, en ocasiones hemos tenido que crear etiquetas sintácticas compuestas. Asimismo, se detallan las distintas fases en la construcción de SF-EPEC.
Palabras clave: Sintaxis superficial, *gold standard*, euskera, anotación de corpus

Abstract: In this paper, we present the process in the construction of SF-EPEC, a 300,000-word corpus syntactically annotated that aims to be a Gold Standard for the surface syntactic processing of Basque. First, the tagset designed for this purpose is described; being Basque an agglutinative language, sometimes complex syntactic tags were needed. We also account for the different phases in the construction of SF-EPEC.
Keywords: Surface syntax, gold standard, Basque, corpus annotation

1 Introduction

Corpora are essential resources in linguistics research. As stated by Sampson (2011), the use of corpora in language research allows a better understanding of language complexity particularly on syntactic issues.

The development of data-driven language processors requires large amounts of texts manually tagged at different levels, which are called gold standard corpora. These are also used to evaluate the output of rule-based processors comparing their results with the gold standard annotation.

Important efforts have been devoted to the construction of syntactically annotated gold standards in several languages such as English (Marcus, Marcinkiewicz, and Santorini, 1993; Silveira et al., 2014), Spanish (Mille et al., 2009), German (Scheible et al., 2011), Norwegian (Solberg et al., 2014), Swedish (Nilsson and Hall, 2005), or Finnish (Voutilainen, Purtonen, and Muhonen, 2012).

Similarly, our effort was led to annotate syntactically the Reference Corpus for the Processing of Basque EPEC (Aduriz et al., 2006a). This syntactically annotated corpus, hereafter SF-EPEC, aims to be a Gold Standard for the development and evaluation of shallow syntactic analyzers for Basque. Specifically, SF-EPEC has as an immediate goal the evaluation of SF-Grammar, a rule-based surface syntactic analyzer for Basque (Arriola, 2015).

Previously, Aduriz and Díaz de Ilarraza (2013) established the theoretical and practical issues for the shallow syntactic annotation in Basque. The annotation process of SF-EPEC was largely inspired in Voutilainen, Purtonen, and Muhonen (2012). The authors specify different steps for the process of corpora annotation, which include tasks such us (i) specifying a tentative annotation model and guidelines; (ii) applying the model to a large sample of example sentences and if necessary refining the model and the guidelines; or (iii) evaluating the applicability by means of the double-blind annotation routine.

Likewise, the methodology for the annotation of SF-EPEC comprised the following

ISSN 1135-5948

steps:

1. A random sample of full sentences –consisting of 3% of the corpus–was extracted for it to be manually annotated.
2. During the annotation of this sample, a discussion phase took place so as to decide how to annotate some specific phenomena. An annotation guideline was drawn up with the decisions taken.
3. Then, taking into account the redefined tagset and the annotation guidelines, three different coders annotated a sample corpus of about 11,500 ambiguous tokens in parallel, and the inter-annotator agreement was measured.
4. Finally, the whole corpus was annotated by two linguists.

After introducing our strategy for building the Gold Standard for surface syntax and related work, Section 2 explains the basic resources for this syntactic annotation. In Section 3, we describe the tagset designed to annotate syntactic functions. Section 4 is devoted to the manual annotation, i.e. the discussion phase and inter-annotator agreement. Finally, some conclusions are presented in Section 5.

2 *Framework for the annotation*

The IXA research group[1] is working on a robust parsing scheme that provides syntactic annotation in an incremental fashion (see Figure 1).

The information contained in the lexical database for Basque EDBL (Aldezabal et al., 2001) constitutes the basis for our analyzers. It consists of 121,823 entries divided into (i) dictionary entries, (ii) inflected verb forms, and (iii) dependent morphemes, all of them with their respective morphological information.

In the morphosyntactic analysis, first a tokenizer divides the text into a sequence of tokens. Then, the robust morphological analyzer MORFEUS (Alegria et al., 1996) gives to each word form every possible analysis, without taking into account the context in which it appears; that way each word form of the whole corpus is assigned its corresponding analysis at the segmentation level.

[1] http://ixa.eus

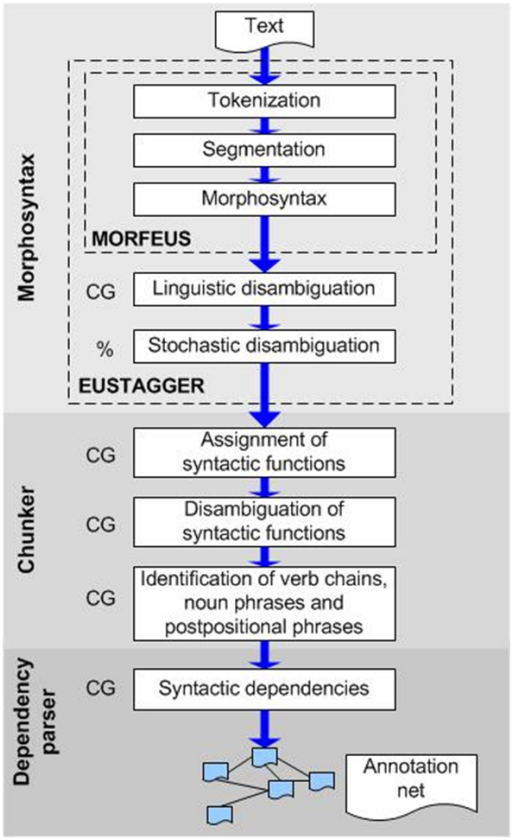

Figure 1: General framework

```
"<$.>" <PUNT_PUNT>"
"<Zalantzak>"
  "zalantza" IZE ARR DEK ABS NUMP MUGM
  "zalantza" IZE ARR DEK ERG NUMS MUGM
  "zalantza" IZE ARR DEK ERG MG
"<argitu>"
  "argitu" ADI SIN AMM PART ASP BURU
  "argitu" ADI SIN AMM PART
"<zituzten>"
  "*edun" ADL B1 NR_HK NK_HK ERL MEN ERLT
  "*edun" ADL B1 NR_HK NK_HK ERL MEN ZHG
  "*edun" ADL B1 NR_HK NK_HK
  "ukan" ADT B1 NR_HK NK_HK ERL MEN ERLT
  "ukan" ADT B1 NR_HK NK_HK ERL MEN ZHG
  "ukan" ADT B1 NR_HK NK_HK
```

Figure 2: Morphological analysis of the sentence *Zalantzak argitu zituzten* 'They clarified the doubts'

Figure 2 shows the analysis provided by MORFEUS for the sentence *Zalantzak argitu zituzten* 'They clarified the doubts' expressed in a Constraint Grammar (CG) style, in which every word form is associated with one or more reading lines. Each line corresponds to a possible interpretation, which provides the word form's lemma, part-of-speech, number, case markers, definiteness and other

morphological information.

Then, the lemmatizer-tagger EUSTAGGER (Aduriz et al., 2003) performs the automatic disambiguation at two levels: first, a rule-based disambiguation is carried out and then the stochastic disambiguation is applied (see Figure 3).

```
"<Zalantzak>"
  "zalantza" IZE ARR DEK ABS NUMP MUGM
"<argitu>"
  "argitu" ADI SIN AMM PART ASP BURU
"<zituzten>"
  "*edun" ADL B1 NR_HK NK_HK
```

Figure 3: Morphological disambiguation

After performing the morphological disambiguation, the next step is to assign the corresponding syntactic tag to each word form. Typically, inflectional suffixes and syntactic functions are closely related in Basque, and therefore most suffixes in the lexical database are assigned their corresponding syntactical function(s) (see Section 3.1). As a result, the output of the morphological analyzer displays these syntactic tags. The syntactic tags at this level refer to shallow syntactic functions. The symbol `@` precedes the abbreviation for the syntactic function. For example, the tags `@OBJ`, `@SUBJ` or `@PRED` stand for object, subject and predicative respectively (Figure 4).

```
"<Zalantzak>"
  "zalantza" IZE ARR DEK ABS NUMP MUGM @OBJ
  "zalantza" IZE ARR DEK ABS NUMP MUGM @SUBJ
  "zalantza" IZE ARR DEK ABS NUMP MUGM @PRED
"<argitu>"
  "argitu" ADI SIN AMM PART ASP BURU
"<zituzten>"
  "*edun" ADL [...] ERL MEN ERLT @+JADLAG_IZLG>
  "*edun" ADL [...] ERL MEN ZHG @+JADLAG_MP_OBJ
  "*edun" ADL B1 NR_HK NK_HK
```

Figure 4: Syntactic tags

However, some word forms lack any suffix or have a suffix with no specify syntactic function as in the past participle *argitu* 'clarified' in the sentence in Figure 4. Those word forms that are not given a syntactic tag by the morphological analyzer are assigned one to their analysis through CG mapping rules (Aduriz and Díaz de Ilarraza, 2013). Similarly, disambiguation is carried out in the case of word forms having more than one possible syntactic function e.g. *zalantzak* and *zituzten* in the sentence in Figure 4. This is also done through a CG grammar (Aduriz, 2000; Arriola, 2015). In the final output, each word form in the sentence keeps a single morphological analysis and a single syntactic tag as shown in Figure 5.

```
"<Zalantzak>"
  "zalantza" IZE ARR DEK ABS NUMP MUGM @OBJ
"<argitu>"
  "argitu" ADI SIN AMM PART ASP BURU @-JADNAG
"<zituzten>"
  "*edun" ADL B1 NR_HK NK_HK @+JADLAG
```

Figure 5: Syntactic disambiguation

Also, a chunk parser provides a partial constituent analysis (Aduriz et al., 2006b) and finally a dependency parser establishes the dependency links (Aranzabe and Díaz de Ilarraza, 2009).

SF-EPEC Gold Standard is aimed to be an essential resource for the evaluation and consequent improvement of the CG grammars that allocate syntactic tags to the word forms in a text.

3 Syntactic tagset

Following the CG formalism, the annotation of syntactic functions in CG is based on the word, understood as the content between two blanks. With this in mind, the main feature of the annotation is that all words need to be provided with a syntactic label (Karlsson et al., 1995). An obvious consequence of this requirement of the CG parser was that, apart from the traditional syntactic functions, specific labels needed to be created for words which in principle do not have 'traditional' syntactic information, such as elements of some multiword expressions.

Being Basque an agglutinative postpositional language, often the syntactic function of a word is given by the suffix attached to it such as a case marker (see Section 3.1a). In (1), the ergative case added to the stem *etxe* 'house' assigns the subject function to the word (*etxeek*).

(1) *etxe-ek*
house-the.PL.ERG
'the houses' (SBJ)

Moreover, subordinating morphemes can be added to finite or non-finite verb forms as well as to main or auxiliary verbs in such a way that each subsequent morpheme gives a piece of the syntactic information. Complex syntactic tags are used for this purpose (see Section 3.1c). For instance, the suffix *-takoan* ('once')

added to the past participle form of a verb allocates the word a complex tag indicating "non-finite verb, subordinate clause functioning as verb complement" as *bukatutakoan* in (2).

(2) *buka-tu-takoan*
finish-ed-once
'once finished' (NFIN SUBR ADV)

Besides, some independent function words—e.g. coordinators (3) or sentence connectors—hold a syntactic function which is inherent to the parts of speech they belong to (see Section 3.1b).

(3) edo
or
'or' (CONJ)

However, not all the lexical words in a sentence are inflected in Basque. For example, it is typically the last element in the noun or postpositional phrase that takes the case marker. In (4), the demonstrative in the final position of the PP takes the inessive case, but the rest of the lexical words (*igande* 'Sunday', *euritsu* 'rainy', *ilun* 'dark') are devoid of a case marker.

(4) *igande euritsu eta ilun hartan*
Sunday rainy and dark that.INE
'in that rainy and dark Sunday'

Therefore, the words lacking a case marker are added a function tag through mapping rules. Some of these syntactic tags are the same as the ones designed for the database, but others are new. In particular 23 specific labels needed to be created for words which in principle do not have 'traditional' syntactic information e.g. elements of some multiword expressions (see Section 3.2). For instance, for the multiword sentence connector *hala eta guztiz ere* 'despite everything' (see (5)), the words *hala, eta*, and *guztiz* are allocated the tag `@HAOS>` denoting they are just components of multiword expression, while the last element *ere* is assigned the function tag for sentence connector `@LOK`.

(5) *hala eta guzti-z ere*
like.that and all-INS too
'despite everything'

Furthermore, some additional tags had to be created during the manual annotation process for specific cases (see Section 4.2).

Bearing all this in mind, we have divided the syntactic tagset developed for the labeling of the Basque corpus in three groups, depending on the step in which they are applied:

- Syntactic tags derived from the lexical database (explained in Section 3.1).
- Tags allocated through mapping rules during the assignment of syntactic functions (explained in Section 3.2).
- Tags created during the manual annotation process for specific cases (Section 4.2).

3.1 Tags from the lexical database

The analyses produced by the morphosyntatic analyzer for Basque MORFEUS are accomplished based on the information included in the lexical database for Basque EDBL. Each entry in EDBL is kept along with its morphosyntactic information. 19 different syntactic tags are used in the lexical database. The following entries holding a syntactic tag:

a) Case markers. As said before, often the syntactic function of a word in a Basque sentence is given by a suffix attached to it such as a case marker. In Table 1 we present some examples of suffixes and their assigned syntactic function.

Function	Meaning	Suffix holding the function
`@SUBJ`	Subject	Ergative and absolutive
`@OBJ`	Direct object	Absolutive
`@ZOBJ`	Indirect object	Dative
`@ADLG`	Verb complement	Locative, directional, origin, comitative, instrumental, cause, goal...
`@PRED`	Predicative	Absolutive
`@IZLG>`	Left noun complement	Genitive locative and genitive

Table 1: Syntactic functions associated to case markers

(6) *Gutun-ek egi-a esan zuten.*
letters-ERG truth-ABS say.PFV AUX.3PL.PST
`@SUBJ` `@OBJ` `@-JADNAG` `@+JADLAG`
SBJ DO NF.VB FIN.AUX
'The letters told the truth.'

In (6), we show which the function tags would be for each word in the sentence *Gutunek egia esaten zuten* 'The letters told the truth'.

b) Some function words. Sentence connectors (*halere*, 'however'), independent

subordinators (*arren*[2], 'although'), and coordinators (*eta*, 'and') hold a syntactic function which is inherent to the part of speech they belong to (Table 2).

Function	Meaning
@LOK	Sentence connector
@PJ	Coordinator
@MP	Independent subordinator

Table 2: Syntactic tags corresponding to function words

We show an example of coordination in (7).

(7) *Peio* *eta* *Iñaki* *hemen* *dira.*
Peio.ABS and Iñaki.ABS here are
@SUBJ @PJ @SUBJ @ADLG @+JADNAG
SBJ CONJ SBJ ADV NF.VB
'Peio and Iñaki are here.'

c) Dependant subordinators (MP) are suffixes which can be added to finite auxiliary verbs (@+JADLAG), non-finite main verbs (@-JADNAG) or finite synthetic verbs (@+JADNAG). The syntactic function that the subordinator assigns to the subordinate clause (verb complement, noun complement, object...) is added to the previous verb-type tag, thus making up a complex tag with the combination of the three elements. For instance, Table 3 shows some complex tags for finite main verbs.

Function	Meaning
@+JADNAG_MP_ADLG	Finite main verb, subordinate clause functioning as a verb complement
@+JADNAG_MP_SUBJ	Finite main verb, subordinate clause functioning as a subject
@+JADNAG_MP_OBJ	Finite main verb, subordinate clause functioning as a direct object

Table 3: Dependant subordinator

In (8), the word *bukatutakoan* 'when finished' holds the complex tag @-JADNAG_MP_ADLG, which stands for "non-finite main verb, subordinate clause functioning as verb complement".

(8) *Buka-tu-takoan* *joan-go* *gara.*
finish-ed-when go.FUT 1PL.PRS
@-JADNAG_MP_ADLG @-JADNAG @+JADLAG
NF.SUBR.ADV NF.VB FIN.AUX
'When finished, we will leave.'

3.2 Tags added in assignment phase

The word forms that are assigned no syntactic tag by the morphological analyzer MORFEUS are allocated one through CG mapping rules (Section 2). Some of the syntactic tags added by this grammar are the same designed for the database (see Section 3.1), but others were created for this stage. In Table 4, we can find some examples of new syntactic tags added in the assignment phase.

Function	Meaning
@KM>	Modifier of the word containing the case marker
@<IA	Postmodifier
@IA>	Premodifier
@<ID	Right determiner
@ID>	Left determiner
<@GRAD	Right grader
@GRAD>	Left grader
@ADILOK>	First element of compound verb
<@ADILOK	Last element of compound verb
@HAOS>	Element of a multiword expression

Table 4: Examples of new syntactic tags added in the assignment phase

Unlike the tags derived from the lexical database, the tags added in the assignment phase need syntactic context to be assigned, and that is why they are attached on the outcome of MORFEUS.

Some words are allocated tags which have no conventional syntactic information. For example, as stated before, it is the last element in the noun or postpositional phrase that takes the case mark in Basque (see (4)). The tag @KM> is assigned to all the nouns lacking a case mark in the phrase, as in *igande* 'Sunday' in (9). Also, the tags @<IA or @IA> are added to noun postmodifiers and premodifiers respectively; for instance, the adjectives *euritsu* 'rainy' and *ilun* 'dark' in (9) get the tag @<IA for postmodifier.

(9) *igande* *euritsu* *eta* *ilun* *hartan*
Sunday.ø rainy.ø and dark.ø that.INE
@KM> @<IA @PJ @<IA @ADLG
HEAD POSTMOD CONJ POSTMOD ADV
'in that rainy Sunday'

Also, some components of compound verbs—such as *min* 'pain' in *min egin*, 'to

[2]Almost all subordinators in Basque are morphemes attached to either finite or non-finite verb forms (see Section 3.1b). Just a few, such as the adversative conjunction *arren* 'despite', are written separately.

hurt' lit. 'do-harm'—are assigned new tags (@ADILOK> or @<ADILOK) since the morphosyntactic information of the compound is usually given by one of the elements. Also, components of other multiword expressions (e.g. *ziur aski*, 'most probably') are added the tag @HAOS> indicating the following element carries the syntactic tag corresponding to the whole expression (see (10)).

(10) *Ziur* *aski* *min* *egin-go* *di-zu.*
sure very harm do-FUT 3SG.SBJ-2SG.IO
@HAOS> @ADLG @ADILOK> @-JADNAG @+JADLAG
> ADV > NF.VB FIN.AUX
'Most probably s/he will hurt you.'

4 Manual annotation

In order to build up the Gold Standard for syntactic functions SF-EPEC we used EPEC, the Reference Corpus for the Processing of Basque (Aduriz et al., 2006a). EPEC is a 300,000-word collection of texts written in standard Basque, which is intended to be a reference corpus for the development and improvement of several NLP tools for Basque. Although small, it is strategic for a less-resourced language like Basque.

EPEC was first morphologically analyzed by means of MORFEUS and then manually disambiguated (Aldezabal et al., 2007). The process of the annotation of the syntactic functions in SF-EPEC consisted in either selecting the correct syntactic function from the different ones provided by the morphological analyzer MORFEUS or adding the correct syntactic tag whenever MORFEUS provided no function or none of the ones provided was correct.

For example, the absolutive case may function either as a subject in intransitive sentences, as an object, or as a predicate. Thus, the word form *zalantzak* in Figure 4 is allocated three syntactic tags. However, in the specific context of *zalantzak* in (11), the correct syntactic function for the annotator to choose would be 'direct object' (absolutive plural).

(11) *Zalantz-ak* *argi-tu* *zituzten.*
doubt-ABS.PL clarify-PTCP 3PL.SUJ·3PL.DO.PST
@OBJ @-JADNAG @+JADLAG
DO NF.VB FIN.AUX
'They clarified the doubts.'

The manual annotation took place in three different stages: the discussion phase, the inter-annotator agreement phase and the annotation of the whole corpora.

4.1 Discussion phase

In order to define the tagset and the criteria for the annotation, a random sample of full sentences—comprising 3% of the corpus—was extracted for it to be manually annotated. The annotation was carried out by a linguist, and the doubts arising during the process were discussed by two more linguists with experience in NLP annotation tasks. Decisions were taken so as to decide how to annotate some specific phenomena. As a result, an annotation guideline was drawn up with the decisions taken in the discussion phase (Aduriz et al., 2015) .

Many of the decisions taken in this stage involved the use of tags previously defined. For instance, we found out that for some tokens the syntactic tags provided by the analyzer did not correspond to their real functions in some specific contexts. In order to solve this problem, the tags were manually added. Sometimes existing tags were added to the tokens, for example, in some multiword expressions such as complex postpositions (*-ren aurrean*, 'in front of') or complex subordinators (*-n arte*, 'until').

(12) *etxe-a-ren* *aurre-a-n*
house-the-GEN front-the-INE
'in front of the house'

In (12) (*etxearen aurrean*, 'in front of the house') the morphological analyzer allocates the noun complement function (@IZLG>) to the first word containing the genitive case (see Section 3.1a). However, in the example above the genitive is part of a complex postposition (*-ren aurrean*, 'in front of') so in the manual annotation the @KM> tag was added to the token containing the genitive case, indicating that it is the following token—*aurrean* 'in front' containing the inessive case marker—that allocates the syntactic function corresponding to the complex postposition.

Similarly, in complex subordinators such as *-n arte* 'until' (see (13)), the subordinator *-n* attached to the finite auxiliary verb is automatically assigned the subordinator function. However, in the manual annotation the @KM> tag was added to this token, indicating that it is the following word that holds the syntactic function corresponding to the whole complex subordinator.

(13) *etorr-i* *de-n* *arte*
come-PTCP has-SUBR until

'until s/he has come'

4.2 New tags

For cases that had not been foreseen in the initial annotation scheme, two new tags were created: `@IS` (isolated noun phrase) and `@FSG` (no syntactic function).

Most of the cases in which a new tag was needed corresponded to phrases belonging to verbless incomplete structures in which it was impossible to determine the syntactic function of the phrase. For isolated noun phrases in contexts such as titles, bibliographical references, mathematical formulae, vocatives, parenthetical structures, dates and places in brackets... the tag 'noun phrase' (`@IS`)[3] was created based on the tagset in the parser Palavras (Bick, 2000).

Also, some tokens such as the numbers in item lists or section headings do not hold a syntactic function. Therefore, the tag @FSG (no syntactic function) was created to annotate these tokens or similar ones that are devoid of a syntactic function.

4.3 Inter-annotator agreement

In order to evaluate the consistency of the annotation guidelines generated so far and the reliability of our corpus, three linguists —two of them with a long experience in several NLP tasks—annotated a part of the corpus. The inter-annotator agreement was measured using Fleiss' kappa (Fleiss, 1971) obtaining 0.945. The observed agreement was 93%. This result shows that our guidelines are clear enough and our tagging is consistent. Besides, they show the reliability of our corpus since agreement is very important for the production of representative text corpora with high-quality linguistic annotation.

Then, we examined the cases in which the annotators disagreed. In 48.55% of the cases, the disagreement was related to the use of the new tags. The disagreements related to the conflictive cases of complex postpositions and complementisers and multiword units were not very common, 7.24% and 2.90% respectively. Nevertheless, 69.56% of the cases where disagreement was found were covered by the guidelines. This suggests that annotators sometimes tended to follow their expertise and intuition rather than the guidelines.

[3] `@IS` stands for Basque *Izen Sintagma* 'Noun Phrase'.

Finally, the whole corpus was annotated bearing in mind all the decisions taken and the expertise gained in the previous stages.

5 Conclusion

Although time-consuming and costly, Gold Standard corpora are essential to develop data-driven language processors as well as to evaluate the output of rule-based processors.

In this paper, we have presented the process in the construction of SF-EPEC a syntactically annotated corpus of 300,000 words aimed to be a Gold Standard for the surface syntactic processing of Basque. Previous to the annotation, a linguistically motivated tagset was designed to account for the morphosyntactic complexity of the Basque language.

The inter-annotator agreement obtained (93%) shows that the tagset developed as well as the criteria established for the annotation are quite sound, and therefore the corpus obtained will be a reliable reference corpus.

Acknowledgments

PROSA-MED: Procesamiento semántico textual avanzado para la detección de diagnósticos, procedimientos, otros conceptos y sus relaciones en informes Médicos (TIN2016-77820-C3-1-R).

References

Aduriz, I. 2000. *EUSMG: Morfologiatik sintaxira Murriztapen Gramatika erabiliz.* Ph.D. thesis, University of the Basque Country (UPV/EHU).

Aduriz, I., I. Aldezabal, I. Alegria, J. M. Arriola, A. Díaz de Ilarraza, N. Ezeiza, and K. Gojenola. 2003. Finite State Applications for Basque. In *EACL'2003 Workshop on Finite-State Methods in Natural Language Processing*, pages 3–11.

Aduriz, I., M. J. Aranzabe, J. M. Arriola, A. Atutxa, A. Díaz de Ilarraza, N. Ezeiza, K. Gojenola, M. Oronoz, A. Soroa, and R. Urizar. 2006a. Methodology and Steps Towards the Construction of EPEC, a Corpus of Written Basque Tagged at Morphological and Syntactic levels for Automatic Processing. *Language and Computers*, 56(1):1–15.

Aduriz, I., M. J. Aranzabe, J. M. Arriola, and A. Díaz de Ilarraza. 2006b. Sintaxi

Partziala. In B. Fernández and I. Laka, editors, *Andolin gogoan: Essays in Honour of Professor Eguzkitza*. pages 31–49.

Aduriz, I., J. M. Arriola, I. Gonzalez-Dios, and R. Urizar. 2015. Funtzio Sintaktikoen Gold Estandarra eskuz etiketatzeko gidalerroak. Technical report, University of the Basque Country (UPV/EHU) UPV/EHU/LSI/TR 01-2015.

Aduriz, I. and A. Díaz de Ilarraza. 2013. Morphosyntactic Disambiguation and Shallow Parsing in Computational Processing of Basque. *Anuario del Seminario de Filología Vasca "Julio de Urquijo"*, pages 1–21.

Aldezabal, I., O. Ansa, B. Arrieta, X. Artola, A. Ezeiza, G. Hernández, and M. Lersundi. 2001. EDBL: a General Lexical Basis for the Automatic Processing of Basque. In *Proceedings of the IRCS Workshop on linguistic databases*, pages 1–10.

Aldezabal, I., K. Ceberio, I. Esparza, A. Estarrona, J. Etxeberria, M. Iruskieta, E. Izagirre, and L. Uria. 2007. EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) segmentazio-mailan etiketatzeko eskuliburua. Technical report, University of the Basque Country (UPV/EHU) UPV/EHU/LSI/TR 11-2007.

Alegria, I., X. Artola, K. Sarasola, and M. Urkia. 1996. Automatic Morphological Analysis of Basque. *Literary and Linguistic Computing*, 11(4):193–203.

Aranzabe, M. J. and A. Díaz de Ilarraza. 2009. Análisis sintáctico computacional del euskera mediante una gramática de dependencias. In *Actas del XI Simposio Internacional de Comunicación Social*, pages 316–320. Centro de Lingüística Aplicada.

Arriola, J. M. 2015. Different Issues in the Design and Implementation of a Rule Based Grammar for the Surface Syntactic Disambiguation of Basque. In *Proceedings of the Workshop on "Constraint Grammar-methods, tools and applications" at NODALIDA 2015*, number 113, pages 1–9. Linköping University Electronic Press.

Bick, E. 2000. *The Parsing System Palavras. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Ph.D. thesis, Aarhus University Press.

Fleiss, J. L. 1971. Measuring Nominal Scale Agreement among many Raters. *Psychological bulletin*, 76(5):378–382.

Karlsson, F., A. Voutilainen, J. Heikkila, and A. Anttila. 1995. *Constraint Grammar, A Language-independent System for Parsing Unrestricted Text*. Mouton de Gruyter.

Marcus, M. P., M. A. Marcinkiewicz, and B. Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330.

Mille, S., A. Burga, V. Vidal, and L. Wanner. 2009. Towards a Rich Dependency Annotation of Spanish Corpora. *Procesamiento del Lenguaje Natural*, (43):325–333.

Nilsson, J. and J. Hall. 2005. Reconstruction of the Swedish Treebank Talbanken. Technical report, Växjö University, Sweden. School of Mathematics and Systems Engineering. MSI report 05067.

Sampson, G. 2011. A Two-way Exchange between Syntax and Corpora. In V. Vander, S. Zyngier, and G. Barnbrook, editors, *Perspectives on Corpus Linguistics*, volume XVI, 256. pages 197–211.

Scheible, S., R. J. Whitt, M. Durrell, and P. Bennett. 2011. A Gold Standard Corpus of Early Modern German. In *Proceedings of the ACL-HLT 25th Linguistic Annotation workshop*, pages 124–128. Association for Computational Linguistics.

Silveira, N., T. Dozat, M.-C. de Marneffe, S. R. Bowman, M. Connor, J. Bauer, and C. D. Manning. 2014. A Gold Standard Dependency Corpus for English. In *Proceedings of LREC 2014, the Ninth International Conference on Language Resources and Evaluation*, pages 2897–2904.

Solberg, P. E., A. Skjærholt, L. Øvrelid, K. Hagen, and J. B. Johannessen. 2014. The norwegian dependency treebank. In *Proceedings of LREC'14, the Ninth International Conference on Language Resources and Evaluation*, pages 789–795.

Voutilainen, A., T. Purtonen, and K. Muhonen. 2012. Outsourcing Parsebanking: The FinnTreeBank Project. In *Shall We Play the Festschrift Game? Essays on the Occasion of Lauri Carlson's 60th Birthday*. Springer, pages 117–131.

Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 133-140 recibido 01-12-2016 revisado 10-01-2017 aceptado 12-02-2017

Lingmotif: A User-focused Sentiment Analysis Tool

Lingmotif: una Herramienta de Análisis de Sentimiento Enfocada en el Usuario

Antonio Moreno-Ortiz
Universidad de Málaga
29071 Málaga, Spain
amo@uma.es

Abstract: In this paper, we describe Lingmotif, a lexicon-based, linguistically-motivated, user-friendly, GUI-enabled, multi-platform, Sentiment Analysis desktop application. Lingmotif can perform SA on any type of input texts, regardless of their length and topic. The analysis is based on the identification of sentiment-laden words and phrases contained in the application's rich core lexicons, and employs context rules to account for sentiment shifters. It offers easy-to-interpret visual representations of quantitative data, as well as a detailed, qualitative analysis of the text in terms of its sentiment. Lingmotif can also take user-provided plugin lexicons in order to account for domain-specific sentiment expression. As of version 1.0, Lingmotif analyzes English and Spanish texts. Lingmotif thus aims to become a general-purpose Sentiment Analysis tool for discourse analysis, rhetoric, psychology, marketing, the language industries, and others.
Keywords: Sentiment analysis, content analysis, discourse analysis, digital humanities.

Resumen: En este artículo se describe Lingmotif, una aplicación de Análisis de Sentimiento multi-plataforma, con interfaz gráfica de usuario amigable, motivada lingüísticamente y basada en léxico. Lingmotif efectúa Análisis de Sentimiento sobre cualquier tipo de texto, independientemente de su tamaño o tema. El análisis se basa en la identificación en el texto de palabras y frases con carga afectiva, contenidas en los diccionarios de la aplicación, y aplica reglas de contexto para dar cabida a modificadores del sentimiento. Ofrece representaciones gráficas fáciles de interpretar de los datos cuantitativos, así como un análisis detallado del texto. Lingmotif también puede utilizar léxicos del usuario a modo de *plugins*, de tal modo que es posible analizar de forma efectiva la expresión del sentimiento en dominios específicos. La versión 1.0 de Lingmotif está preparada para trabajar con textos en español e inglés. De este modo, se conforma como una herramienta de propósito general en el ámbito del Análisis de Sentimiento para el análisis del discurso, retórica, psicología, marketing, las industrias de la lengua y otras.
Palabras clave: Análisis de sentimiento, análisis de contenido, análisis del discurso, humanidades digitales.

1 Introduction[1]

Sentiment Analysis (SA), along with text analytics in general, has experimented increased attention in the last 15 years, no doubt due to the ever-increasing surge of user-generated content (UGC) on the World Wide Web, a vast body of knowledge that companies and organizations seek to sift, probe, and make sense of. Since text is the form that most of this knowledge is encoded as, it is no surprise that text analytics, or text mining, has become the focus of many research efforts.

Such strong interest has resulted in a vast body of technical knowhow, academic publications, and software. Most available SA software, however, is in the form of either code libraries, usually as part of NLP toolkits for developers, such as NLTK (Loper and Bird, 2002), Stanford CoreNLP (Manning et al., 2014), Apache OpenNLP (Morton et al., 2005), or end-user, "black-box", commercial applications and services, mostly focused on the

[1] This research was supported by Spain's MINECO through the funding of project Lingmotif2 (FFI2016-78141-P).

analysis of user-generated content, such as that produced by social networking sites.

These tools usually make use of supervised, Machine Learning techniques, which implies that they either require that users train the classifiers on their own data sets (in the case of developer libraries) or rely on the trained algorithms offered by commercial products.

The disadvantage of the first type of tools is that users are required to possess certain programming skills, whereas the latter offer no indication of what was found in the text that was used to classify a text as positive or negative. Furthermore, such tools are almost invariably geared toward short texts where opinion or sentiment is known to be expressed: user reviews, tweets or other online UGC. Their applicability to longer, multi-topic texts is simply not considered.

However, the automatic identification and analysis of sentiment in texts is interesting not just for sifting online UGC sources, but for many other applications, such as content and discourse analysis. Lingmotif attempts to tackle such needs by taking a radically different approach, and opens a door to a wider range of applications than current Sentiment Analysis tools offer. It can be used as classifier in the "traditional" sense, but it can also be used as a general-purpose text analysis tool that will show the sentiment profile of long texts, identify and clearly display sentiment expressions, provide a number of useful text metrics, compare texts alongside one another, produce analysis of a time series, and more.

Lingmotif is available as desktop application for the Windows, Mac OS, and Linux platforms, and is free for non-commercial uses. Currently, it supports English and Spanish input text, with ongoing development for French, German and Italian.

1.1 Approaches to Sentiment Analysis

Two approaches are distinguished to tackle the automatic analysis of semantic orientation. Most systems, as mentioned above, make use of statistical, Machine Learning techniques, mostly supervised methods, where the SA problem is seen as one of classification: a text is either positive or negative (sometimes finer-grained categories) to be classified under one of these classes. In these systems, a set of tagged examples of the type the classifier is meant to deal with (the training set) is used to train the classifying algorithms. The algorithm is then evaluated against a second set of tagged examples (the evaluation set), and accuracy metrics (in terms of precision and recall) are obtained that allow such systems to be compared in terms of performance. A classic example of such systems is Pang, Lee and Vaithyanathan (2002). Machine Learning classifiers generally work well with the type of content they have been trained for, but their performance drops, almost to chance, when they are used with other types of texts (Taboada et al., 2011). Several approaches have been used to adapt ML-based classifiers to various subject domains (Aue and Gamon, 2005, Choi, Kim and Myaeng, 2009), but the problem remains.

The second approach involves the use of rich lexical sources where sentiment-carrying lexical items are listed. The task of determining the semantic orientation of a text, consists of identifying such items in the input texts, perhaps analyze their context, and perform calculations on the identified items. A classic example of this type of system is Turney (2002).

2 Sentiment Analysis with Lingmotif

Lingmotif is a lexicon-based SA system, since it uses a rich set of lexical sources and analyzes context in order to identify sentiment laden text segments and produce two scores that qualify a text from a SA perspective. In a nutshell, it breaks down a text into its constituent sentences, where sentiment-carrying words and phrases are searched for, identified, and assigned a valence (i.e., a sentiment index). The complete analysis process is explained in section 4 below.

2.1 Levels of analysis

Sentiment Analysis, as a classification task, can take different text units as the object of classification. Traditionally, most SA systems have focused on document-level classification, that is, their function is to classify an input text as positive or negative (Turney, 2002, Pang, Lee and Vaithyanathan, 2002). Fewer systems have taken sentences (Wiebe and Riloff, 2005) or clauses (Wilson, Wiebe and Hwa, 2004, Thet et al., 2009) as classification segments. The reason why most systems perform document-level classification is simply that they are designed to classify short documents, traditionally, user reviews. Lingmotif analyzes text at the sentence-level, which, from a linguistic point of view, leaves much to be desired, since many sentiment indicators operate extra-sententially.

2.2 Sentiment shifters

Sentiment shifters, or contextual valence shifters, were first proposed by Polanyi and Zaenen (2006) as a mechanism to account for the modification (or *shift*) of the sentiment of a given lexical unit by means of its surrounding context. Since then, they have been implemented in a number of lexicon-based SA systems: Kennedy and Inkpen, 2006, Moreno Ortiz et al., 2010, Taboada et al., 2011. Sentiment can be altered by context in different ways: it can be intensified, diminished, or it can be inversed altogether. Negation is probably the most relevant shifter (Wiegand et al., 2010), since it usually inverts the polarity of the lexical item it modifies, but intensification and downtoning need also be addressed. In section 3.2 below, we describe Lingmotif's CVS system.

2.3 Analysis modes

Lingmotif uses a simple, but efficient GUI that allows users to select input and options, and launch the analysis (see Figure 1). Results are generated as an HTML document, which is saved to a predefined location and automatically sent to the user's default browser for immediate display.

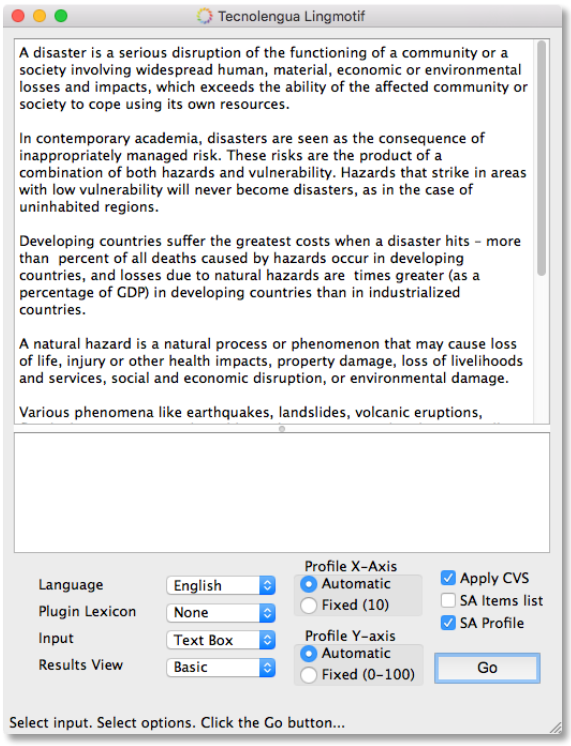

Figure 1: Lingmotif's GUI

Internally, the application generates results as an XML document containing all the relevant data; this XML document is then parsed against one of several available XSL templates, and transformed into the final HTML and Javascript. This interface allows users to simply type or paste a text in the input text area, or load a number of text files to be analyzed. Lingmotif works in either single-document or multi-document mode.

2.3.1 Single-document mode

Whether in single or multi-document mode, Lingmotif will always produce a number of metrics for each individual text, which we list below:

- TSS: Text Sentiment Score: the text's overall sentiment score.
- TSI: Text Sentiment Intensity: the proportion of sentiment vs non-sentiment items. These two are shown graphically (Figure 2)

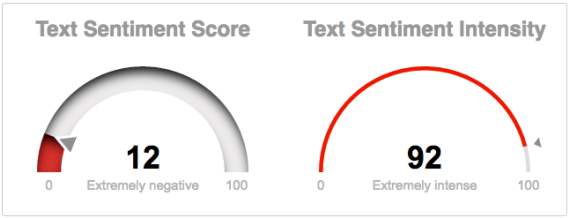

Figure 2: TSS and TSI gauges

- Sentiment Profile: a graphical representation of the text's sentiment "flow". See Figure 3.

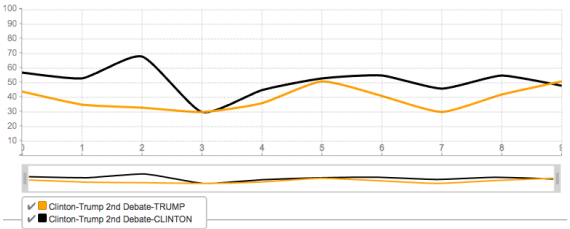

Figure 3: Sentiment profile (parallel mode)

- Text Analysis: several text metrics: number of tokens, types, sentences, lexical and function words, etc.
- Quantitative Sentiment Analysis: a breakdown of the figures that were obtained in order to come up with the TSS and TSI (See Figure 4).
- Detailed Sentiment Analysis: a display of the input text where sentiment items are color coded according to their polarity and

specific data for each are (optionally) displayed (See Figure 5).

Text Analysis

Text Stats			Sentences	Words by Function		Words by Form	
Tokens	Types	T/T Ratio		Lexical	Grammatical	Single Words	Multiwords
6898	1293	18.74%	637	2916	3982	6409	489

Sentiment Analysis

TSS	TSI	CVS segments	Positive items	Negative items	Neutral items	Positive Score	Negative Score
38	88	95	169	298	2416	379	693

Figure 4: Quantitative data

Detailed Sentiment Analysis

Well, I actually agree_with that.

I agree_with everything she said.

I began this campaign because I was so_tired_of seeing such foolish things happen to our country.

This is_a_great country.

This is_a_great land.

I have gotten_to_know the people of the country over the last year-and-a-half that I have been doing this as a politician.

Figure 5: Detailed sentiment analysis

The most relevant results are the TSS and TSI. These two scores qualify a text in terms of its orientation (TSS) and intensity (TSI). Both are displayed by means of visual, animated gauges at the top of the results page. These gauges also include a category for each, from "extremely negative" to "extremely positive", which makes numeric results readily interpretable by the user (see Figure 2). We describe how these scores are obtained in section 4.

However, for long texts, the Sentiment Profile is a powerful that can provide a quick insight into the text's internal structure and organization in terms of sentiment expression. This graph is interactive: hovering the data points will display the lexical items that make up that particular text segment. The quantitative data tables are also quite useful when comparing texts (see next section), since it readily offers useful common text metrics, such as type/token ratio.

2.3.2 Multi-document analysis

Multi-document mode is enabled simply by loading multiple files or one multi-document file (i.e., a file where each line is assumed to be a –short– document, such as tweets or user reviews).

When in multi-document mode, Lingmotif will analyze documents one by one, generating one HTML file for each, although they will not be displayed on the browser, just saved to the output folder. When the analysis is finished, a single results page will be displayed. This page is a summary of results, and is different from the single-document results page: the gauges for TSS and TSI are now the average for the analyzed set and the detailed analysis section contains a quantitative analysis of each of the files in the set. The first column in this table shows the title of the document (file name without extension) as a hyperlink to the HTML file for that particular file.

Multi-document mode has several modes of operation:

- Classifier (default): a stacked bar graph and data table are offered showing classification results based on their TSS category. The graph offers a visualization of results; both its legend and the graph itself are interactive (see Figure 6).

TOTAL DOCUMENTS								
469								
TSS CLASSIFICATION								
Extremely	Very	Fairly	Slightly	Neutral	Slightly	Fairly	Very	Extremely
29	40	53	48	207	28	32	14	18
THREE-WAY CLASSIFICATION								
Negative				Neutral	Positive			
170 (36.25%)				207 (44.14%)	92 (19.62%)			
BINARY CLASSIFICATION								
Negative					Positive			
274 (58.42%)					195 (41.58%)			

Figure 6: Classifier data table

- Series: the set of loaded files is assumed to be in order, chronological (time series) or otherwise. Each data point in the Sentiment Analysis Profile represents one document. The data point is the average TSS for that particular document.
- Parallel: produces a graph with one line for each file (this mode is limited to 15 documents). This is useful to compare sentiment flow in texts side by side.
- Merge: this option merges all loaded individual files in one single text.

3 *Lexical resources*

Lingmotif's performance is directly proportionally to the quality of its lexical resources. The creation of our current lexical resources has been our focus for many years. Work on Spanish the lexicon started with the Sentitext project (Moreno-Ortiz et al., 2010) and was further expanded, refined adapted to Lingmotif's format during the Lingmotif 1 project, along with the creation of the English resources.

For each language, Lingmotif requires the following resources:

- A full-coverage core sentiment lexicon which contains both single words and

multiword expressions.

- A set of context rules, where sentiment shifters are defined using a template approach.
- A part of speech tagger.
- A lemmatizer, used for generating inflected forms during lexicon imports or updates.
- Optionally, a plugin lexicon can be used to account for domain-specific sentiment expression.

3.1 Core lexicon

The core lexicon is the most important resource. A lexical item in a Lingmotif lexicon (whether a single word or a multiword expression) is defined by a specification of its form, part of speech, and valence. The valence is an integer from -5 to -2 for negatives and 5 to 2 for positives. The item's form can either be a literal string or a lemma (represented between angled brackets). As for the part-of-speech specification, Lingmotif uses the Penn Treebank tag set for all languages. A wildcard (`ALL`) can be used for cases where all possible parts of speech for that lemma share the same valence. Figure 7 below shows some examples.

```
in_cold_blood,RB,-4
<kill>_time,VB,0
<kill>,ALL,-3
broke,JJ,-2
```

Figure 7: Lingmotif lexicon format

The creation process of Lingmotif's core lexicons basically involves merging freely available sentiment lexicons, adapt the merged list to our format and refine it using corpus analysis techniques and sheer heuristics. For English, we merged items from The Harvard General Inquirer (Stone and Hunt, 1963), MPQA (Wilson, Wiebe and Hoffmann, 2005), and Bing Liu's Opinion Lexicon (Hu and Liu, 2004). These resources were expanded by using a thesaurus and derivational generation rules. These resources, however, are characterized by their lack of attention to multiword expressions. We then used a number non SA-specific lexical resources, including common idioms from Wiktionary, which we tagged manually for valence. Ultimately, Lingmotif's lexicons are the result of intensive lexicographical work. A similar processed was followed for the Spanish language. The English lexicon contains over 77,000 entries (word forms) and nearly 500 context rules. The Spanish lexicon contains 207,000 word forms and over 300 context rules.

Sentiment disambiguation is currently dealt with using exclusively formal features: part-of speech tags and multi-word-expressions. MWEs usually include words that may or may not have the same polarity of the expression. including such expressions can solve disambiguation for many cases. For example, we can classify as negative the word "kill" and then include phrases such as "kill time" with a neutral valence. When this is not possible, the options are to include it with the more statistically probable polarity or simply leave it out when the chances of getting the item with one polarity or another are similar.

3.2 Context rules

Context rules are Lingmotif's mechanism to deal with sentiment shifters. They work by specifying words or phrases that can appear in the immediate vicinity of the identified sentiment word. Basically, we use the same approach as Polanyi and Zaenen (2006) or Kennedy and Inkpenn (2006), or Taboada (2011): we use simple addition or subtraction (of integers on a -5 to 5 scale in our case).

In Lingmotif, every rule specifies the following:

- The part of speech and polarity of the sentiment word.
- The form, location (left or right), and span (in number of words) of the shifter.
- The result of the rule application.

Currently, Lingmotif uses over 400 such rules for each language. Table 1 below shows examples of all types of sentiment shifters according to the effect they produce on the resulting text segment.

Shift type	Example Context Rule
Inversion	`NN,-,avoid*,LR,5,INV0`
	`JJ,+-,not,L,2,INV0`
Intensification	`JJ,+-,seriously,L,2,INT3`
	`VB,+-,may_well,L,1,INT1`
Downtoning	`NN,-,mild,L,2,DOW1`
	`NN,+-,a_bit,L,2,DOW1`

Table 1: Context rules types and examples

Lingmotif's context rules were compiled by extensive corpus analysis, studying concordances of common polarity words (adjectives, verbs, nouns, and adverbs), and then testing the rules against texts to further improve

and refine them.

When a context rule is matched, the resulting text segment is marked as a single unit and assigned the calculated valence, as specified in the rule. Multiple context rule matching is possible and not handled at present: as soon as a rule is matched, no further rules are searched and the rule is applied. It would definitely be interesting to improve this by establishing a priority system for rules.

3.3 Plugin lexicons

As many researchers have pointed out (Aue and Gamon, 2005, Turney 2002, Read, 2005, Pang and Lee, 2008), sentiment is very often dependent on topic, or domain. Attributes such as size, weight, or location, for example, can be regarded as positive or negative, or neither, depending on whether we are discussing electronic gadgets, hotels or movies.

Being a general-purpose SA system, Lingmotif provides a flexible mechanism to adapt to specific domains by means of user-provided lexicons. Lexical information contained in plugin lexicons overrides Lingmotif's core lexicon. When a plugin lexicon is selected for analysis, the plugin lexicon is searched first. If a word or phrase is found there, the core lexicon will not be searched for that item, and its information in the plugin lexicon will be used. Thus, plugin lexicons can be used to provide domain-specific sentiment items, but also to override polarity assignment in the core lexicon, for whatever reasons.

Plugin lexicons have exactly the same format as the core lexicon. In order to import a plugin lexicon, it must first be created as a UTF-8 encoded CSV file, which is then imported. Updating a plugin lexicon simply involves modifying the source CSV file and importing it again. Any number of plugin lexicons can be created in Lingmotif, but only one can be used for a given analysis.

4 Analysis process

As mentioned above, Lingmotif's results are obtained by identifying sentiment-laden words and multiword expressions, analyzing their contexts for sentiment shifters, and weighing sentiment against non-sentiment items. In this section, we describe this process in detail.

The analysis process is the following:

1. Preprocessing: text is scanned for common abbreviations, contractions and misspelling.
2. Tokenization: both sentence-level and word-level.
3. Multiword identification: n-grams are matched against the list multiword expressions contained in the plugin lexicon (if selected) and core lexicon. Identified MWEs are marked and assigned their valence.
4. Polarity words identification: individual words are looked up in the lexicons and assigned their valence if found.
5. Context rule matching: identified polarity words and MWEs are matched against the list of context rules. If a rule matches the word's context, the whole text segment is marked and tagged as a unit of type CVS (context valence shifter). This process is repeated twice, in order to account for cumulative sequences of shifters, such as "very very good".
6. TSS and TSI calculation and category assignment.
7. Generation of internal XML document, which contains the results data for the input text.
8. Generation of Javascript code for the graphical components.
9. Generation of the final results HTML document.

Calculation of the Text Sentiment Intensity (TSI) and Text Sentiment Score (TSS) metrics deservers further discussion. Valences found in lexical units are added and weighed against the number of neutral *lexical* units. Therefore, we do not use the simpler "term-counting method" (Kennedy and Inkpen, 2006): a particularly intense unit can have more weight in the overall score than two less intense units. Also, function words do not enter into the equation. Lingmotif offers all the figures employed to come up with the final scores in the "Text Analysis" section (see Figure 4 above).

First TSI is calculated as the proportion of sentiment vs non-sentiment items (or rather, their added scores). TSI calculation takes text length into account in order to capture the fact that the ratio of sentiment to non-sentiment items is necessarily lower the longer the text. A *max_tsi* factor is created which places stronger weight in sentiment words in longer texts. TSI is then calculated as

$$TSI = \frac{pos_{score}\% + neg_{score}\%}{\max\ \_tsi}$$

TSS is then calculated as

$$TSS = \frac{\sum \begin{cases} v * sw, if\ v \neq 50 \\ v, otherwise \end{cases}}{v_n * 100}$$

where *v* is the value of each lexical unit, and *sw* is the sentiment weight, a factor inversely proportional to the *max_tsi* previously calculated.

5 Performance evaluation

Even though Lingmotif was not conceived as a classifier, it can compete with ML-based classifiers in terms of performance. In this section, we employ some readily available SA-tagged data sets to evaluate Lingmotif's performance, obtaining outstanding results. Table 2 below shows the confusion matrix and precision, recall and f-measure figures summarizing the evaluation results against the STS-Gold data set (Saif et al., 2013), which is a data set specifically designed to serve as a gold standard in Sentiment Analysis of Twitter text.

	POS	NEG	TOTAL
POS	522	97	619
NEG	254	1125	1379
	Evaluation		
	Precision	Recall	F-measure
	0.92	0.82	0.87

Table 2: STS-Gold data set results

It must be mentioned that Lingmotif's is not a binary classifier, therefore a number of the documents were actually classified as binary. Specifically, 294 of the 1,379 negative documents and 145 of 719 positive documents. Neutral documents are simply coerced randomly as positive or negative in order to be able to compare results alongside binary classifiers.

Results were similar with other Twitter data sets, such as UMICH SI650[2] or the Stanford Twitter Sentiment Test Set (STS-Test) (Go, Bhayani, and Huang, 2009), which we show in Table 3 below.

	POS	NEG	TOTAL
POS	139	22	161
NEG	37	132	169
	Evaluation		
	Precision	Recall	F-measure
	0.86	0.78	0.82

Table 3: STS-Test data set results

[2] https://inclass.kaggle.com/c/si650winter11/data

6 Conclusions

Being an end-user tool, evaluating Lingmotif requires more than accuracy figures exclusively. Aspects such as usability and adequacy to specific tasks should also be discussed. We believe that the application also addresses such aspects successfully. All in all, Lingmotif is a platform that offers many possibilities for the analysis of texts from a Sentiment Analysis perspective. Its lexicon-based approach, coupled with a careful curation of its resources results in highly accurate results.

Even so, there are many ways in which it can be improved. Specifically, sentiment disambiguation is only partially dealt with. Current context rules are limited in their expressive power, and would no doubt benefit if semantic categories could be specified, rather than simply words or lemmas. In general, deeper semantic analysis of context would be necessary to improve on current results, both at the sentence level and at the text level.

7 References

Aue, A., and M. Gamon. 2005. Customizing Sentiment Classifiers to New Domains: A Case Study. In *Recent Advances in Natural Language Processing (RANLP-05)*. Borovets, Bulgaria.

Choi, Y., Y. Kim, and M. Sung-Hyon. 2009. Domain-Specific Sentiment Analysis Using Contextual Feature Generation. In *Proceeding of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion*, 37–44. Hong Kong, China: ACM.

Go, A., R. Bhayani, and L. Huang. 2009. Twitter Sentiment Classification Using Distant Supervision. CS224N Project Report. Stanford.

Hu, M., and B. Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 168–77. Seattle, WA, USA: ACM.

Kennedy, A., and D. Inkpen. 2006. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence*, 22 (2): 110–25.

Loper, E., and S. Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, 1, 63–70. ETMTNLP '02. Stroudsburg, PA, USA: ACL.

Manning, C., M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 55–60. ACL.

Moreno-Ortiz, A., Á. Pérez Pozo, and S. Torres Sánchez. 2010. Sentitext: Sistema de Análisis de Sentimiento para el Español. *Procesamiento de Lenguaje Natural,* 45: 297–98.

Morton, T., G. Bierner, J. Kottmann, and J. Baldridge. 2005. OpenNLP: A Java-Based NLP Toolkit. Available at https://opennlp.apache.org.

Pang, B., and L. Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval,* 2 (1–2): 1–135.

Pang, B., L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, 10, 79–86. ACL.

Polanyi, L., and A. Zaenen. 2006. Contextual Valence Shifters. In *Computing Attitude and Affect in Text: Theory and Applications, Shanahan*, James G., Qu, Y., and Wiebe, J., 20:1–10. The Information Retrieval Series. Dordrecht, The Netherlands: Springer.

Read, J. 2005. Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification. In *Proceedings of the ACL Student Research Workshop*, 43–48. ACLstudent '05. Stroudsburg, PA, USA: ACL.

Saif, H., M. Fernández, Y. He, and H. Alani. 2013. Evaluation Datasets for Twitter Sentiment Analysis: A Survey and a New Dataset, the STS-Gold. Turin, Italy.

Stone, P. J., and E. B. Hunt. 1963. A Computer Approach to Content Analysis: Studies Using the General Inquirer System. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference*, 241–56. AFIPS '63 (Spring). New York, NY, USA: ACM.

Taboada, M., J. Brooks, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2): 267–307.

Thet, T. T., J. Na, C. S. G. Khoo, and S. Shakthikumar. 2009. Sentiment analysis of movie reviews on discussion boards using a linguistic approach. In *Proceeding of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion*. 81–84. Hong Kong, China: ACM.

Turney, P. D. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting of the ACL*, 417–24. Philadelphia, USA.

Wiebe, J., and E. Riloff. 2005. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In *Computational Linguistics and Intelligent Text Processing*, 486–97. Lecture Notes in Computer Science 3406. Springer Berlin Heidelberg.

Wiegand, M., A. Balahur, B. Roth, D. Klakow, and A. Montoyo. 2010. A Survey on the Role of Negation in Sentiment Analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, 60–68. NeSp-NLP '10. Stroudsburg, PA, USA: ACL.

Wilson, T., J. Wiebe, and P. Hoffmann. 2009. Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis. *Computational Linguistics*, 35(3): 399–433.

Wilson, T., J. Wiebe, and R. Hwa. 2004. Just How Mad Are You? Finding Strong and Weak Opinion Clauses. In *Proceedings of the 19th National Conference on Artifical Intelligence*, 761–67. AAAI'04. San Jose, California: AAAI Press.

Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 141-148 recibido 01-12-2016 revisado 10-01-2017 aceptado 10-02-2017

Analizando opiniones en las redes sociales*

Analysing Opinions in Social Networks

Javi Fernández, Fernando Llopis, Patricio Martínez-Barco, Yoan Gutiérrez, Álvaro Díez
Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante
Apdo. de Correos 99, E-03080, Alicante, Spain
{javifm,llopis,patricio,ygutierrez,adiez}@dlsi.ua.es

Resumen: La Web 2.0 ha focalizado la importancia de la información, no en unos pocos expertos en un tema, sino en una multitud de opiniones vertidas por usuarios a través de diversos medios en las redes sociales. Debido a ello, han cobrado un mayor interés los sistemas que son capaces de determinar qué es lo que piensan los usuarios sobre un determinado concepto, agregando diferentes fuentes de datos y aplicando cálculos de polaridad de las opiniones, que permiten determinar y comparar esos conceptos con otros similares. En este artículo describimos *Social Analytics*, nuestra visión sobre cómo deberían funcionar este tipo de sistemas, con una interfaz simple y optimizada que permita responder las necesidades de los usuarios.
Palabras clave: Análisis de opiniones, minería de opiniones, recuperación de información, redes sociales, turismo

Abstract: Web 2.0 has focused the importance of information, not on a few experts on a topic, but on a multitude of opinions expressed by users through various media on social networks. Due to this, there has been a increasing interest in systems that are able to determine what users think about a certain concept, by adding different sources of data and applying polarity calculations, to determine and compare these concepts with similar ones. In this paper we describe *Social Analytics*, our vision on how these kind of system should work, with a simple and optimized interface to meet the needs of the users.
Keywords: Sentiment analysis, opinion mining, information retrieval, social networks, tourism

1 Introducción

En la película *The Naked Jungle* se relataba la historia de una joven americana, Eleanor Parker, que se iba a casar con un terrateniente instalado en América del Sur, Charlton Heston. La joven llegaba con cierto temor a una jungla ante el peligro que podían suponer enormes animales salvajes. Lo que ella ignoraba es que el peligro real lo iban a suponer unas diminutas hormigas, que agrupadas en forma de millones podían suponer una marabunta que arrasara todo a su paso.

En la misma línea, con la Web 2.0 todo ha cambiado. Están empezado a perder valor las opiniones de grandes expertos en un tema concreto frente a las de cientos de usuarios que, con menos experiencia y conocimientos, opinan sobre los mismos temas. Si hablamos sobre teoría de la relatividad, sin duda alguna acudiríamos a los escritos de Einstein o discípulos para tratar de obtener más información. Es probable que nadie, o al menos casi nadie, pusiera en duda sus postulados frente a las opiniones de cientos de estudiantes de primero de universidad. Pero en cada vez más ámbitos estos cientos de opiniones pueden pesar más ante la toma de decisiones.

El sector turístico es un sector donde los conceptos web 2.0 se han impuesto. En una búsqueda del tipo «visitar iglesias de alicante capital» provocaría un resultado parecido al

* Este trabajo ha sido parcialmente financiado por el Ministerio de Educación, Cultura y Deporte (MECD FPU014/00983), y la Universidad de Alicante, la Generalitat Valenciana y el Gobierno Español a través de los proyectos TIN2015-65136-C2-2- R, TIN2015-65100-R, PROMETEOII/2014/001 y FUNDACIONBBVA2-16PREMIOI.

ISSN 1135-5948

mostrado en la Figura 1. La primera y la tercera opción propuesta corresponden a páginas oficiales, la primera al organismo de turismo de la ciudad y la tercera a la página del Ayuntamiento. Sin embargo, la segunda es la habitualmente seleccionada en un segundo click. Esta segunda corresponde a la famosa web de reservas e información turística *TripAdvisor*[1]. Dicha web contiene pequeñas opiniones de muchos usuarios. Los responsables de hoteles ya temen más unas cuantas opiniones negativas en una web de ese tipo que un análisis demoledor realizado por un experto y publicado en un diario o revista especializada.

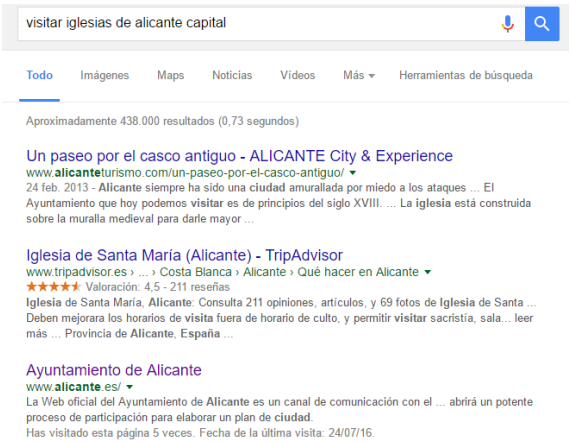

Figura 1: Retorno de un buscador de Internet

Al mismo nivel, aparecen las opiniones que los usuarios pueden verter habitualmente en las redes sociales. No hay mejor publicidad para un destino turístico que un usuario de una red social publique un comentario positivo sobre un lugar, y más todavía si es acompañado de una fotografía que inmortaliza el momento. Una playa, un atardecer, un copioso plato de arroz, un helado junto a un comentario positivo realizado por una persona a la que sigues en redes sociales es la mejor publicidad. De la misma forma, un comentario negativo acerca de la suciedad o inseguridad pueden ser auténticamente demoledores.

Pero al igual que ocurría en *The Naked Jungle*, una sola hormiga no produce ningún temor. De hecho, la aparición de una de ellas puede responder a un hecho anecdótico. La unión de ellas en forma de marabunta supone todo lo contrario. En el mundo de la Web 2.0 cada vez es más importante conocer en tiempo real las opiniones de la gente sobre lo que les interesa, les gusta o desprecian.

Por ejemplo, en el contexto del mundo del turismo (aunque sería aplicable a muchos más ámbitos) existe una doble visión: por una parte, la del usuario que pretende realizar una reserva; y por otro lado la del gestor del producto, que desea vender el suyo. Los objetivos del primero serían:

- No quiero equivocarme, ni pagar de más.
- Quiero conocer muchas opiniones, pero no tener que leerlas todas.
- Quiero que esas opiniones provengan del mayor número de fuentes posible.
- Quiero opiniones de expertos, pero también de gente como yo o en mis mismas circunstancias.

Y los objetivos del segundo:

- Quiero saber cómo conservar a mis clientes y ganar nuevos.
- Quiero conocer lo que piensa mi cliente sobre mí y sobre mi competencia.
- Quiero descubrir nuevos aspectos sobre mis potenciales clientes.

Ese conocimiento debe circunscribirse a la más inmediata actualidad, de nada me sirve saber que hace tres años un hotel era muy cómodo o que en el mismo período de tiempo un cliente pensara que la limpieza del local no era óptima. Como decía Lope de Vega «Lo que cuenta no es mañana, sino hoy. Hoy estamos aquí, mañana tal vez, nos hayamos marchado». Además, otro aspecto también de mucho interés es el concepto de medición y valoración. Como indica J.P. Rayo «Lo que no se mide no se puede gestionar». Necesitamos unos valores que nos determinen si mejoramos con respecto a días anteriores y cuál es nuestra posición con respecto a nuestra competencia.

En este artículo se presenta el sistema *Social Analytics*, que en base a las anteriores premisas analiza las redes sociales *Twitter*[2] e *Instagram*[3], obteniendo mediciones de reputación de ciertos conceptos parametrizables en base al número de opiniones positivas y negativas recibidas. El sistema determina la valoración en base a una serie de fórmulas

[1] https://www.tripadvisor.es

[2] https://twitter.com

[3] https://www.instagram.com

que tienen en cuenta los aspectos de positividad y negatividad de las menciones, así como la influencia de los que las realizan. La información se agrupa de forma gráfica y en texto resumido.

El artículo se estructura en los siguientes apartados. El apartado 3 describirá la arquitectura del sistema *Social Analytics*. En el apartado 4 se realizará una descripción de modelo utilizado para determinar si un comentario se puede calificar como positivo, negativo o neutral. En el apartado 5 se detallarán las fórmulas utilizadas para determinar cómo se mide la reputación de un concepto. La sección 6 describirá la evaluación realizada. Por último, el apartado 7 recoge una serie de conclusiones y trabajos que se están realizando actualmente para mejorar el sistema.

2 Trabajos relacionados

Existen otros sistemas para visualizar datos de redes sociales, muchos de ellos centrados únicamente en estadísticas o únicamente en análisis de opiniones (Marcus et al., 2011; Hao et al., 2011; Wang et al., 2012). Nuestro sistema contiene visualizaciones de los dos tipos y además ofrece un valor de reputación para poder realizar estudios en el tiempo y poder realizar comparaciones. Entre los datos que extraemos podemos destacar: el texto del mensaje, la fecha de publicación, el autor, los usuarios mencionados, el lugar donde se ha escrito el mensaje, y el lugar de origen del autor. Este último dato es de formato libre y los usuarios pueden rellenarlo con lugares mal escritos o que no representen lugares reales. Obtener un lugar real a partir de estos datos es un problema muy complejo (Hecht et al., 2011; Peregrino, Tomás, y Llopis, 2013). Nuestra aproximación para este problema, junto a toda su arquitectura, se describe en la sección 3.

Por otro lado, es fundamental detectar si es positivo, negativo o si no se puede obtener un detalle claro de esa polaridad, en muchas ocasiones porque no la tiene. Se han utilizado diferentes técnicas tanto el análisis y estudio de esa polaridad en texto. El modelo propuesto por Hu y Liu (2004), se basa en un conjunto de adjetivos base o semilla que amplía utilizando las relaciones de sinonimia y antonimia proporcionadas por *WordNet* (Miller, 1995). *WordNet-Affect* es un modelo descrito en (Strapparava y Valitutti, 2004) y que ya utiliza seis categorías básicas de emociones (alegría, tristeza, miedo, sorpresa, ira y disgusto) que también son expandidas utilizando *Wordnet*. El mayor problema que tienen estas propuestas es que no utilizan el contexto local para definir esa polaridad. *Social Analytics* utiliza un módulo de cálculo de la polaridad basado en el uso de n-gramas ampliado, con el fin de añadir contexto a las palabras y mantener parte de su secuencialidad. Este enfoque se describe en el apartado 4.

3 Arquitectura del sistema

El sistema se ha dividido en tres módulos principales: *escucha*, *procesamiento* y *presentación*, que a su vez utilizan tres bases de datos diferentes (ver Figura 2). Describiremos cada una de estas partes con detalle a continuación:

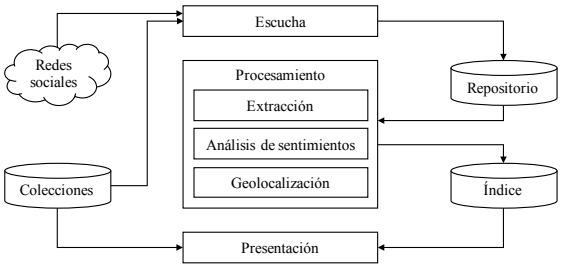

Figura 2: Arquitectura el sistema

3.1 Bases de datos

Primero describiremos las bases de datos utilizadas en nuestro sistema. Esto nos servirá para entender mejor cómo hemos representado los conceptos a analizar en las redes sociales y la información que finalmente queremos obtener.

- **Colecciones**. En esta base de datos se almacenan los términos de interés para los usuarios. Cuando hablamos de términos nos referimos a conjuntos de palabras, hashtags o menciones a usuarios de las redes sociales. Ya que varios términos pueden representar a un mismo concepto, los agruparemos en lo que hemos llamado *entidades*. Al mismo tiempo, también agruparemos las entidades en *colecciones*, que son agrupaciones de entidades dentro de una misma categoría.

  Por ejemplo, si estuviéramos interesados en lo que se comenta en las redes sociales sobre el Partido Popular (PP) y el Partido Socialista (PSOE) (dos partidos políticos conocidos en el panorama

político español), crearíamos dos entidades, una para cada partido. Estas entidades podrían tener diferentes términos por los que se las podría identificar, como "partido popular", #partidopopular y @ppopular para el primero, o psoe, #psoe y @psoe para el segundo. A su vez, estas entidades, se podrían agrupar en una colección a la que podríamos llamar *Política*.

- **Repositorio**. Esta es una base de datos temporal en la que se almacenan los mensajes y comentarios encontrados en las redes sociales. Aquí, los datos se almacenan rápidamente y sin procesar, y se eliminan una vez procesados.
- **Índice**. Esta también es una base de datos de comentarios generados en redes sociales. Pero, a diferencia de la base de datos anterior, los datos se almacenan procesados e indexados de manera que realizar estadísticas y análisis sea lo más óptimo y eficiente posible. Cabe destacar que no se almacenan los mensajes completos, sólo se indexan.

3.2 Módulo de escucha

Este módulo es el encargado de la descarga de mensajes de las redes sociales. Se realizan búsquedas periódicas de los términos añadidos previamente utilizando la *API*[4] de cada red social para obtener mensajes que contengan dichos términos. La frecuencia de búsqueda dependerá de los límites de la API y del número de términos que tengamos en el sistema. Algunas APIs permiten obtener estos mensajes en *streaming*, esto es, ofreciéndonos los mensajes según se van publicando, sin necesidad de hacer búsquedas periódicas. En ambos casos, los mensajes obtenidos se almacenan el anteriormente mencionado *Repositorio*.

3.3 Módulo de procesamiento

Este módulo realiza toda la parte de extracción de datos, detección de la localización y análisis de opiniones de los mensajes descargados mediante el módulo de escucha. Como hemos comentado antes, extraemos el texto del mensaje, la fecha de publicación, el autor, los usuarios mencionados, el lugar donde se ha escrito el mensaje, y el lugar de origen del usuario. Para este último punto hemos utilizado una aproximación muy simple, indexando una base de datos de lugares (*Geonames*[5]), realizando búsquedas en ella y quedándonos con el mejor resultado.

Utilizando el texto también obtenemos su polaridad, es decir, si el autor está escribiendo de forma positiva, negativa o neutral. Describiremos la aproximación utilizada con detalle en la sección 4.

3.4 Módulo de presentación

Este módulo se refiere a la interfaz de usuario. En la Figura 3 se puede ver la vista principal de un tablero de mando típico de Social Analytics. En el ejemplo podemos ver una comparativa de la reputación de los principales partidos políticos en España durante las elecciones de 2016.

De un sólo vistazo se puede acceder a los datos desde diferentes puntos de vista:

- Número de menciones que se han realizado sobre cada entidad. También es posible ver la evolución de este número en el tiempo.
- La audiencia a la que han llegado esas menciones (el número de seguidores de los autores de los mensajes).
- La reputación de cada entidad, según el número de menciones positivas, negativas y neutras (ver sección 5).
- Las palabras y hashtags más repetidos.
- Los usuarios que más han publicado.
- Los usuarios más mencionados.
- Los lugares en los que se han escrito los mensajes.
- Los lugares de origen de los usuarios que han publicado más mensajes.

Una de las grandes ventajas del sistema es que toda la información a considerar se puede parametrizar en base a diferentes aspectos, como el rango de fechas, la polaridad, o la fuente (*Twitter*, *Instagram* o ambas). El usuario también puede elegir en el mismo tablero si quiere cambiar alguna aspecto del informe y considerar solamente algunos elementos, como ver solo menciones positivas o negativas, ver las menciones de un usuario,

[4]Siglas de *Application Programming Interface*, o *Interfaz de programación de aplicaciones*

[5]http://www.geonames.org

Figura 3: Cuadro de mando principal de Social Analytics

ver las menciones donde aparezcan ciertas palabras o hashtags, el lugar donde se ha escrito el mensaje, etc. Podemos ver otro ejemplo en la Figura 4.

Figura 4: Palabras, hashtags y lugares más repetidos

Además, el sistema incorpora un modo «versus» que permite contrastar la reputación de una o varias entidades frente a otras, con similar información, tal como se muestra en la Figura 5. Así, de un vistazo podemos saber quién va ganando en una confrontación de parámetros, o qué se dice de uno y otro en el mismo período de tiempo.

4 *Análisis de las menciones*

Para detectar la polaridad de los mensajes hemos utilizado una aproximación híbrida (Fernández et al., 2013), que crea un lexicón a partir de un corpus etiquetado, y construye un clasificador utilizando técnicas de aprendizaje automático. Hemos evitado el uso de herramientas lingüísticas, con el fin de minimizar la propagación de errores externos a nuestra aproximación. A continuación detallaremos esta aproximación con más detalle.

4.1 Extracción de términos

Como no queremos perder información subjetiva del texto original, realizamos una normalización muy básica. Utilizar una normalización más compleja puede inducir más errores que se propagarían a los resultados finales. Comenzamos convirtiendo los textos a minúsculas, eliminando nombres de usuario y URLs. Después, realizamos una eliminación de caracteres repetidos parcial: si el mismo carácter se repite más de tres veces, el resto de repeticiones se elimina. De esta manera, las palabras se normalizan, pero todavía es posible reconocer si las palabras originales tienen caracteres repetidos.

Una vez los textos están normalizados, extraemos las palabras que contiene. Después, obtendremos nuevos términos agrupándolos según la posición en la que aparecen en el texto. La mayoría de aproximaciones utilizan *n-gramas* para mantener parte de la secuencialidad de los textos, pero de manera muy estricta, ya que obligan a que los términos

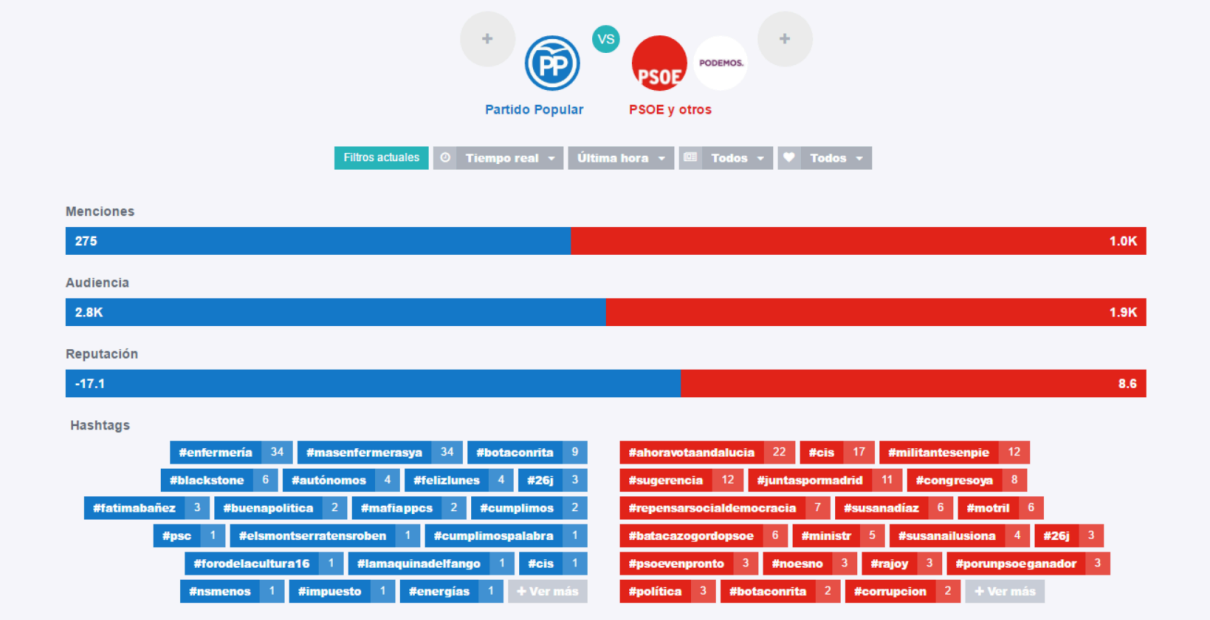

Figura 5: Modo «versus» para comparar pares de entidades

aparezcan siempre juntos. Por eso hemos decidido utilizar una aproximación basada en *skipgrams*, en la que los términos no tienen que aparecer juntos, sino que se permite que haya una cierta distancia entre ellos. Más específicamente, en un *k-skip-n-gram* se obtienen n-gramas en los que se permite una distancia máxima entre términos de k.

4.2 Puntuación de los términos

Nuestro lexicón consiste en una lista de skipgrams, donde cada uno tiene asociados diferentes valores de polaridad, indicando cómo es de fuerte la relación entre ese término y cada polaridad. También calcularemos una puntuación general para ver cómo es de importante (o frecuente) ese término en el corpus. Para construir este lexicón, necesitamos un corpus donde cada texto esté anotado con su polaridad. El peso de los skipgrams dependerá del número de veces que el skipgram aparece en textos de cada polaridad, y la distancia entre los términos de esas apariciones.

La ecuación 1 muestra la forma de calcular la puntuación general, donde T representa el conjunto de textos del corpus, t es un texto del conjunto T, $o_{s,t}$ representa una ocurrencia del skipgram s en el texto t, y *skip* es una función que nos indica la distancia entre los términos de $o_{s,t}$.

$$score(s) = \sum_{t \in T} \sum_{o_{s,t} \in t} \frac{1}{skip(o_{s,t}) + 1} \quad (1)$$

En la ecuación 2 se calcula la relación entre cada término y cada polaridad. Aquí, p representa una polaridad y T_p es el conjunto de textos del corpus etiquetados con esa polaridad p.

$$score(s, p) = \sum_{t \in T_p} \sum_{o_{s,t} \in t} \frac{1}{skip(o_{s,t}) + 1} \quad (2)$$

4.3 Aprendizaje supervisado

Utilizamos técnicas de aprendizaje automático para clasificar la polaridad de nuevos textos. Cada uno de los mensajes del corpus se utilizan como *instancias* de entrenamiento, y las polaridades etiquetadas se utilizan como *categorías*. Sin embargo, al contrario que las mayoría de aproximaciones, empleamos también las polaridades como *características* del modelo. El peso para estas características para cada texto se calcula según se especifica en la ecuación 3, donde S_t es el conjunto de skipgrams encontrados en el texto t.

Para construir nuestro modelo hemos elegido las máquinas de soporte vectorial, y más específicamente, la implementación por defecto de *LibSVM*[6].

5 Medición de la reputación

A partir de la cantidad de opiniones positivas, neutrales y negativas detectadas, y la

[6] www.csie.ntu.edu.tw/ cjlin/libsvm/

$$weight(p,t) = \sum_{o_{s,t} \in S_t} \left(\frac{1}{k(o_{s,t})+1} \cdot \frac{score(s,p)}{score(s,p)+1} \cdot \frac{score(s,p)}{score(s)}\right) \quad (3)$$

cantidad de gente a la que han llegado estas opiniones, podemos calcular un valor de reputación para cada entidad en un período de tiempo concreto. Esta puntuación es un valor numérico, acotado dentro del intervalo $[-1,+1]$, donde -1 sería la peor valoración y $+1$ sería la mejor valoración dada por el sistema, que se calcula utilizando la ecuación 7.

$$pos = \sum_{p \in P_{e,t,+}} a_p \quad (4)$$

$$neg = \sum_{p \in P_{e,t,-}} a_p \quad (5)$$

$$neu = \sum_{p \in P_{e,t,0}} a_p \quad (6)$$

$$v_{e,t} = \frac{2 \cdot pos + neu - 2 \cdot neg}{2 \cdot (pos + neu + neg) + d_t} \quad (7)$$

Donde e es la entidad que estamos valorando; t es un período de tiempo determinado; $P_{e,t,+}$, $P_{e,t,0}$ y $P_{e,t,-}$ representan los conjuntos de publicaciones que contienen una mención a la entidad e en el periodo de tiempo t con una polaridad positiva, neutral y negativa respectivamente; a_p es la audiencia o número de usuarios a los que ha llegado la publicación p; y d_t es la duración en milisegundos del periodo t..

La suma de la duración del período en el denominador de la ecuación (d_t) es una forma de dar una mayor valoración a las publicaciones que hayan llegado a más seguidores. Por ejemplo, si una entidad tiene menciones negativas que ha llegado a 100 personas en un minuto, su valoración sería de $-3 \cdot 100/(3 \cdot 100 + 60000) = -0,005$. Sin embargo, si las publicaciones ha llegado a 10.000 personas, la valoración sería de $-3 \cdot 10000/(3 \cdot 10000 + 60000) = -0,333$.

Cabe destacar que en nuestro sistema consideramos las menciones neutrales ($p \in P_{e,t,0}$) como algo positivo, ya que el hecho de que una entidad sea mencionada en las redes sociales aumenta su valoración. De esta forma, si todas las valoraciones son neutrales, la reputación será mayor que cero, mientras que si no hay ninguna mención sería exactamente cero.

6 Evaluación

La evaluación del sistema se he enfocado en dos entornos diferentes. Por un lado se evaluó la eficacia del sistema a la hora de determinar si la polaridad detectada era la correcta. Las pruebas se realizaron a través de los datos de la tarea abierta *Sentiment Analysis in Twitter* del workshop *TASS 2015* (Villena-Román et al., 2015). Los resultados obtenidos en la misma se pueden observar en la Tabla 1, comparándolos con los resultados obtenidos con bigramas ($n = 2$) y trigramas ($n = 3$).

	Parámetros	F1
Bigramas	$n = 2$	0.636
2-skip-k-grams	$n = 2, k = 1$	0.642
	$n = 2, k = 2$	0.646
	$n = 2, k = 3$	**0.647**
	$n = 2, k = max$	**0.647**
Trigramas	$n = 3$	0.624
3-skip-k-grams	$n = 3, k = 1$	0.623
	$n = 3, k = 2$	0.630
	$n = 3, k = 3$	0.637
	$n = 3, k = max$	**0.639**

Tabla 1: Resultados de la evaluación

Se puede observar un beneficio en la utilización del modelo de skipgrams. La mejor puntuación se ha obtenido con $n = 2$ y $k = 3$ (o $k = max$) comparando con los resultados con bigramas, con una mejora del 1,7 %, y con $n = 3$ $k = max$ respecto a los resultados con trigramas, con una mejora del 2,4 %. Por lo tanto, se puede intuir que existen expresiones que identifican polaridad, cuyas palabras no aparecen explícitamente juntas en los textos, que el modelo de n-gramas no ha podido descubrir pero sí el modelo de skipgrams.

Por otro lado se realizó una prueba a nivel experiencia del usuario. Para ello se convocó a los responsables de turismo de las localidades de la provincia de Alicante. Se realizó un pequeño seminario de uso de la herramienta y se permitió a los diferentes usuarios utilizarla durante un tiempo. Entre las principales conclusiones de nuestro experimento podemos destacar las siguientes:

- El sistema era lo suficientemente intuiti-

vo y los datos que se mostraban eran de gran interés y utilidad para su trabajo.

- Los usuarios no entraron en valorar la calidad del clasificador de polaridad, sólo necesitaban una visión global de las opiniones de los usuarios.
- Había especial interés en los cambios bruscos de polaridad, y gracias a la lista de palabras y hashtags más repetidos, era posible adivinar cuál era la causa de dicho cambio.
- Su mayor preocupación era como afectaban determinados eventos o sucesos de forma positiva o negativa a un destino. Uno de los ejemplos más llamativos fue un accidente de autobús con numerosas víctimas mortales que se produjo en las inmediaciones de la ciudad de Valencia. Esto focalizó una gran cantidad de menciones con polaridad negativa a la ciudad e incluso a la festividad que tenía lugar en ese momento (Fallas de Valencia).

7 Conclusiones y trabajos futuros

Social Analytics es una herramienta que ya se está utilizando por el gobierno valenciano para monitorizar los avances de cada uno de los municipios que forman la comunidad en el apartado turístico. El sistema ha fomentado la competitividad y el deseo de mejorar las campañas, al mostrar información sobre la reputación de las localidades comparadas.

Como trabajo futuro incorporaremos más fuentes de información, algo que cada vez es más complicado, ya que las plataformas de las redes sociales son cada vez más conscientes del valor que tiene la información de la que disponen y se muestran cada vez menos predispuestas a permitir el acceso libre a la misma. También nos planteamos mejorar la evaluación de nuestra medida de reputación, comparando nuestros resultados con diferentes sondeos y encuestas realizadas en otros medios.

Bibliografía

Fernández, J., Y. Gutiérrez, J. M. Gómez, P. Martínez-Barco, A. Montoyo, y R. Muñoz. 2013. Sentiment Analysis of Spanish Tweets Using a Ranking Algorithm and Skipgrams. En *XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN 2013)*, páginas 133–142.

Hao, M., C. Rohrdantz, H. Janetzko, U. Dayal, D. A. Keim, L.-E. Haug, y M.-C. Hsu. 2011. Visual sentiment analysis on twitter data streams. En *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, páginas 277–278. IEEE.

Hecht, B., L. Hong, B. Suh, y E. H. Chi. 2011. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. En *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, páginas 237–246. ACM.

Hu, M. y B. Liu. 2004. Mining and summarizing customer reviews. En *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, páginas 168–177. ACM.

Marcus, A., M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, y R. C. Miller. 2011. Twitinfo: aggregating and visualizing microblogs for event exploration. En *Proceedings of the SIGCHI conference on Human factors in computing systems*, páginas 227–236. ACM.

Miller, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Peregrino, F. S., D. Tomás, y F. Llopis. 2013. Every move you make i'll be watching you: geographical focus detection on twitter. En *Proceedings of the 7th Workshop on Geographic Information Retrieval*, páginas 1–8. ACM.

Strapparava, C. y A. Valitutti. 2004. Wordnet affect: an affective extension of wordnet. En *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, volumen 4, páginas 1083–1086.

Villena-Román, J., J. García-Morera, M. A. García-Cumbreras, E. Martínez-Cámara, M. T. Martín-Valdivia, y L. A. Urena-López. 2015. Overview of tass 2015. En *TASS 2015 Workshop on Sentiment Analysis at SEPLN*, volumen 1397, páginas 13–21.

Wang, H., D. Can, A. Kazemzadeh, F. Bar, y S. Narayanan. 2012. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. En *Proceedings of the ACL 2012 System Demonstrations*, páginas 115–120. Association for Computational Linguistics.

Tesis

Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 151-154 recibido 29-11-2016 revisado 12-01-2017 aceptado 09-02-2017

Author Profiling en Social Media: Identificación de Edad, Sexo y Variedad del Lenguaje*

Author Profiling in Social Media: Age, Gender and Language Variety Identification

Francisco Manuel Rangel Pardo
Universitat Politècnica de València
Camino de Vera s/n. 46022 Valencia (Spain)
kico.rangel@gmail.com

Resumen: Tesis doctoral escrita por Francisco Manuel Rangel Pardo en la Universitat Politècnica de València, bajo la dirección del PhD. Paolo Rosso. La tesis fue defendida el 3 de junio de 2016 en la misma universidad, ante el tribunal compuesto por los doctores Núria Bel de la Universitat Pompeu Fabra, Raquel Martínez Unanue de la Universidad Nacional de Educación a Distancia (UNED) y Rafael Berlanga Llavorí de la Universitat Jaume I. La tesis fue calificada con la puntuación de Sobresaliente *Cum Laude*.
Palabras clave: Author profiling, identificación edad, identificación sexo, identificación variedad lenguaje, social media, big data, emograph, representación de baja dimensionalidad

Abstract: PhD thesis written by Francisco Manuel Rangel Pardo at the Universitat Politècnica de València, under the supervision of PhD. Paolo Rosso. The thesis was defended on June 3rd 2016, with the commitee formed by the doctors Núria Bel from Universitat Pompeu Fabra, Raquel Martínez Unanue from Universidad Nacional de Educación a Distancia (UNED) and Rafael Berlanga Llavorí from Universitat Jaume I. The thesis was graded with Excellent *Cum Laude*.
Keywords: Author profiling, age identification, gender identification, language variety identification, social media, big data, emograph, low-dimensionality representation

1 Introducción

La posibilidad de conocer rasgos de una persona a partir únicamente de los textos que escribe se ha convertido en un área de gran interés denominada *author profiling* (LaVanguardia, 17/10/2016). Ser capaz de inferir de un usuario su sexo, edad, idioma nativo o los rasgos de su personalidad, simplemente analizando sus textos, abre todo un abanico de posibilidades desde el punto de vista forense, de la seguridad o del marketing.

Además, la proliferación de los medios sociales, que favorece nuevos modelos de comunicación y relación humana, potencia este abanico de posibilidades hasta cotas nunca antes vistas. La idiosincrasia inherente a estos medios sociales hace de ellos un entorno de comunicación especial, donde la libertad de expresión, la informalidad y la generación espontánea de temáticas y tendencias propician el acercamiento a la realidad diaria de las personas en su uso de la lengua. Sin embargo, esa misma idiosincrasia hace que en muchas ocasiones la aplicación de técnicas lingüísticas de análisis no sea posible, o sea extremadamente costoso.

La mayoría de aproximaciones propuestas por los investigadores para abordar las diversas tareas de *author profiling*, se basan en la frecuencia de uso de determinadas características (e.g. categorías gramaticales, palabras vacías o signos de puntuación), o en modelos que emplean n-gramas. Nuestra hipótesis, especialmente cuando hablamos de medios sociales donde no hay censura y prima la libertad de expresión, es que los usuarios expresan sus emociones de manera diferente dependiendo de ciertos rasgos de su persona. Nuestro objetivo es profundizar en el modo en que los usuarios expresan dichas emociones en el marco de su discurso, no sólo tomando en consideración su frecuencia relativa de

* Este trabajo ha sido parcialmente financiado por Autoritas Consulting SA (http://www.autoritas.net)

ISSN 1135-5948

aparición, sino también su posición con y en relación con el resto de elementos del discurso, y analizar cómo puede esto ayudar a determinar su edad y sexo, independientemente del medio social y del idioma. Para hacerlo, hemos propuesto EmoGraph, una representación basada en grafos, debido a su capacidad para modelar y analizar estructuras complejas como el lenguaje, que debido a la idiosincrasia propia de los medios sociales, hace compleja la aplicación de técnicas elaboradas de análisis sintáctico.

Además, hemos querido investigar si la expresión de emociones permitiría diferenciar entre hablantes de diferentes variedades de una misma lengua, por ejemplo españoles, mexicanos o argentinos, o portugueses y brasileños. Nuestra hipótesis es que la variación entre lenguas se basa más en aspectos léxicos, y así lo hemos corroborado tras comparar EmoGraph con representaciones basadas en patrones, representaciones distribuidas y una representación que toma en consideración el vocabulario completo, pero reduciendo su dimensionalidad a únicamente 6 características por clase y que se erige idónea para su aplicación en entornos *big data* como los medios sociales.

Podemos resumir los objetivos de la tesis en los siguientes:

1. Proponer una representación que permita:
   a) modelar la estructura del discurso y la expresión de las emociones en el mismo, tomando en consideración no sólo su frecuencia de aparición sino su posición con y en relación al resto de elementos del discurso;
   b) verificar la hipótesis de que el modo en que el usuario articula su discurso y expresa en él sus emociones, sirve para determinar su edad y sexo;
   c) comprobar la independencia de la hipótesis con respecto al medio social; y
   d) con respecto al idioma.
2. Investigar si la expresión de las emociones es un rasgo diferenciador entre lenguas similares o variedades de una misma lengua, o si por el contrario, variaciones léxicas aportan más información a la tarea;
   a) comprobar la adecuación de las representaciones distribuidas a la tarea;
   b) proponer una representación que reduzca la dimensionalidad frente a las comunmente utilizadas basadas en n-gramas.
3. Crear los recursos necesarios para investigar las cuestiones planteadas y un marco de evaluación común que permita comparar las propuestas de diferentes investigadores construyendo así un estado del arte homogéneo, comparable y reproducible.

2 Estructura

La tesis se ha organizado en torno a 8 capítulos y 2 apéndices. En ellos se trata de responder a las preguntas motivadas en el apartado anterior. Concretamente, la estructura de capítulos y lo que en ellos se trata, es la siguiente:

1. Introducción. En este capítulo introducimos la oportunidad que brindan los nuevos medios sociales y la necesidad de ser capaces de obtener información sobre las personas que participan en ellos, introduciendo así el concepto de *author profiling*.

2. Identificación de edad y sexo. En este capítulo efectuamos una revisión exhaustiva al estado del arte en identificación de edad y sexo, describiendo las representaciones propuestas, los corpus disponibles, las medidas de evaluación utilizadas y los resultados alcanzados. Además, sobre la base de las teorías psicolingüísticas actuales, realizamos un estudio estadístico relativo al uso de las categorías gramaticales por medio social y por sexo.

3. Author profiling en el PAN. En este capítulo realizamos una descripción detallada de los tres años en los que organizamos la tarea de identificación de edad y sexo en el PAN[1]. La organización del PAN ha propiciado la creación de recursos como corpus etiquetados con edad y sexo en idiomas diferentes al inglés, la definición de un marco común de evaluación y la generación de un estado del arte consistente, reproducible y útil para la comparación (Rangel et al., 2013; Rangel et al., 2014; Rangel et al., 2015).

[1] http://www.pan.webis.de

4. Identificación de emociones en medios sociales. En este capítulo abordamos la tarea de identificación de emociones en medios sociales e investigamos su relación con el *author profiling*. Nuestra hipótesis central es que la expresión de las emociones tiene una fuerte correlación con nuestro sexo y edad, algo que en el estado del arte no ha sido abordado. Comenzamos así con una revisión del estado del arte en procesamiento afectivo, desde la perspectiva de la generación de recursos y desde la perspectiva de la identificación automática de emociones en texto, para posteriormente analizar la utilidad de la expresión de las emociones para la identificación de tendencias, o su relación con la ironía, la edad y el sexo. En este capítulo presentamos nuestra investigación en identificación de edad y sexo a partir del mismo conjunto de características que utilizamos para la identificación de las emociones, sentando las bases de la hipótesis central de esta tesis.

5. EmoGraph: Una aproximación basada en grafos. En este capítulo investigamos con mayor profundidad cómo el modo en que los usuarios expresan las emociones sirve para conocer su edad y su sexo. Para ello, tratamos de modelar cómo los usuarios estructuran su discurso y cómo las emociones se enmarcan en el mismo, utilizando grafos debido a su potencia para representar y analizar estructuras complejas, como en este caso el lenguaje. Tras una revisión del estado del arte en uso de grafos para diversas tareas de procesamiento del lenguaje natural, decidimos aprovechar los grafos para extraer el conocimiento relacional entre las partes del discurso y las emociones, y obtener un esquema de asignación de pesos para el aprendizaje automático de los modelos (Rangel y Rosso, 2016). Para finalizar el capítulo investigamos en la robustez del método ante diversos medios sociales e idiomas.

6. Identificación del lenguaje nativo y de las variedades del lenguaje. En este capítulo se proporciona una visión detallada del estado del arte relativo a dos tareas relacionadas: la identificación del idioma nativo de un usuario que escribe en una segunda lengua, y la discriminación entre variedades de una misma lengua. El objetivo del capítulo es presentar el transfondo de una tarea que tiene la doble vertiente de la clasificación de textos y el *author profiling*, y sobre la que deseamos contrastar una segunda hipótesis: la variación entre lenguas similares o variedades de una misma lengua, se debe más a cambios léxicos que al modo en que sus usuarios expresan las emociones.

7. Aproximaciones para la identificación de variedades del lenguaje. En este capítulo investigamos la adecuación de la representación propuesta para identificar la edad y el sexo a partir del modo en que los usuarios articulan su discurso y expresan las emociones a la tarea de discriminar entre usuarios que hablan variedades de una misma lengua, o lenguas muy similares. Así mismo proponemos representaciones alternativas (Rangel, Rosso, y Franco-Salvador, 2016) que permiten contrastar nuestra hipótesis de que en este caso, el léxico tiene un componente discriminativo mayor.

8. Conclusiones y trabajo futuro. En este capítulo presentamos las conclusiones al trabajo que hemos llevado a cabo en el marco de nuestro doctorado, subrayando los principales descrubrimientos que soportan nuestras hipótesis, las principales contribuciones al estado del arte y marcando las directrices de trabajos futuros que pueden derivarse del mismo.

Apéndice I. Author profiling en PAN 2014. En este apéndice se muestran las tablas de significación estadística en una comparación pareada de los sistemas participantes de la tarea del PAN 2014, además de las distancias entre la edad predecida y la edad real de los autores.

Apéndice II. Author profiling en PAN 2015. En este apéndice se muestran las tablas de significación estadística en una comparación pareada de los sistemas participantes en la tarea del PAN 2015, así como una comparativa de resultados entre idioma.

3 Contribuciones

Se pueden destacar las siguientes contribuciones:

1. Hemos propuesto la representación EmoGraph para modelar el estilo discursivo y la expresión de las emociones en textos, y la hemos aplicado a la identificación de edad y sexo. Además, hemos verificado su aplicabilidad y robustez a diferentes medios sociales e idiomas.

2. Hemos investigado la aplicabilidad de la representación EmoGraph en la tarea de identificación de variedades de una misma lengua, comprobando que la expresión de emociones y el estilo discursivo no varía de modo discriminativo. Para verificar nuestra hipótesis de que las variaciones se producen principalmente a nivel léxico, hemos analizado varias representaciones: una basada en patrones (IG-WP), dos basadas en representaciones distribuidas sobre el conocido modelo de Skip-gramas continuos, y la representación de baja dimensionalidad (LDR) que hemos propuesto y que permite trabajar de manera eficiente en entornos *big data*.

3. Hemos creado los recursos necesarios para llevar a cabo la investigación, concretamente:

   a) Con respecto a la identificación de edad y sexo, hemos colaborado en la organización y creación de un marco de evaluación en la tarea de identificación de edad y sexo del PAN en el CLEF, lo que ha permitido crear un conjunto de corpus recopilados de diferentes medios sociales (Twitter, blogs, revisiones online, redes sociales) y en diferentes idiomas (inglés, español, holandés e italiano), etiquetados con edad y sexo.

   b) Con respecto a la identificación de emociones en medios sociales, y concretamente en Twitter, hemos compilado el corpus Barcenas con tuits tratando un caso de corrupción política ocurrido en España entre el 9 de julio y el 2 de octubre de 2013, con un total de 4.397.023 tuits en español.

   c) Con respecto a la identificación de emociones en medios sociales y su relación con el sexo, así como con el uso de la ironía, hemos generado el corpus EmIroGeFB con comentarios de Facebook anotados con las seis emociones básicas de Ekman, la presencia/ausencia de ironía, y el sexo de los autores de los comentarios. El corpus se enmarca dentro de tres temáticas (política, fútbol, famosos) y consta de 1.200 comentarios en español.

   d) Por último, con respecto a la identificación de la variedad de lenguaje, hemos construido el corpus HispaBlogs con posts escritos en blogs personales en cinco variedades del español: Argentina, Chile, España, México y Panamá. El corpus consta de dos particiones de 2.400 y 1.000 autores por partición.

Bibliografía

LaVanguardia. 17/10/2016. Los textos escritos delatan el sexo. http://www.lavanguardia.com/cultura/20161017/411054860677/linguistica-estudio-determina-sexo-autor-articulo-upf.html.

Rangel, F. y P. Rosso. 2016. On the impact of emotions on author profiling. *Information processing & management*, 52(1):73–92.

Rangel, F., P. Rosso, I. Chugur, M. Potthast, M. Trenkmann, B. Stein, B. Verhoeven, y W. Daelemans. 2014. Overview of the 2nd author profiling task at pan 2014. En L. Cappellato N. Ferro M. Halvey, y W. Kraaij, editores, *CLEF 2014 labs and workshops, notebook papers. CEUR-WS.org*, volumen 1180, páginas 898–927.

Rangel, F., P. Rosso, y M. Franco-Salvador. 2016. A low dimensionality representation for language variety identification. En *17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing*. Springer-Verlag, LNCS.

Rangel, F., P. Rosso, M. Moshe Koppel, E. Stamatatos, y G. Inches. 2013. Overview of the author profiling task at pan 2013. En P. Forner R. Navigli, y D. Tufis, editores, *CLEF 2013 labs and workshops, notebook papers. CEUR-WS.org*, volumen 1179, páginas 352–365.

Rangel, F., P. Rosso, M. Potthast, B. Stein, y W. Daelemans. 2015. Overview of the 3rd author profiling task at pan 2015. En L. Cappellato N. Ferro G. Jones, y E. San Juan, editores, *CLEF 2015 labs and workshops, notebook papers. CEUR Workshop Proceedings. CEUR-WS.org*, volumen 1391.

Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 155-158 recibido 29-11-2016 revisado 11-01-2017 aceptado 02-02-2017

Análisis de la complejidad y simplificación automática de textos. El análisis de las estructuras complejas en euskera*

Readability Assessment and Automatic Text Simplification. The Analysis of Basque Complex Structures

Itziar Gonzalez-Dios
Grupo Ixa, Universidad del País Vasco (UPV/EHU)
Manuel Lardizabal 1, 20018 Donostia
itziar.gonzalezd@ehu.eus

Resumen: Tesis doctoral titulada "Euskarazko egitura sintaktiko konplexuen analisirako eta testuen sinplifikazio automatikorako proposamena / Readability Assessment and Automatic Text Simplification. The Analysis of Basque Complex Structures", defendida por Itziar Gonzalez Dios en la Universidad del País Vasco (UPV/EHU) y elaborada bajo la dirección de las doctoras Arantza Díaz de Ilarraza (Departamento de Lenguajes y Sistemas Informáticos) y María Jesús Aranzabe (Departamento de Lengua Vasca y Comunicación). La defensa tuvo lugar el 23 de junio de 2016 ante el tribunal formado por los doctores Kepa Sarasola (Presidente, Universidad del País Vasco (UPV/EHU)), Ricardo Etxepare (Secretario, Centre National de la Recherche Scientifique-IKER) y Giulia Venturi (Vocal, Instituto di Linguistica Computazionale Antonio Zampolli - Consiglio Nazionale delle Ricerche) y la tesis obtuvo la mención Cum Laude y Doctor Internacional.
Palabras clave: análisis de la complejidad o lecturabilidad, simplificación automática de textos, sintaxis, euskera

Abstract: Ph.D. thesis entitled "Euskarazko egitura sintaktiko konplexuen analisirako eta testuen sinplifikazio automatikorako proposamena / Readability Assessment and Automatic Text Simplification. The Analysis of Basque Complex Structures" written by Itziar Gonzalez Dios at the University of Basque Country (UPV/EHU) under the supervision of the Ph.D. Arantza Díaz de Ilarraza (Languages and Computer Systems Department) and Ph.D. María Jesús Aranzabe (Basque Language and Communication Department). The viva voce was held on the 23rd June 2016 and the members of the commission were the Ph.D. Kepa Sarasola (President, University of Basque Country (UPV/EHU)), Ph.D. Ricardo Etxepare (Secretary, Centre National de la Recherche Scientifique-IKER) and Ph.D. Giulia Venturi (Vocal, Instituto di Linguistica Computazionale Antonio Zampolli - Consiglio Nazionale delle Ricerche) and the thesis obtained the mention Cum Laude and International Doctor.
Keywords: readability assessment, automatic text simplification, syntax, Basque

1 Introducción de la tesis

Esta tesis doctoral se ha realizado en el grupo Ixa[1] de la Universidad del País Vasco (UPV/EHU) y se han tratado dos líneas de investigación: el análisis de la complejidad de textos o lecturabilidad (*readability assessment* en inglés) y la simplificación automática de textos (*automatic text simplification* en inglés). Concretamente, se ha analizado la complejidad sintáctica del euskera con el objetivo de diseñar un sistema de simplificación automática de textos.

Millones de textos se producen a diario en nuestra sociedad, pero estos textos no son accesibles para todos por diversos motivos: por ejemplo, algunas estructuras son complejas para las personas con enfermedades cognitivas o con alteraciones en el lenguaje y, también para las personas que aprenden lenguas extranjeras. A este último colectivo además

* Este tesis doctoral ha sido realizada con una beca predoctoral del Gobierno Vasco. Referencia: BF1-2011-392

[1] `http://ixa.eus/Ixa`

ISSN 1135-5948

se le añade como problema el desconocimiento del vocabulario. Pero la complejidad de los textos no es solo un problema que afecta al ser humano, sino que también afecta a las aplicaciones avanzadas del Procesamiento del Lenguaje Natural (PLN). Estas aplicaciones no procesan efectivamente oraciones largas y complejas y, por ello, mediante la simplificación automática se pretende mejorar su rendimiento.

Esta tesis tiene dos partes: 1) el análisis de las estructuras sintácticas complejas del euskera para realizar propuestas de simplificación desde un punto de vista lingüístico y 2) la producción de los recursos lingüísticos necesarios para implementar el análisis de la complejidad y la simplificación automática de textos de manera general desde una perspectiva computacional.

2 Estructura de la tesis

Con respecto a la primera parte, se dedica el primer capítulo a presentar la motivación, los objetivos y los objetos de estudio; en el segundo capítulo, se realiza un resumen de los sistemas de análisis de lecturabilidad y de los sistemas de simplificación automática explicando los diferentes tipos de simplificación, arquitecturas, técnicas y métodos de evaluación además de los recursos necesarios para ambas tareas. En el tercer capítulo se presenta el análisis lingüístico de las estructuras sintácticas complejas del euskera y se explican las propuestas para simplificar las oraciones coordinadas, yuxtapuestas, las compuestas subordinadas (sustantivas, adjetivas o relativas y adverbiales) y las aposiciones. Además, se estudian las estructuras parentéticas que contienen información biográfica que dan lugar al sistema *Biografix* que se explica en el sexto capítulo.

El cuarto capítulo se presenta como un nexo entre el análisis lingüístico que constituye la primera parte y el más computacional de la segunda parte y se expone cómo automatizar el análisis lingüístico. Para ello, se presentan las decisiones tomadas en referencia: i) a los tipos y niveles de simplificación; y ii) al algoritmo para la selección del tipo de simplificación atendiendo al usuario. Además se presentan las herramientas básicas de análisis lingüístico automático necesarias para llevar a cabo dicho proceso (la cadena de análisis de análisis desarrollada en el grupo Ixa y las herramientas básicas *Mugak* (Aranzabe, Díaz de Ilarraza y Gonzalez-Dios, 2013) y *Aposizioak* (Gonzalez-Dios et al., 2013)) desarrolladas en esta tesis.

La segunda parte está compuesta por los capítulos quinto y sexto. En el quinto capítulo se detalla el sistema de lecturabilidad *Erre-Xail* (Gonzalez-Dios et al., 2014) que discierne si un texto es simple o complejo teniendo en cuenta 96 ratios con información lingüística y técnicas de aprendizaje automático. Este sistema se utiliza como preproceso para saber si el texto de entrada debe ser simplificado o no. Si el texto es complejo, será simplificado por el sistema *EuTS* (*Euskarazko Testuen Sinplifikatzailea* [Simplificador de textos en euskera]) (Aranzabe, Díaz de Ilarraza y Gonzalez-Dios, 2012) cuya propuesta y módulos se presentan en el sexto capítulo. *EuTS* aplica las reglas lingüísticas para la simplificación sintáctica presentadas en el tercer capítulo. Como caso de estudio, se describe *Biografix* (Gonzalez-Dios, Aranzabe y Díaz de Ilarraza, 2014), que siguiendo las operaciones de *EuTS*, realiza la simplificación sintáctica de las estructuras parentéticas en euskera, castellano, alemán, francés, italiano, gallego y catalán.

Para evaluar nuestra propuesta de análisis de complejidad y simplificación, en el septimo capítulo se presenta el corpus de los textos simplificados en euskera: *Euskarazko Testu Sinplifikatuen Corpusa (ETSC)/ Corpus of Basque Simplified Texts (CBST)*. Este corpus contiene textos simplificados según las aproximaciones estructural e intuitiva y se ha anotado siguendo el esquema de anotación definido en esta tesis.

Finalmente, en el octavo capítulo se presentan las contribuciones de la tesis y el trabajo futuro. En los apéndices de la tesis, se recogen la lista de las estructuras adverbiales analizadas, las reglas de simplificación y la lista de las operaciones encontradas en ambas aproximaciones con el objeto de que sean obligatorias al aumentar el corpus o en futuras simplificaciones.

3 Contribuciones más relevantes

A continuación se presentan las contribuciones de la tesis según las dos líneas de investigación exploradas y los recursos generados.

3.1 Análisis de la complejidad y lecturabilidad

En lo referente al análisis de la complejidad, se han determinado las estructuras sintácticas consideradas complejas en euskera basándonos en los trabajos realizados en otras lenguas y en nuestro análisis lingüístico. Dichas estructuras son las oraciones coordinadas y yuxtapuestas, las compuestas subordinadas, las aposiciones y las estructuras parentéticas que contienen información biográfica. Para que las oraciones que contienen estas estructuras sean simplificadas, se ha determinado que deben tener una extensión mínima de dos sintagmas además del verbo.

A su vez, se ha implementado el sistema de análisis de lecturabilidad llamado *Erre-Xail* que determina si los textos son simples o complejos. Para ello, se han definido 96 ratios que se dividen en los grupos globales, lexicales, morfológicos, morfosintácticos, sintácticos y pragmáticos. En los experimentos realizados con la herramienta *Weka*[2], el mejor resultado de clasificación (93,50 % de precisión) se ha obtenido con la combinación de las características léxicales, morfológicas, morfosintácticas y sintácticas y con el clasificador SMO (máquinas de vectores de soporte).

3.2 Simplificación automática de textos

En lo referente a la simplificación automática de textos, se ha realizado el esquema general del sistema *EuTS*. Este sistema está basado en reglas lingüísticas, realiza dos tipos de simplificación a nivel sintáctico (sustitución sintáctica y la simplificación sintáctica) y adapta los textos a tres niveles diferentes (simplificación sintáctica superficial, simplificación natural y simplificación absoluta) adecuados a los niveles de conocimiento del euskera y a las necesidades de las aplicaciones del PLN.

En la sustitución sintáctica, las estructuras adverbiales de menor frecuencia se sustituyen por equivalentes de mayor frecuencia (Gonzalez-Dios, Aranzabe y Díaz de Ilarraza, 2015). De este modo, se consiguen los textos del nivel llamado simplificación sintáctica superficial, que son más accesibles pero que mantienen la estructura general. Este tipo de simplificación está completamente implementado y se ha evaluado cuantitativamente y cualitativamente. Cuantitativamente, el 79,63 % de las sustituciones realizadas han sido correctas, y de ellas en el 88,64 % de los casos las frases generadas has sido gramaticalmente correctas. Cualitativamente, el 75,00 % de las oraciones has resultado más fáciles de comprender para los usuarios de dicho nivel.

En la simplificación sintáctica se han establecido las operaciones y se ha recopilado toda la información lingüística necesaria para su implementación. Dichas operaciones son división, reconstrución, reordenación y correción. Mediante su aplicación se obtienen frases más cortas y, a su vez, la estructura sintáctica original desaparece. Según el nivel del usuario, los textos se adaptan a los niveles de simplificación natural o simplificación absoluta.

Asimismo, se ha implementado *Biografix* que prueba las operaciones y reglas definidas en nuestro estudio con las estructuras parentéticas que contienen información biográfica en 8 idiomas. De este modo, se ha comprobado que las reglas definidas para el euskera pueden ser adaptadas y reutilizadas en otras lenguas.

3.3 Recursos

Tres son los recursos más importantes creados en esta tesis: 1) el corpus de los textos simplificados en euskera y las herramientas básicas 2) *Mugak* y 3) *Aposizioak*.

El corpus de los textos simplificados recoge dos aproximaciones de simplificación de textos: la estructural y la intuitiva. Esto significa que cada frase original del corpus ha sido simplificada según ambas aproximaciones. Para simplificar los textos según la aproximación estructural, una traductora jurada ha seguido directrices de lectura fácil y para la aproximación intuitiva, una profesora de euskera se ha basado en su experiencia e intuición. Para realizar el análisis de este corpus se ha creado un esquema de anotación que se compone de las siguientes ocho macrooperaciones: eliminación, fusión, división, transformación, inserción, reordenación, ninguna operación y otras. Mediante este esquema, se han analizado las operaciones realizadas en cada una de las aproximaciones, y se ha creado una lista con las operaciones (distintas realizaciones de las macrooperaciones)

[2]Hall M., Frank E., Holes G., Pfahringer B., Reutemann P., y Witten I.H. The WEKA Data Mining Software: an Update. ACM SIGKDD Explorations Newsletter, 11(1):10-18, 2009.

más comunes como, por ejemplo, dividir oraciones coordinadas o recuperar los elementos elididos.

En cuanto a las herramientas básicas que se han implementado, cabe mencionar que *Mugak* detecta los límites de las oraciones basándose en información lingüística (Aranzabe, Díaz de Ilarraza y Gonzalez-Dios, 2013) y que *Aposizioak* (Gonzalez-Dios et al., 2013) detecta y clasifica las aposiciones y sus componentes. Estas dos herramientas son indispensables para realizar la operación de división.

Bibliografía

Aranzabe, M. J., A. Díaz de Ilarraza, y I. Gonzalez-Dios. 2012. First Approach to Automatic Text Simplification in Basque. En L. Rello y H. Saggion, editores, *Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) workshop (LREC 2012)*, páginas 1–8.

Aranzabe, M. J., A. Díaz de Ilarraza, y I. Gonzalez-Dios. 2013. Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque. *Procesamiento del Lenguaje Natural*, 50:61–68.

Gonzalez-Dios, I. 2014a. Euskarazko testuak errazten: euskal testuen sinplifikazio automatikoa [Making Basque Texts Easier: Automatic Simplification of Basque Texts]. En I. Aduriz y R. Urizar, editores, *Euskal Hizkuntzalaritzaren egungo zenbait ikerlerro. Hizkuntzalari euskaldunen I. topaketa*. Udako Euskal Unibertsitatea, páginas 135–149.

Gonzalez-Dios, I. 2014b. Simplificación automática de textos en euskera [Automatic Simplification of Basque Texts]. En L. A. Ureña López J. A. Troyano Jiménez F. J. Ortega Rodríguez, y E. Martínez Cámara, editores, *Actas de las V Jornadas TIMM, Cazalla de la Sierra, España, 12-JUN-2014, publicadas en http://ceur-ws.org*, páginas 45–50.

Gonzalez-Dios, I., M. J. Aranzabe, A. D. de Ilarraza, y A. Soraluze. 2013. Detecting Apposition for Text Simplification in Basque. *Computational Linguistics and Intelligent Text Processing*. Springer, páginas 513–524.

Gonzalez-Dios, I., M. J. Aranzabe, y A. Díaz de Ilarraza. 2013. Testuen sinplifikazio automatikoa: arloaren egungo egoera [Automatic Text Simplification: State of Art]. *Linguamática*, 5(2):43–63.

Gonzalez-Dios, I., M. J. Aranzabe, y A. Díaz de Ilarraza. 2014. Making Biographical Data in Wikipedia Readable: A Pattern-based Multilingual Approach. *Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA 2014)*, páginas 11–20, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Gonzalez-Dios, I., M. J. Aranzabe, y A. Díaz de Ilarraza. 2015a. Perpaus adberbialen agerpena, maiztasuna eta kokapena EPEC-DEP corpusean [Presence, frequency and Position of Basque Adberbial Clauses in The BDP corpus]. Informe técnico, University of the Basque Country (UPV/EHU) UPV/EHU/LSI/TR 02-2015.

Gonzalez-Dios, I., M. J. Aranzabe, y A. Díaz de Ilarraza. 2015b. Simplifying Basque Texts: the Shallow Syntactic Substitution Simplification. *Proceedings of the 7th Language & Technology Conference.*, páginas 450–454.

Gonzalez-Dios, I., M. J. Aranzabe, A. Díaz de Ilarraza, y H. Salaberri. 2014. Simple or Complex? Assessing the Readability of Basque Texts. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, páginas 334–344, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 159-162 recibido 01-12-2016 revisado 11-01-2017 aceptado 06-02-2017

Detección de Opinion Spam usando PU-Learning

Opinion Spam detection using PU-Learning

Donato Hernández Fusilier
Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València
Camino de Vera s/n 46022, Valencia, España
doherfu@doctor.upv.es

Resumen: Tesis doctoral realizada por Donato Hernández Fusilier en la Universitat Politècnica de València, dirigida por los Doctores Paolo Rosso (Universitat Politècnica de València, España, Manuel Montes-y-Gómez (Instituto Nacional de Astrofísica, Óptica y Electrónica, México) y Rafael Guzmán (Universidad de Guanajuato, México). La defensa se efectuó el 20 de enero de 2016 en Valencia. El tribunal estuvo conformado por la Dra. Raquel Martinez Unanue de la Universidad Nacional de Educación a Distancia, Madrid como vocal, por el Dr. Carlos David Martinez Hinajeros de la Universitat Politècnica de València como secretario y por el Dr. Rafael Berlanga LLavori de la Universitat Jaume I, Castelló como presidente. La tesis obtuvo una calificación de Sobresaliente.
Palabras clave: Análisis de Opiniones, aprendizaje supervisado, aprendizaje no supervisado, clasificadores, PU-Learning.

Abstract: Doctoral thesis written by Donato Hernández Fusilier at the Universitat Politècnica of València, directed by Ph.D. Paolo Rosso (Universitat Politècnica of València, Spain), Ph.D. Manuel Montes-y-Gómez (National Institute of Astrophisics, Optics and Electronics, México) and Ph.D. Rafael Guzmán (University of Guanajuato, México). The defense took place on January 20, 2016 in Valencia. The doctoral committee was integrated by the following doctors: Ph.D. Raquel Martinez Unanue of National Distance Learning University, Madrid as panel member, Ph.D. Carlos David Martinez Hinajeros de la Universitat Politècnica of València as secretary and by Ph.D. Rafael Berlanga LLavori of the Universitat Jaume I, Castelló as president. The thesis was graded as Outstanding.
Keywords: Analysis of Opinions, supervised learning, unsupervised learning, classifiers, PU-Learning.

1 Introducción

Con el uso general de las tecnologías de información es cada vez más común que los usuarios de servicios y los consumidores de productos escriban sus opiniones a favor o en contra de los productos o servicios que adquirieron. Estas referencias escritas comúnmente en foros, blogs y en general en las redes sociales, sirven de ayuda a otros consumidores que desean adquirir algunos productos o servicios similares. También sirven a los fabricantes o prestadores de servicios para identificar nuevas áreas de oportunidad por parte de los consumidores y les permite saber no solo la opinión sobre los mismos, sino además ver sus usos, costumbres, satisfacción, etcétera. Los consumidores utilizan las opiniones para recibir información sobre los productos, tales como calidad y utilidad, también son utilizadas para proporcionar datos sobre su propia experiencia con el producto a otros consumidores. Por otro lado, los fabricantes utilizan estos comentarios positivos o negativos, para identificar características que son importantes para los consumidores. Estas características son entonces incluidas en la comercialización y desarrollo de nuevos productos o servicios.

ISSN 1135-5948

Con este tipo de opiniones se presenta el problema del Opinión Spam que, en otras palabras, son aquellas opiniones falsas, escritas deliberadamente para promover o desacreditar un producto o servicio. Son opiniones escritas por personas que no han adquirido producto o servicio alguno, pero que fueron contratados para escribir opiniones engañosas. Estas opiniones falsas hacen creer a los posibles usuarios, que el producto o servicio es bueno o, según sea la causa para la cual fueron inducidos.

2 Objetivos

Desarrollar un método semi-supervisado, basado en el método de PU-Learning, para la detección de opiniones falsas, que considere tanto las características temáticas como estilísticas y que además sea robusto, adecuado y efectivo en escenarios reales que presentan una escasez de datos etiquetados.

Encontrar los atributos que consideren información tanto del contenido como del estilo de escritura de las opiniones.

Establecer el efecto de la polaridad de las opiniones, mediante el entrenamiento con una polaridad y la prueba con otra polaridad.

Probar el método propuesto en ambientes de dominios cruzados, donde se entrena con opiniones de un dominio y se prueba con opiniones de otro dominio diferente.

3 Organización

Este trabajo de tesis está organizado en seis capítulos: en el Capítulo 1 se describe la introducción al tema de la detección de opiniones falsas, los siguientes tres capítulos (Capitulo 2, Capítulo 3 y Capítulo 4), corresponden a las publicaciones realizadas sobre el nuevo método de detección de opiniones falsas basado en PU-Learning que se propone, a este nuevo método le llamamos PU-Learning*. En el Capítulo 5 se realiza la discusión de los resultados obtenidos y por último en el Capítulo 6 se presentan las conclusiones y el trabajo futuro.

Una breve síntesis del contenido de los capítulos presentados se muestra a continuación:

Capítulo 2: Se presenta el trabajo:"Using PU-Learning to detect deceptive Opinion Spam"publicado en el workshop WASSA 2013 (Hernández et al., 2013). En este artículo se utiliza el método de PU-Learning* por primera vez, evaluándolo en un conjunto de opiniones positivas (falsas y verdaderas) sobre hoteles situados en el centro de la ciudad de Chicago. En esta publicación se compara el nuevo método con otros que requieren un conjunto completamente etiquetado de opiniones para su funcionamiento y se obtienen valores de f-measure promedio de hasta 0.840.

Capítulo 3: Se presenta el artículo "Detecting positive and negative opinions using PU-Learning"publicado en la revista Information Processing & Management (Hernández et al., 2015a), donde se evalúa el método de PU-Learning* en un conjunto de opiniones que contiene tanto opiniones positivas como negativas de 20 hoteles del área del centro de Chicago. En este artículo se compara la precisión obtenida usando el método de PU-Learning* contra el método de PU-Learning tradicional, empleando diferentes tipos de atributos. También se realiza un análisis de significancia estadística entre los atributos empleados, para determinar la mejor combinación que nos lleva a obtener resultados con la precisión más alta.

Capítulo 4: Se presenta la publicación de la conferencia CICLing 2015 "Detection of opinion spam with character n-grams"(Hernández et al., 2015b), donde se evalúa y compara el método de PU-Learning* para detectar las opiniones falsas utilizando diferentes tipos de atributos, particularmente n-gramas de palabras y n-gramas de caracteres. En este trabajo se presenta también un análisis sobre la robustez de los n-gramas de caracteres en la clasificación con pocos datos de entrenamiento.

Capítulo 5: Se presenta una descripción completa de los corpus utilizados en los experimentos realizados para la evaluación del método de PU-Learning*, así como los resultados obtenidos en los experimentos diseñados para probar su eficacia. También se realizan pruebas de dominios cruzados, donde se entrena con un dominio diferente al dominio con el que se realiza la prueba.

Capítulo 6: Se presentan las conclusiones del desarrollo del trabajo de tesis, algunas ideas para desarrollar en el futuro y la lista de las publicaciones generadas.

4 Contribuciones

La principal contribución de este trabajo es haber diseñado el método llamado PU-Learning*, que permite detectar opiniones falsas partiendo de un conjunto pequeño de

instancias etiquetadas como opiniones falsas y otro conjunto más grande de instancias no-etiquetadas. Los resultados experimentales nos permiten concluir que sí es posible detectar de una manera más efectiva las opiniones falsas usando el método propuesto.

El método de PU-Learning* se puede describir de una manera breve por medio del algoritmo 1. En este algoritmo se puede apreciar como se van extrayendo del conjunto no etiquetado, aquellas instancias a las cuales el clasificador les asigna una etiqueta positiva. Al final de un número finito de iteraciones, solo permanecen en el conjunto de instancias no-etiquetadas, aquellas que pasaran a formar parte del conjunto de instancias negativas.

Algoritmo 1 PU-Learning* para la detección de Opinion Spam

$i \leftarrow 1$
$|W_0| \leftarrow |U_1|$
$|W_1| \leftarrow |U_1|$
while $|W_i| <= |W_{i-1}|$ **do**
  $C_i \leftarrow Generate\_Classifier(P, U_i)$
  $U_i^L \leftarrow C_i(U_i)$
  $W_i \leftarrow Extract\_Positives(U_i^L)$
  $U_{i+1} \leftarrow U_i - W_i$
  $i \leftarrow i + 1$
Return Classifier C_i

Mediante la aplicación del algoritmo 1 y usando los atributos apropiados, se obtiene como resultado la clase faltante, la de las opiniones verdaderas.

Adicionalmente los experimentos realizados nos permitieron formular las siguientes conclusiones respecto al método PU-Learning*:

- El método de PU-Learning* es más efectivo para la detección de opiniones falsas que el método PU-Learning tradicional; esto se debe en gran manera al uso de criterios más conservadores y exigentes para la selección de las opiniones (presumiblemente) falsas del conjunto no-etiquetado.
- El método de PU-Learning* es efectivo para la detección de opiniones falsas de ambas polaridades, sin embargo, los resultados obtenidos fueron mejores en la detección de opiniones falsas positivas. Este comportamiento del método propuesto coincide con trabajos previos, y se origina a partir de la mayor diversidad y mayor nivel de especificidad, de las opiniones falsas negativas.
- El método de PU-Learning* es una solución adecuada para la detección de opiniones falsas en escenarios de dominios cruzados, particularmente cuando los dominios fuente y objetivo son afines, es decir, cuando estos muestran una alta intersección en el vocabulario empleado.

Adicional al diseño del método PU-Learning*, una segunda contribución de este trabajo es la propuesta de una representación de los documentos apropiada para esta tarea, que incorpora tanto aspectos de su contenido como de su estilo. Esta representación fue realizada a través de los n-gramas de caracteres. Los resultados obtenidos con esta representación son superiores a los obtenidos con la representación tradicional de bolsa de palabras, permitiendo concluir que la información estilística es también importante para la detección de opiniones falsas.

Los experimentos realizados con ambas representaciones, también indican que para modelar adecuadamente el estilo de escritura de las opiniones falsas y verdaderas es necesario disponer de grandes conjuntos de entrenamiento; usando conjuntos pequeños las diferencias en los resultados de ambas representaciones fueron estadística-mente no significativas.

Bibliografía

Hernández, D., R. Guzmán, M. M. y Gómez, y P. Rosso. 2013. Using PU-learning to detect deceptive opinion spam. En *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, volumen 1606 de *Association for Computational Linguistics*, páginas 38–45, Atlanta, Georgia. Association for Computational Linguistics.

Hernández, D., M. M. y Gómez, P. Rosso, y R. Guzmán. 2015a. Detecting positive and negative deceptive opinions using PU-learning. *Information Processing & Management*, 51(4):433–443.

Hernández, D., M. M. y Gómez, P. Rosso, y R. Guzmán. 2015b. Detection of opinion spam with character n-grams. En A. Gelbukh, editor, *Proc. of 16th International*

conference on Intelligent Text Processing and Computational Linguistics (CICLing-2015), Lecture Notes in Artificial Intelligence, páginas 285–294, Cairo, Egypt, April. Springer-Verlag. Vol. 9042 part II.

Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 163-166 recibido 01-12-2016 revisado 11-01-2017 aceptado 06-02-2017

Detección de reutilización de código fuente monolingüe y translingüe

Crosslingual and monolingual source code re-use detection

Enrique Flores Sàez
Universitat Politècnica de València
Camí de Vera, s/n, Valencia
enflosae@gmail.com

Resumen: Tesis doctoral en Informática realizada por Enrique Flores Sàez en la Universitat Politècnica de València bajo la dirección de la Dra. Lidia Moreno Boronat y del Dr. Paolo Rosso. El acto de defensa tuvo lugar el lunes 30 de Mayo de 2016 ante el tribunal formado por los doctores Fidel Cacheda (Universidad de A Coruña), Juan Manuel Torres (Universidad de Avignon Murcia) y Greg Kondrak (Universidad de Alberta). Obtuvo mención internacional y calificación de Sobresaliente.
Palabras clave: Detección de reutilización en códigos fuente, detección de reutilización monolingüe, detección de reutilización translingüe, detección de plagio

Abstract: PhD Thesis in Computer Science written by Enrique Flores Sàez at the Universitat Politècnica de València under the supervision of PhD. Lidia Moreno Boronat and PhD. Paolo Rosso. The author was examined on 30th May 2016 by a committee formed by the PhD Fidel Cacheda (University of A Coruña), Juan Manuel Torres (University of Avignon) and Greg Kondrak (University of Alberta). The thesis obtained the grade of Excellent and received the international mention.
Keywords: Source code re-use detection, monolingual re-use detection, crosslingual re-use detection, plagiarism detection

1 Introducción

La detección automática de reutilización en códigos fuente consiste en determinar si un (fragmento de) código ha sido creado considerando otro como fuente. El plagio y las ramificaciones en proyectos software son dos ejemplos de tipos de reutilización en códigos fuente. Con la llegada de la Web y los medios electrónicos ha crecido enormemente la facilidad de acceso a códigos fuente para ser leídos, copiados o modificados. Esto supone una gran tentación para programadores que, con propósitos de reducir costes (temporales o económicos), deciden utilizar códigos fuente previamente depurados y probados. Este fenómeno ha causado que expertos en lenguajes de programación estudien el problema.

La gran cantidad de recursos disponibles en la Web hace imposible un análisis manual de códigos fuente sospechosos de haber sido reutilizados. Por ello, existe una necesidad urgente de desarrollar herramientas automáticas capaces de detectar con precisión los casos de reutilización. Basándose en técnicas del procesamiento del lenguaje natural y recuperación de información, las herramientas de detección automáticas de reutilización son capaces de realizar multitud de comparaciones de códigos fuente de forma eficiente.

Uno de los principales objetivos de esta tesis es abordar el problema de la reutilización de códigos fuente escritos en un lenguaje de programación como si se tratase de un texto escrito en un lenguaje natural. Por este motivo, en este trabajo de investigación se han estudiado los principales modelos aplicados a reutilización de textos en lenguaje natural, se han implementado y adaptado para aplicarse sobre códigos fuente. En esta tesis se describen los modelos implementados para la detección de reutilización en código fuente. Los modelos presentados en esta tesis siguen el siguiente esquema: Inicialmente, una colección de códigos fuente es tratada mediante un preproceso concreto. Tras este preproceso se obtiene un conjunto de características que representan a cada código fuente, y finalmente, se aplica una medida de similitud obteniéndose como resultado un listado de pares de códigos fuente junto al valor de similitud entre cada par.

ISSN 1135-5948

En esta tesis proponemos un conjunto de modelos que pueden aplicarse indistintamente a nivel monolingüe o translingüe. Es decir, se pueden comparar dos códigos que están escritos en el mismo, o en distinto, lenguaje de programación. Inicialmente, hemos evaluado los modelos tanto en un escenario académico como en un escenario de detección a gran escala. Finalmente, nuestras mejores propuestas se han comparado con otras propuestas del estado de la cuestión dentro de un mismo marco de evaluación.

Estas pruebas de nuestros modelos se han realizado mediante millones de comparaciones tanto a nivel monolingüe como translingüe empleando diversas técnicas que fueron efectivas al aplicarlas sobre textos escritos en lenguaje natural. Con los recursos utilizados en esta tesis hemos configurado dos escenarios (monolingües y translingües) de evaluación que son un referente para que actuales y futuros trabajos de investigación puedan ajustar y comparar sus propuestas.

2 Estructura de la tesis

- Estado de la cuestión: En este capítulo se resume como se ha abordado la detección de reutilización en el área de la investigación. Primeramente se repasa la detección de reutilización en textos y seguidamente sobre códigos fuente tanto a nivel monolingüe como translingüe.
- Recursos: Este capítulo describe los corpus recopilados, procesados y etiquetados manualmente en este trabajo de tesis.
- Modelos propuestos: Este capítulo propone diferentes modelos para resolver la problemática de la detección automática de reutilización en códigos fuente. Estos modelos están orientados a considerar los códigos fuente como si de un texto se tratara.
- Experimentación: Este capítulo está dividido en dos partes. La primera parte describe la experimentación realizada a nivel monolingüe, para detectar la reutilización entre códigos fuente escritos en un mismo lenguaje de programación. La segunda parte representa la contribución más novedosa de este trabajo de investigación, la detección de reutilización translingüe.
- Evaluación en la competición (CL-) SOCO: En el marco de esta tesis se han organizado dos competiciones de detección de reutilización en código fuente, una monolingüe y la otra translingüe. En ambos casos se estudian los resultados conseguidos por los participantes respeto a los resultados de mayor rendimiento desarrollados en esta tesis.
- Conclusiones y trabajos futuros: Este capítulo resume las principales aportaciones de este trabajo de investigación. También se enuncian las posibles líneas a seguir en trabajos futuros.

3 Principales contribuciones

La principal dificultad para la investigación en detección automática de reutilización en códigos fuente es la escasez de colecciones de códigos fuente que estén etiquetadas con casos de reutilización. De este modo, hemos propuesto y adaptado un conjunto de modelos que mostraron un buen rendimiento en la detección de reutilización y similitud en textos. Los resultados obtenidos por estos modelos han mostrado ser válidos y más eficaces que las demás aproximaciones del estado de la cuestión.

Durante esta tesis se han recopilado una serie de recursos que han ayudado a la investigación de causas y tipos de reutilización. Un ejemplo de ello es el corpus A&T++ con casos de reutilización monolingüe en C y Java, y que a posteriori hemos enriquecido con casos de reutilización translingüe. Hemos etiquetado manualmente los corpus monolingües. En el corpus A&T++, en primer lugar, hemos comparado con la herramienta JPlag siendo los resultados de ambos modelos con desempeño superior a ésta. Finalmente, comparamos todos los modelos propuestos. También añadimos a la comparación un ensamble de modelos utilizando clasificadores conocidos. El ensamble de los modelos, muestra una ligera mejoría en rendimiento. Por otra parte, se han identificado las modificaciones más frecuentes para evadir su detección. Ha resultado importante conocer y diferenciar los mecanismos/modificaciones que se realizan tanto a nivel monolingüe como a nivel translingüe (Flores et al., 2011; Flores et al., 2014; Flores, Moreno, y Rosso, 2016; Flores et al., 2012; Flores et al., 2011).

Se han analizado los cd́igos fuente en busca de casos de reutilización en la primera fase

de la competición Google Code Jam. El objetivo es descubrir si es un entorno propicio para la reutilización, si la reutilización es muy abundante y de que forma se realiza (Flores et al., 2015). El corpus Google Code Jam supone el mayor corpus disponible de códigos fuente con casos de reutilización conocidos. Esto supone una mayor dificultad para descartar los casos de reutilización. Solo seleccionamos los tres lenguajes más populares, C/C++, Java y Python, por lo que existe posibilidad de ampliación. Se han etiquetado manualmente 360 pares de códigos fuente muy similares.

En la experimentación realizada con la colección Google Code Jam hemos realizado más de 34 millones de comparaciones. Comprobamos que es posible aplicar y detectar casos de reutilización en un corpus a gran escala con nuestra propuesta. Además, comparamos con la herramienta JPlag donde observamos que, en líneas generales, se observa un desempeño similar en tareas de mayor cantidad de líneas de código fuente, mientras que en los códigos cortos detectamos más casos de reutilización.

Por otra parte, a nivel translingüe hemos recorrido una línea experimental similar a la monolingüe donde primeramente comprobamos que es importante considerar el código fuente completo para obtener el mejor rendimiento. Se han realizado estudios sobre otro tipo de escenario como plantea el repositorio de códigos fuente Rosettacode.org. Sobre este repositorio hemos estudiado el desempeño de los modelos propuestos recuperando códigos fuente relacionados y también traducidos (Flores et al., 2014; Flores et al., 2015).

Rosettacode.org consiste en un "banco" de códigos fuente con implementaciones de algoritmos conocidos. Hemos recopilado los códigos fuente de C, Java y Python. Tras un etiquetado manual se encontraron casos de reutilización translingüe en este corpus que además hemos enriquecido mediante herramientas de traducción automática.

La experimentación realizada con el corpus Rosettacode.org de gran tamaño comparando los modelos propuestos en dos escenarios de recuperación de información supuso más de cinco millones de comparaciones de códigos fuente por cada modelo de un total de nueve modelos. En general, los modelos muestran un mejor rendimiento en la recuperación de códigos fuente paralelos que en los comparables. Otro experimento que hemos realizado es simular un escenario de reutilización translingüe. Para ello consideramos los pares de códigos fuente paralelos como casos de reutilización y los comparables como no reutilizados. Comparando los modelos en este escenario de reutilización simulado, los modelos basados en características léxicas muestran mejores resultados. Este comportamiento sigue el comportamiento observado en las comparaciones a nivel monolingüe.

Como trabajo paralelo a esta investigación, y con la intención de crear un marco de evaluación común para otros trabajos de investigación, hemos organizado dos competiciones de detección de reutilización automática en códigos fuente. En estas competiciones se han ofrecido tanto recursos para entrenar como para evaluar a disposición de los participantes para el desarrollo y mejora de sus sistemas propuestos. Estos marcos de evaluación han sido propuestos dentro del laboratorio PAN para la detección de reutilización en código fuente (SOCO) y para la detección de reutilización translingüe en código fuente (CL-SOCO) (Flores et al., 2014; Flores et al., 2015; Flores et al., 2015).

Finalmente, hemos realizado una comparación de los modelos de mejor rendimiento propuestos en esta tesis con los sistemas propuestos en las competiciones aprovechando los marcos de evaluación que se facilitan en SOCO y CL-SOCO. Hemos entrenado los modelos utilizando los corpus de entrenamiento de la competición mediante validación cruzada. Comparando los resultados de SOCO con los que se han obtenido con los dos modelos propuestos, hemos podido apreciar que ambos modelos obtienen mejores resultados en términos de la medida F_1. Del mismo modo, en la comparativa translingüe con los sistemas presentados en CL-SOCO, el modelo SoCo-LSA obtiene los mejores resultados siendo estos significativamente mejores, mientras que el modelo SoCo-NG obtiene resultados similares a los mejores de la competición.

4 Conclusiones y trabajos futuros

El objetivo principal de esta tesis ha sido desarrollar modelos eficaces para la detección automática de reutilización en códigos fuente. Para ello, hemos tenido en cuenta distintos escenarios donde es posible que suceda la reutilización de códigos fuente. La re-

utilización/colaboración entre compañeros en el ámbito académico es tan importante como la reutilización en recursos externos como Webs, respositorios, etc. Se han realizado experimentos tanto a nivel monolingüe como translingüe para comprobar la validez de los modelos propuestos en situaciones distintas. Así pues, el trabajo de investigación se estructuró de la siguiente manera: *a)* recopilación de colecciones que contienen casos de reutilización monolingüe como translingüe; *b)* desarrollo y adaptación de modelos de detección de similitud en textos para códigos fuente; *c)* modificación de estos modelos para escenarios translingües; *d)* Estudio de las modificaciones para evitar la detección; *e)* análisis de la detección tanto a nivel académico como a gran escala; *f)* comparación del rendimiento de los modelos propuestos con distintos ensambles de estos modelos; *g)* comparación de los modelos en escenarios de recuperación translingüe; *h)* simulación de un escenario de reutilización translingüe a gran escala; *i)* contribución a la creación de marcos de evaluación para la detección de reutilización en códigos fuente; y *j)* comparación de los modelos desarrollados con propuestas del estado de la cuestión tanto a nivel monolingüe como translingüe.

Identificamos como trabajos futuros las siguientes líneas de investigación: Identificación del origen de la reutilización; detección intrínseca de reutilización; y detección de fragmentos reutilizados.

Bibliografía

Flores, E., A. Barrón-Cedeño, P. Rosso, y L. Moreno. 2011. Detecting Source Code Reuse across Programming Languages. *Póster en Conferencia Sociedad Española para el Procesamiento del Lenguaje Natural.*

Flores, E., P. Rosso, E. Villatoro-Tello, L. Moreno, R. Alcover, y V. Chirivella. 2015. PAN@FIRE: Overview of CL-SOCO track on the Detection of Cross-Language SOurce COde Re-use. En *7th International Workshop of the Forum for Information Retrieval Evaluation*, Gandhinagar, India.

Flores, E., A. Barrón-Cedeño, L. Moreno, y P. Rosso. 2014. Cross-language source code re-use detection. En *3a Conferencia Española de Recuperación de Información*, páginas 145–156, A Coruña, España.

Flores, E., A. Barrón-Cedeño, L. Moreno, y P. Rosso. 2015. Uncovering source code reuse in large-scale academic environments. *Computer Applications in Engineering Education*, 23(3):383–390.

Flores, E., A. Barrón-Cedeño, P. Rosso, y L. Moreno. 2011. Towards the detection of cross-language source code reuse. En *Natural Language Processing and Information Systems*, volumen 6716 de *Lecture Notes in Computer Science*. Springer, páginas 250–253.

Flores, E., A. Barrón-Cedeño, P. Rosso, y L. Moreno. 2012. DeSoCoRe: Detecting Source Code Re-use across programming languages. En *Conference of the North American Chapter of the ACM: HLT*, páginas 1–4, Montreal, Canadá. Association for Computational Linguistics.

Flores, E., M. Ibarra-Romero, L. Moreno, G. Sidorov, y P. Rosso. 2014. Modelos de recuperación de información basados en n-gramas aplicados a la reutilización de código fuente. En *3a Conferencia Española de Recuperación de Información*, páginas 185–188, A Coruña, España.

Flores, E., L. Moreno, y P. Rosso. 2016. Detecting source code re-use with ensemble models. En *4a Conferencia Española en Recuperación de Información*, Granada, España.

Flores, E., A. B.-C. no, L. Moreno, y P. Rosso. 2015. Cross-Language Source Code Re-Use Detection Using Latent Semantic Analysis. *Journal of Universal Computer Science*, 21(13):1708–1725.

Flores, E., P. Rosso, L. Moreno, y E. Villatoro-Tello. 2014. PAN@FIRE: Overview of SOCO track on the detection of SOurce COde re-use. En *6th Forum for Information Retrieval Evaluation.*

Flores, E., P. Rosso, L. Moreno, y E. Villatoro-Tello. 2015. On the detection of source code re-use. En *Forum for Information Retrieval Evaluation*, páginas 21–30.

Información General

SEPLN 2017

XXXIII CONGRESO INTERNACIONAL DE LA SOCIEDAD ESPAÑOLA PARA EL PROCESAMIENTO DEL LENGUAJE NATURAL

Universidad de Murcia – Murcia (España)
20-22 de septiembre 2017
http://sepln2017.um.es/

1 Presentación

La XXXIII edición del Congreso Anual de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) se celebrará los días 20, 21 y 22 de septiembre de 2017 en la Universidad de Murcia.

La ingente cantidad de información disponible en formato digital y en las distintas lenguas que hablamos hace imprescindible disponer de sistemas que permitan acceder a esa enorme biblioteca que es Internet de manera cada vez más estructurada.

En este mismo escenario, hay un interés renovado por la solución de los problemas de accesibilidad a la información y de mejora de explotación de la misma en entornos multilingües. Muchas de las bases formales para abordar adecuadamente estas necesidades han sido y siguen siendo establecidas en el marco del procesamiento del lenguaje natural y de sus múltiples vertientes: Extracción y recuperación de información, Sistemas de búsqueda de respuestas, Traducción automática, Análisis automático del contenido textual, Resumen automático, Generación textual y Reconocimiento y síntesis de voz.

2 Objetivos

El objetivo principal del congreso es ofrecer un foro para presentar las últimas investigaciones y desarrollos en el ámbito de trabajo del Procesamiento del Lenguaje Natural (PLN) tanto a la comunidad científica como a las empresas del sector. También se pretende mostrar las posibilidades reales de aplicación y conocer nuevos proyectos I+D en este campo.

Además, como en anteriores ediciones, se desea identificar las futuras directrices de la investigación básica y de las aplicaciones previstas por los profesionales, con el fin de contrastarlas con las necesidades reales del mercado. Finalmente, el congreso pretende ser un marco propicio para introducir a otras personas interesadas en esta área de conocimiento

3 Áreas Temáticas

Se anima a grupos e investigadores a enviar comunicaciones, resúmenes de proyectos o demostraciones en alguna de las áreas temáticas siguientes, entre otras:

- Modelos lingüísticos, matemáticos y psicolingüísticos del lenguaje
- Desarrollo de recursos y herramientas lingüísticas
- Gramáticas y formalismos para el análisis morfológico y sintáctico
- Semántica, pragmática y discurso
- Resolución de la ambigüedad léxica
- Generación textual monolingüe y multilingüe
- Traducción automática
- Síntesis del habla
- Sistemas de diálogo
- Indexado de audio
- Identificación idioma
- Extracción y recuperación de información monolingüe y multilingüe
- Sistemas de búsqueda de respuestas
- Evaluación de sistemas de PLN

ISSN: 1135-5948

- Análisis automático del contenido textual
- Análisis de sentimientos y opiniones
- Análisis de plagio
- Minería de texto en blogosfera y redes sociales
- Generación de Resúmenes
- PLN para la generación de recursos educativos
- PLN para lenguas con recursos limitados
- Aplicaciones industriales del PLN

4 Formato del Congreso

La duración prevista del congreso será de tres días, con sesiones dedicadas a la presentación de artículos, pósters, proyectos de investigación en marcha y demostraciones de aplicaciones. Además, prevemos la organización de talleres-workshops satélites para el día 19 de septiembre.

5 Comité ejecutivo SEPLN 2017

Presidente del Comité Organizador

- Rafael Valencia García (Universidad de Murcia)

Colaboradores

- Pascual Cantos Gómez (Universidad de Murcia)
- Jesualdo Tomás Fernández Breis (Universidad de Murcia)
- Francisco García Sánchez (Universidad de Murcia)
- Gema Alcaraz Mármol (Universidad de Castilla la Mancha)
- Ángela Almela Sánchez-Lafuente (Universidad Politécnica de Cartagena)

6 Consejo Asesor

Miembros:

- Manuel de Buenaga Rodríguez (Universidad Europea de Madrid, España)
- Pascual Cantos Gómez (Universidad de Murcia)
- Sylviane Cardey-Greenfield (Centre de recherche en linguistique et traitement automatique des langues, Lucien Tesnière. Besançon, Francia)
- Irene Castellón Masalles (Universidad de Barcelona, España)
- Arantza Díaz de Ilarraza (Universidad del País Vasco, España)
- Antonio Ferrández Rodríguez (Universidad de Alicante, España)
- Alexander Gelbukh (Instituto Politécnico Nacional, México)
- Koldo Gojenola Galletebeitia (Universidad del País Vasco, España)
- Xavier Gómez Guinovart (Universidad de Vigo, España)
- José Miguel Goñi Menoyo (Universidad Politécnica de Madrid, España)
- Ramón López Cózar (Universidad de Granada)
- Bernardo Magnini (Fondazione Bruno Kessler, Italia)
- Nuno J. Mamede (Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa, Portugal)
- M. Antonia Martí Antonín (Universidad de Barcelona, España)
- Mª Teresa Martín Valdivia (Universidad de Jaén, España)
- Patricio Martínez Barco (Universidad de Alicante, España)
- Paloma Martínez Fernández (Universidad Carlos III, España)
- Raquel Martínez Unanue (Universidad Nacional de Educación a Distancia, España)
- Ruslan Mitkov (University of Wolverhampton, Reino Unido)
- Manuel Montes y Gómez (Instituto Nacional de Astrofísica, Óptica y Electrónica, México)
- Lluis Padró Cirera (Universidad Politécnica de Cataluña, España)
- Manuel Palomar Sanz (Universidad de Alicante, España)
- Ferrán Pla (Universidad Politécnica de Valencia, España)
- Germán Rigau Claramunt (Universidad del País Vasco, España)
- Horacio Rodríguez Hontoria (Universidad Politécnica de Cataluña, España)
- Leonel Ruiz Miyares (Centro de Lingüística Aplicada de Santiago de Cuba, Cuba)
- Kepa Sarasola Gabiola (Universidad del País Vasco, España)
- Emilio Sanchís (Universidad Politécnica de Valencia, España)
- Encarna Segarra (Universidad Politécnica de Valencia, España)
- Thamar Solorio (University of Houston, Estados Unidos de América)

- Maite Taboada (Simon Fraser University, Canadá)
- Mariona Taulé (Universidad de Barcelona, España)
- Juan-Manuel Torres-Moreno (Laboratoire Informatique d'Avignon / Université d'Avignon, Francia)
- José Antonio Troyano Jiménez (Universidad de Sevilla, España)
- L. Alfonso Ureña López (Universidad de Jaén, España)
- Rafael Valencia García (Universidad de Murcia, España)
- Mª Felisa Verdejo Maíllo (Universidad Nacional de Educación a Distancia, España)
- Luis Villaseñor-Pineda (Instituto Nacional de Astrofísica, Óptica y Electrónica, México)

7 Fechas importantes

Fechas para la presentación y aceptación de comunicaciones:

- Fecha límite para la entrega de comunicaciones: 15 de marzo de 2017.
- Notificación de aceptación: 1 de mayo de 2017.
- Fecha límite para entrega de la versión definitiva: 15 de mayo de 2017.
- Fecha límite para propuesta de talleres y tutoriales: 15 de enero de 2017.

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/home-2/revista/instrucciones-autor/

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 : .....
IBAN

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 : .....

IBAN _____|_____|_____|_____|_____|_____

En.....a.....de.....de.....

(firma)

Sociedad Española para el Procesamiento del Lenguaje Natural. SEPLN

Sr. Director de:
Entidad : .....
Núm. Sucursal : .....
Domicilio : .....
Municipio : ..... Cód. Postal : .....
Provincia : .....
Tipo cuenta
(corriente/caja de ahorro) : .....

Ruego a Vds. que a partir de la fecha y hasta nueva orden se sirvan de abonar a la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) los recibos anuales correspondientes a las cuotas vigentes de dicha asociación.

Les saluda atentamente

Fdo: .....
(nombre y apellidos del firmante)

.....de .....de.....

Cuotas de los socios*: 18 €* (residentes en España) o 2*4 €* (socios residentes en el extranjero).
Nota: La parte inferior debe enviarse al banco o caja de ahorros del socio

Información Adicional

Funciones del Consejo de Redacción

Las funciones del Consejo de Redacción o Editorial de la revista SEPLN son las siguientes:

- Controlar la selección y tomar las decisiones en la publicación de los contenidos que han de conformar cada número de la revista
- Política editorial
- Preparación de cada número
- Relación con los evaluadores y autores
- Relación con el comité científico

El consejo de redacción está formado por los siguientes miembros

L. Alfonso Ureña López (Director)
Universidad de Jaén
laurena@ujaen.es
Patricio Martínez Barco (Secretario)
Universidad de Alicante
patricio@dlsi.ua.es
Manuel Palomar Sanz
Universidad de Alicante
mpalomar@dlsi.ua.es
Felisa Verdejo Maíllo
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:

Manuel de Buenaga	Universidad Europea de Madrid (España)
Sylviane Cardey-Greenfield	Centre de recherche en linguistique et traitement automatique des langues (Francia)
Irene Castellón	Universidad de Barcelona (España)
Arantza Díaz de Ilarraza	Universidad del País Vasco (España)
Antonio Ferrández	Universidad de Alicante (España)
Alexander Gelbukh	Instituto Politécnico Nacional (México)
Koldo Gojenola	Universidad del País Vasco (España)
Xavier Gómez Guinovart	Universidad de Vigo (España)
José Miguel Goñi	Universidad Politécnica de Madrid (España)
Bernardo Magnini	Fondazione Bruno Kessler (Italia)
Nuno J. Mamede	Instituto de Engenharia de Sistemas e Computadores (Portugal)

Cartas al director

Sociedad Española para el Procesamiento del Lenguaje Natural
Departamento de Informática. Universidad de Jaén
Campus Las Lagunillas, Edificio A3. Despacho 127. 23071 Jaén
secretaria.sepln@ujaen.es

Más información

Para más información sobre la Sociedad Española del Procesamiento del Lenguaje Natural puede consultar la página web http://www.sepln.org.
Los números anteriores de la revista se encuentran disponibles en la revista electrónica: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/issue/archive
Las funciones del Consejo de Redacción están disponibles en Internet a través de http://www.sepln.org/category/revista/consejo_redaccion/
Las funciones del Consejo Asesor están disponibles Internet a través de la página http://www.sepln.org/home-2/revista/consejo-asesor/
La inscripción como nuevo socio de la SEPLN se puede realizar a través de la página http://www.sepln.org/socios/inscripcion-para-socios/

Información General