sociedad española para el Procesamiento del Lenguaje Natural

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

ISSN: 1135-5948

Artículos

Tesis

Información General

© 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1135-5948

Comité Editorial

Consejo de redacción

L. Alfonso Ureña López	Universidad de Jaén	laurena@ujaen.es	(Director)
Patricio Martínez Barco	Universidad de Alicante	patricio@dlsi.ua.es	(Secretario)
Manuel Palomar Sanz	Universidad de Alicante	mpalomar@dlsi.ua.es	
Felisa Verdejo Maillo	UNED	felisa@lsi.uned.es	

ISSN: 1135-5948
ISSN electrónico: 1989-7553
Depósito Legal: B:3941-91
Editado en: Universidad de Jaén
Año de edición: 2015
Editores: Mariona Taulé Delor, Universidad de Barcelona, mtaule@ub.edu
Mª Teresa Martín Valdivia, Universidad de Jaén, maite@ujaen.es
Publicado por: Sociedad Española para el Procesamiento del Lenguaje Natural
Departamento de Informática. Universidad de Jaén
Campus Las Lagunillas, EdificioA3. Despacho 127. 23071 Jaén
secretaria.sepln@ujaen.es

Consejo asesor

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)
M. Antonia Martí	Universidad de Barcelona (España)
M. Teresa Martín Valdivia	Universidad de Jaén (España)
Patricio Martínez-Barco	Universidad de Alicante (España)
Raquel Martínez	Universidad Nacional de Educación a Distancia (España)
Leonel Ruiz Miyares	Centro de Linguística Aplicada de Santiago de Cuba (Cuba)
Ruslan Mitkov	Universidad de Wolverhampton (Reino Unido)

© 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

Manuel Montes y Gómez	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)
Lidia Moreno	Universidad Politécnica de Valencia (España)
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)
Kepa Sarasola	Universidad del País Vasco (España)
Emilio Sanchís	Universidad Politécnica de Valencia (España)
Thamar Solorio	University of Houston (Estados Unidos de America)
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 Maillo	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

Marina Lloberes	Universidad de Barcelona (España)
Enrique Puertas	Universidad Europea (España)
Eugenio Martínez Cámara	Universidad de Jaén (España)
Salud M. Jiménez Zafra	Universidad de Jaén (España)

© 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

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 54 de la revista *Procesamiento del Lenguaje Natural* contiene trabajos correspondientes a dos apartados diferenciados: comunicaciones científicas y resúmenes de

© 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

tesis. 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 22 trabajos para este número de los cuales 20 eran artículos científicos y 2 correspondían a resúmenes de tesis. De entre los 20 artículos recibidos 11 han sido finalmente seleccionados para su publicación, lo cual fija una tasa de aceptación del 55%.

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. 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 2015
Los editores

© 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

sociedad española para el Procesamiento del Lenguaje Natural

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

ISSN: 1135-5948

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 54th issue of the *Procesamiento del Lenguaje Natural* journal contains scientific papers and doctoral dissertation summaries. All of these were accepted by the traditional peer reviewed

© 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

process. We would like to thank the Advisory Committee members and additional reviewers for their work.

Twenty-two papers were submitted for this issue of which twenty were scientific papers and two dissertation summaries. From these twenty papers, we selected eleven (55% 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 2015
Editorial board

© 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

sociedad española para el Procesamiento del Lenguaje Natural

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

ISSN: 1135-5948

Artículos

Tesis

Información General

© 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

Artículos

Procesamiento del Lenguaje Natural, Revista nº 54, marzo de 2015, pp 13-20 recibido 20-10-14 revisado 27-01-15 aceptado 11-02-15

A Combination based on OWA Operators for Multi-label Genre Classification of web pages

Una combinación basada en operadores OWA para la Clasificación de Género Multi-etiqueta de páginas web

Chaker Jebari
Colleges of Applied Sciences
P. O. Box 14, P.C. 516, Sultanate of Oman
jebarichaker@yahoo.fr

Resumen: En este trabajo se presenta un nuevo método para la identificación de género que combina clasificadores homogéneos utilizando OWA (promedio ponderado) Pedimos operadores. Nuestro método utiliza caracteres n-gramas extraídos de diferentes fuentes de información, tales como URL, título, encabezados y anclajes. Para hacer frente a la complejidad de las páginas web, se aplicó MLKNN como un clasificador multi-etiqueta, en el que una página web puede verse afectada por más de un género. Los experimentos llevados a cabo usando un conocido corpus multi-etiqueta muestran que nuestro método logra buenos resultados.
Palabras clave: OWA, combinación, multi-etiqueta, clasificadores, género, página web.

Abstract: This paper presents a new method for genre identification that combines homogeneous classifiers using OWA (Ordered Weighted Averaging) operators. Our method uses character n-grams extracted from different information sources such as URL, title, headings and anchors. To deal with the complexity of web pages, we applied MLKNN as a multi-label classifier, in which a web page can be affected by more than one genre. Experiments conducted using a known multi-label corpus show that our method achieves good results.
Keywords: OWA, combination, multi-label, classifier, genre, web page.

1 Introduction

As the World Wide Web continues to grow exponentially, the classification of web pages becomes more and more important in web searching. Web page classification, assigns a web page to one or more predefined classes.

According to the type of the class, the classification can be divided into sub-problems: topic classification, sentiment classification, genre classification, and so on. Currently, search engines use keywords to classify web pages. Returned web pages are ranked and displayed to the user, who is often not satisfied with the result. For example, searching for the keyword "Java" will provide a list of web pages containing the word "Java" and belonging to different genres such as "tutorial", "exam", "Call for papers", etc. Therefore, web page genre classification could be used to improve the retrieval quality of search engines (Stein and Meyer, 2008).

Generally speaking, a genre is a category of artistic, musical, or literary composition characterized by a particular style, form, or content, but more specialized characterizations have been proposed (Santini, 2007).

According to Shepherd and Watters (1998), the genres found in web pages (also called cyber-genres) are characterized by the triple <content, form, functionality>. The content and form attributes are common to non-digital genres and refers to the text and the layout of the web page respectively. The functionality attribute concerns exclusively digital genres and describes the interaction between the user and the web page.

A common fact for all defintions is that genre and topic are orthogonal, meaning that documents addressing the same topic can be of different genres and vice versa. Following this

ISSN 1135-5948 © 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

way, we can say that a document genre describes a style of writing and/or presentation rather than the document topic. This style can be captured by exploiting the structure of the document rather than its content.

It is worth noting that a web page is a complex object that is composed of different sections belonging to different genres. For example, a conference web page contain information on the conference, topics covered, important dates, contact information and a list of hyperlinks to related information. This complexity need to be captured by a multi-label classification scheme in which a web page can be assigned to multiple genres.

In this paper we used character n-grams extracted from different sources such as URL, title, headings and hyperlinks. Our constribution is to use OWA (Ordered Weighted Averaging) operators to combine the outputs of three homogenous classifiers: contextual, logical and hyperlink classifiers.

The contextual classifier uses the URL which defines the location of a web page. It is composed of three parts: host name (domain), directory path and file name (Berners-Lee, Fielding, and Masinter, 1998). The URL is not expensive to obtain and it is one of the more informative sources about the genre of the web page. URLs are often meant to be easily recalled by humans, and web sites that follow good design techniques will encode useful words that describe their resources in the web site's host name (domain). Web sites that present a huge amount of information often break their contents into web pages. This information structuring is also accompanied with URLs structuring. For example, if the file extension is PDF, PS or DOC, then the document is long and it can be a paper, a book, a thesis, a manual, etc. Another example, if the file name contain some genre specific words like faq, cv, how, thesis, etc., we can easily recognize the genre of the web page.

The structure of a web page were used to identify the genre (Crowston and Williams, 1997; Jebari and Ounalli, 2004).

Jebari and Ounalli (2004) investigated the usefulness of the internal, also called logical structure to identify the genre of a web page. They used words included in the title and headings to extract the internal structure.

The hyperlink structure has been investigated by Crowston and Williams (1997) to identify the form of the web page and therefore can help to identify its genre.

In our work we have used the hypertext structure in different way than used by the previous researches. In our work we have used the character n-grams and the words contained in hyperlinks contrary to many other researches that use the number of internal and external links, number of images, etc. (Crowston and Williams, 1997; Boese and Howe, 2005; Lim, Lee, and Kim, 2005).

The remainder of the paper is organized as follows. Section 2 reviews previous works on genre classification of web pages. Section 3 describes the multi-label classification. Section 4 presents a brief overview about classifier combination and describes in details OWA operators. Section 5 describes our method. Section 6 evaluates and compares our method with other previous works. Finally, Section 7 concludes our paper and suggests future research directions.

2 Related works

A broad number of studies on genre classification of web documents have been proposed in the literature (Santini, 2007). These studies differ with respect to the following three factors: 1) the features used to represent the web document, 2) the classification methods used to identify the genre of a given web document and 3) the list of genres used in the evaluation, called also genre palette.

Many types of features have been proposed for automatic genre classification. These features can be grouped on four groups. The first group refers to surface features, such as function words, genre specific words, punctuation marks, document length, etc. The second group concerns structural features, such as Parts Of Speech (POS), Tense of verbs, etc. The third group is the presentation features, which mainly describe the layout of document. Most of these features concerns HTML documents and cannot be extracted from plain text documents. Among these features we quote the number of specific HTML tags and links. The last group of features is often extracted from metadata elements (URL, description, keywords, etc.) and concerns only structured documents.

Once a set of features has been extracted it is necessary to choose a classification method, which are often based on machine learning

techniques such as Naive bayes, SVM, K-nearest neighbor, decision trees, neural networks, centroid-based techniques, etc. (Mitchell, 1997). Broadly speaking, classification methods can be divided into two main categories: single-label and multi-label methods (Tsoumakas, Katakis, and Vlahavas, 2010). In single label methods, a document is associated to only one label, whereas, in multi-label methods, a document is assigned to a set of labels.

The third factor concerns the list of genres used for the evaluation. Many genre corpora[1] (KI-04, KRYS-I, 20-genre, SANTINIS, etc.) have been compiled and used to evaluate genre identification tasks. These corpora differ with respect to the number of genres, the types of genres and the number of documents associated to each genre.

Table 1 presents an overview of features, machine learning techniques and corpora used in web genre classification.

Autor	**Features**	**Machine learning**	**Corpora**
(Meyer and stein, 2004)	HTML tag frequencies, classes of words (names, dates, etc.), frequencies of punctuation marks and POS tags	Discriminant Analysis	KI-04
(Lim, Lee, and Kim, 2005)	POS tags, URL, HTML tags, token information, most frequent function words, most frequent punctuation marks, syntactic chunks	K-Nearest Neighbor	The corpus consists of 1224 documents distributed across 15 genres (home page, public, commercial, bulletin, link collection, image collection, FAQ, discussion, product specification, etc.)
(Kennedy and Shepherd, 2005)	Content features (common words, Meta tags), form features (e.g. number of images), and functionality features (e.g., number of links, use of JavaScript).	neural network	The corpus is composed of 321 web pages classified as home pages or as noise pages (not home page). The home pages are classified into three subgenres (corporate home pages, personal home pages and organization home pages.
(Santini, 2007)	Most frequent English words, HTML tags, POS tags, punctuation symbols, genre-specific core vocabulary	SVM	SANTINIS
(Vidulin, Lustrek, and Gams, 2009)	Surface features (unction words, genre-specific words, sentence length). Structural features (POS tags, sentence type). Presentation features describe the formatting of a document through the HTML tags. Context features describe the context in which a web page was found (e.g. URL, hyperlinks, etc.).	AdaBoost	20-genre
(Kim and Ross, 2008)	Image features (extracted from the visual layout of the first page) Style features: genre-prolific words. Textual features are represented by a bag of words extracted from the content of the PDF document.	Naive bayes, SVM, Random Forest	KRYS-I
(Jebari, 2008)	Words extracted from URL, title, headings and anchors	Centroid-based	KI-04 and WebKB
(Kanaris and Stamatatos, 2009)	Character n-grams extracted from text and structure	SVM	20-genre
(Mason, 2009)	Character n-grams extracted from the textual content	SVM	20-genre
(Abramson and Aha, 2012)	Character n-grams extracted from URL	SVM	Syracuse and SANTINIS corpora

Table 1: Overview of previous works

[1]http://www.webgenrewiki.org/index.php5/Genre_Collection_Repository

3 Multi-label classification

In traditional single-label classification, a classifier is built and trained using a set of examples associated with just one single label l of a set of disjoint labels $L=\{l_1, l_2, \ldots l_i, \ldots\}$, where $|L|>1$. Moreover, in multi-label classification, the examples can be associated with a set of labels $Y \subseteq L$. In the literature, different methods have been proposed to be applied to multi-label classification problems. These methods are grouped into two main categories: problem transformation and algorithm transformation (Tsoumakas, Katakis, and Vlahavas, 2010).

Problem transformation methods are algorithm independent and transform a multi-label learning problem into one or more single-label learning problems. The most widely used transformation methods are Binary Relevance BR, Label Power Set (LP) and Random k-labelsets method (RA*k*EL). The algorithm transformation methods extend existing learning algorithms to deal with multi-label data directly. Several transformation methods have been proposed in the literature such as BR-SVM, BPMLL and MLKNN.

MLKNN is an instance-based learner (Zhang and Zhou, 2007), it learns a single classifier h_i for each label $l_i \in L$. However, instead of using the standard k-nearest neighbor (KNN) classifier as a base learner, it implements h_i by means of a combination of KNN and Bayesian inference. Given an example x, it finds the k nearest neighbors of x in the training data and counts the number of occurrences of l_i among these neighbors. Considering this number, y, as information in the form of a realization of a random variable Y, the posterior probability of $l_i \in L$ is given by:

$$P(l_i \in L/Y = y)=\frac{P(Y = y/l_i \in L)\cdot P(l_i \in L)}{P(Y = y)} \quad (1)$$

This, leads to the following classification:

$$H_i(x)= \{(l_i, f(l_i)), \ldots, (l_i, f(l_i)), \ldots\} \quad (2)$$

Where $f(l_i)$ is the posterior probability of $l_i \in L$ defined in the previous equation.

The prior probability $P(l_i \in L)$ as well as the conditional probability $P(Y = y/l_i \in L)$ are estimated from the training data in terms of corresponding relative frequencies.

4 OWA Operators

Based on the assumption that each source of information provides a different view point, a combination has the potential of providing better results than any single method. There are various methods to combine such classifiers (Kuncheva, 2004). These methods can be classified according to the classifier used. Generally, classifiers can be combined at different levels: abstract level, ranking level and measurement level (Kang and Kim, 1995). In abstract level, combination methods combine simple class labels. In ranked level, combination methods combine ranked lists of class labels ordered according to the degree of membership of the input pattern. In the measurement level, combination methods combine values provided by individual classifiers as a measure of the degree of membership of the input pattern to each class. Among the three categories, the combination of classifiers at the measurement level is expected to be the most effective, since it uses all information available.

Different types of aggregation operators are found in the literature to combine the information produced by measurement level classifiers (Beliakov, Pradera, and Calvo, 2007). A very common aggregation operator is the Ordered Weighted Averaging (OWA) operator which is first introduced in (Yager, 1988).

Broadly speaking, a mapping F: $[0,1]^n \rightarrow [0, 1]$ is called an OWA operator of dimension n if it is associated with a weighting vector $W=[w_1, \ldots, w_i, \ldots, w_n]$, such that $w_i \in [0, 1]$, $\sum_i w_i =1$ and $F(a_1, \ldots, a_n) = \sum_i w_i b_i$ where b_i is the i-th largest element in the collection $a_1, \ldots, a_n$. Yager suggested two methods to identify the weights w_i's. The first approach uses learning techniques and the second one uses fuzzy linguistic quantifiers to gives semantics to the weights. Herrera and Verdegay (Herrera and Verdegay, 1996) defined a quantifier function as follows:

$$Q(r)=\begin{cases} 0 & r \prec a \\ \frac{r-a}{r-b} & r \in [a,b] \\ 1 & r \succ b \end{cases} \quad (3)$$

Where $a, b \in [0, 1]$ are two parameters.

Using this quantifier function, Yager (1988) computes the weight w_i as follows:

$$w_i = Q\left(\frac{i}{n}\right) - Q\left(\frac{i-1}{n}\right), \text{ for } i=1, 2, \ldots, n \qquad (4)$$

Where n is the number of classifiers to combine. According to Yager (Yager, 1988), using the quantifier function defined above, we can identify 5 common OWA operators which are: Minimum, Maximum, Average, Vote1 and Vote2.

5 Proposed approach

This section describes how a web page is represented and how a new web page is classified.

5.1 Web page representation

To represent a web page, our approach performs five pre-processing steps:

Step1. This step consists in extracting the content of the elements URL, title, headings and anchors.

Step2. In this step, our method processes the content of each element separately, by removing digits, special characters (., :, /, ?, &, -, _, $, #, etc.) and stop words that differ according to the element. For the URL element we removed the stop words (http, www, etc.), since they are commonly used in all URLs. For the rest of the elements (title, headings and anchor) we removed the known stop words such as: the, of, for, etc.

Step3. This step consists in extracting words and character n-grams from all elements (URL, title, headings and anchors). A character n-grams is a set of n contiguous characters. For example, from the string 'myCV', we can extract 3 different 2-grams (my, yc, cv), 2 different 3-grams (myc, ycv) and one 4-gram (myCV). In our approach we extracted all character n-grams of length between 2 and 5, since they can capture all genre specific words in the URL.

Step4. One of the main challenges of text classification tasks is the high dimenstionality. A typical text will contain a hundreds of features, hence it is extremely difficult to produce an accurate classification without any dimension reduction. Many dimension reduction techniques have been proposed in the literature (Yang and Pedersen, 1997). In this paper we used the Document frequency thresholding technique. Given a term *t*, this technique computes the document frequency *DF* by counting the number of documents in which the term *t* occurs. Then reduce the terms whose document frequency is less than a predefined threshold. In this study, we decided to keep only URL words and character n-grams that appear in at least 100 web pages. For the other elements (title, headings and anchors), we removed words and character n-grams that appears in less than 10 web pages.

Step5. Using the Vector Space Model (VSM) (Salton and Buckley, 1988), a web page is represnetated by a vector where each term is assoictaed with a weight using the *TFIDF* weighting formula (Sebastiani, 2002).

5.2 Classification of a new web page

Given a new webpage p_i, our approach applies the five pre-processing steps described in the previous section to extract character n-grams from different sources (URL, title, headings and anchors). A web page p_i is represented by three vectors. The first vector cp_i, called contextual vector, contains character n-grams extracted from the URL. The second vector lp_i, called logical vector, contains character n-grams extracted from title and headings. The third vector hp_i, called hyperlink vector and contains character n-grams extracted from the anchors. The vectors cp_i, lp_i and hp_i are used to perform contextual, logical and hyperlink classifications named respectively $CC(cp_i)$, $LC(lp_i)$ and $HC(hp_i)$.

For a predefined set of genres $G=\{g_1, \ldots, g_i, \ldots, g_m\}$, the contextual, logical and hyperlink classifications are defined as follows:

$$\begin{aligned} CC(cp_i) &= \{(g_1, \alpha_1), \ldots, (g_i, \alpha_i), \ldots, (g_m, \alpha_m)\} \\ LC(lp_i) &= \{(g_1, \beta_1), \ldots, (g_i, \beta_i), \ldots, (g_m, \beta_m)\} \\ HC(hp_i) &= \{(g_1, \lambda_1), \ldots, (g_i, \lambda_i), \ldots, (g_m, \lambda_m)\} \end{aligned} \qquad (5)$$

Where α_i, β_i and λ_i are the similarities between the web page p_i and the genre g_i, for the contextual, logical and hyperlink classification respectively. This similarity is calculated using the cosine formula.

In order to provide a final classification, our approach combines the contextual, logical and hyperlink classifications using the different OWA operators.

For a given web page p_i, the final classification $C(p_i)$ is defined as follows:

$$\begin{aligned} C(p_i) &= OWA_j(CC(cp_i), LC(lp_i), HC(hp_i)) \\ &= \{(g_1, OWA_j(\alpha_1, \beta_1, \lambda_1)), \ldots, \\ &\quad (g_i, OWA_j(\alpha_i, \beta_i, \lambda_i)), \ldots, \\ &\quad (g_m, OWA_j(\alpha_m, \beta_m, \lambda_m))\} \end{aligned} \qquad (6)$$

Where OWA_j is one of the five OWA operators introduced in Section 4.

6 Experimentation

Our experimentation methodology is to experiment contextual, logical, hyperlink and combined separately. In our experimentation we used MLKNN classifier. This classifier is already implemented in the Mulan toolkit[2]. In our experimentation, we followed the k-cross-validation procedure which consists of randomly splitting the corpus into k equal parts. Then we used k-1 parts for testing and the remaining one part for training. This process is performed k times and the final performance is the average of the k individual performances. Due to the small number of web pages in each genre, we decided to use 3-cross-validation.

6.1 Corpus

In this paper we used the corpus 20-genre (Vidulin, Lusterk, and Gams, 2007). For the best of my knowledge, 20-genre is the only multi-label genre corpus available at the moment. This corpus consists of 1539 English web pages classified into 20 genres as shown in Table 2.

Genre	#pages	Genre	#pages
Blog	83	Index	308
Adult	79	Informative	318
Children's	113	Journalistic	206
Commercial/ Promotional	193	Official	85
Community	82	Personal	133
Content Delivery	207	Poetry	76
Entertainment	126	Prose Fiction	75
Error Message	90	Scientific	98
FAQ	71	Shopping	81
Gateway	119	User Input	96

Table 2: Composition of 20-genre corpus

6.2 Evaluation metrics

The evaluation of multi-label classifiers requires different evaluation metrics from those used in single-label classifiers. In a single-label classification, conventional metrics such as accuracy, precision, and recall are used to verify that an example is correctly or incorrectly classified. However, performance evaluation in multi-label classification is much complicated than traditional single-label setting, as each example can be associated with multiple labels simultaneously. Several multi-label evaluation metrics have been proposed in the literature (Tsoumakas, Katakis, and Vlahavas, 2010).

In this study, we used the following metrics: Hamming Loss, Micro-averaged precision, One-Error, Coverage and Ranking Loss.

Hamming Loss (HL) evaluates how many times an example-label pair is misclassified. The smaller the value of HL, the better the performance. The performance is perfect when the value of HL is 0.

Micro-averaged precision (MP) is the precision averaged over all the example/label pairs. The higher the value of the MP, the better the performance.

One-Error (OE) evaluates how many times the top-ranked label is not in the set of relevant labels of the example. The smaller the value of OE, the better the performance.

Coverage (CV) evaluates how far, on average, we need to go down the list of ranked labels in order to cover all the relevant labels of the example.The smaller the value of CV, the better the performance.

Ranking Loss (RL) evaluates the average fraction of label pairs that are reversely ordered for the particular example. The smaller the value of RL, the better the performance, so the performance is perfect when RL=0.

6.3 Results and discussion

6.3.1 Experiment1

In this experiment, we evaluate the contextual (CC), logical (LC) and hypertext (HC) classifiers using character n-grams and bag of words (BOW) representations. The results are reported in Table 3.

		HL	OE	RL	CV	MP
CC	Grams	0.082	0.700	0.312	7.126	0.602
	BOW	0.085	0.712	0.344	7.110	0.550
LC	Grams	0.081	0.412	0.215	8.774	0.901
	BOW	0.080	0.415	0.300	9.005	0.805
HC	Grams	0.081	0.560	0.280	8.123	0.720
	BOW	0.084	0.670	0.320	8.250	0.680

Table 3: Results achieved by contextual, logical and hypertext classifiers

By considering each classifier seperatly, we can conclude that using character n-grams achieves better results in comparison with BOW representation. Overall, the logical

[2] http://mulan.sourceforge.net/index.html

classfier (LC), reported the best results with respect to all all metrics except the Coverage metric which is better for contextual and logical classifiers. This is because, the majority of the significant genre words or grams are found in the title and heading sections. Moreover, the contextual classifier achieves the lowest results due to the lack of genre specific words in the URL.

6.3.2 Experiment2

To evaluate the combined classifier, we used different OWA operators described in Section 4. The results achieved are presented in Table 4. Overall, the best results are achieved using Avg operator with respect to all metrics except the Coverage metric where the highest value is reported by the Vote1 operator. Moreover, we observe that the results obtained using character n-grams are much better in comparison with BOW representation.

		HL	OE	RL	CV	MP
Min	Grams	0.101	0.098	0.088	9.100	0.760
	BOW	0.201	0.102	0.090	8.550	0.720
Max	Grams	0.116	0.085	0.094	8.885	0.815
	BOW	0.186	0.082	0.090	8.900	0.770
Avg	Grams	0.065	0.054	0.082	9.118	0.941
	BOW	0.070	0.066	0.090	9.002	0.935
Vote1	Grams	0.095	0.088	0.092	7.778	0.885
	BOW	0.092	0.090	0.096	7.320	0.820
Vote2	Grams	0.058	0.055	0.082	8.226	0.920
	BOW	0.060	0.066	0.099	8.100	0.905

Table 4: Results achieved using different OWA operators

6.4 Comparison with similar works

In this section we compare our proposed method with three previous studies (See Table 5). This studies uses the multi-label corpus 20-genre.

Study	Classifier	MP
Our work	MLKNN	0.94
(Vidulin, Lustrek, and Gams, 2009)	AdaBoost	0.35
(Mason, 2009)	SVM	0.70
(Kanaris and Stamatatos, 2009)	SVM	0.74

Table 5: Classifier used and performance achieved by some previous works

As shown in the above table, our method achieves the best results. We should mention that the other studies are based on single-label classifiers such as SVM and AdaBoost, whereas in our study we used MLKNN classifier which is a multi-label classification method. It is worth noting also that all the studies used character n-grams except (Vidulin et al., 2009). So, we can confirm that using character n-grams we obtain better results rather than using other kind of features. Morover, using a multi-label classifier we can achieve better classification performance in comparison with single-label classifiers such as SVM and AdaBoost.

7 Conclusion and future work

In this paper, we proposed a combination of multi-label genre classifications using OWA operators. Our method exploits the character n-grams extracted from different sources such as URL, title, headings and links. The experiments conducted using a known multi-labeled corpus show that using character n-grams achieves better results than using bag-of-words. As part of the future work, we plan to evaluate our approach using other data sets, preferably with more examples. Morover, we plan to test other combination methods such as Dempester-shafer theory of evidence and Behavior Knowledge Space.

References

Abramson, M., and D. W. Aha. 2012. What's in a URL? Genre Classification from URLs. *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Beliakov, G., A. Pradera and T. Calvo. 2007. *Aggregation Functions: A Guide for Practitioners*. Springer-Verlag.

Berners-Lee, T., R. Fielding, and L. Masinter. 1998. RFC2396: *Uniform Resource Identifiers (URI): Generic Syntax*, RFC editor, USA.

Boese, E., and A. Howe. 2005. Genre Classification of Web Documents. *In Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-05)*, USA.

Crowston, K., and Williams, M. 1997. Reproduced and Emergent Genres of Communication on the World Wide Web. *In Proceedimgs of the 30th Hawaii International Conference on System Sciences*, USA.

Herrera, F., and J. L. Verdegay. 1996. *Genetic algorithms and soft computing*. PhysicaVerlag, Heidelberg, Germany.

Jebari, C. 2008. Catégorisation Flexible et Incrémentale avec raffinage de pages web par genre. PhD thesis, Tunis University, Tunisia.

Jebari, C., and H. Ounalli. 2004. The Usefulness of Logical Structure in Flexible Document Categorization. *In Proceeding of the International Conference on Computational Intelligence*, Turkey.

Kanaris, I., and E. Stamatatos. 2009. Learning to Recognize Webpage Genres. *Information Processing and Management journal*, 45(5): 499-512.

Kang, H. J. and J.H. Kim. 1995. Dependency relationship based decision combination in multiple classifier systems. *In Proceedings of the 14th International Joint Conf on Artificial Intelligence*.

Kennedy, A., and M. Shepherd. 2005. Automatic Identification of Home Pages on the Web. *In the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05)*, USA.

Kessler, B., G. Nunberg, and H. Schutze. 1997. Automatic detection of text genre. *In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Morristown, NJ, USA.

Kim, Y. and S. Ross. 2008. Examining Variations of Prominent Features in Genre Classification. *In the Proceedings of the 41th Annual Hawaii International Conference on System Sciences (HICSS'08)*, USA.

Kuncheva, LI. 2004. Combining Pattern Classifiers Methods and Algorithms. John Wiley & Sons.

Lim, C. S., K. J. Lee, and G. C. Kim. 2005. Multiple sets of features for automatic genre classification of web documents. *Information Processing and Management Journal*, 41(5): 11263-1276.

Mason, J. 2009. An n-gram Based Approach to the Automatic Classification of Web Pages by Genre. PhD thesis, Dalhousie University, Canada.

Meyer, S. E., and B. Stein. 2004. Genre Classification of Web Pages. *In Proceedings of the 27th German Conference on Artificial Intelligence.*

Mitchell, T. 1997. *Machine Learning*, McGraw Hill.

Salton, G., and Buckley, C. 1988. Term weighting approaches in automatic text retrieval. *Information Processing and Management 24*(5): 513–523.

Santini, M., 2007. Automatic Identification of Genre in Web Pages. PhD thesis, University of Brighton, UK.

Sebastiani, F. 2002. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1), pp. 1-47.

Shepherd, M. and C. Watters. 1998. Evolution of Cybergenre. *In proceedings of the 31th Hawaiian International Conference on System Sciences*, USA.

Stein, B., and S. E. Meyer. 2008. Retrieval Models for Genre Classification. *Scandanivian Journal of Information Systems*. 20(1):93-119.

Tsoumakas, G., I. Katakis, I. Vlahavas. 2010. Mining Multi-Label Data. *Data Mining and Knowledge Discovery Handbook*, Springer.

Vidulin, V., M. Luštrek, and M. Gams. 2007. Using Genres to Improve Search Engines, *1st International Workshop: Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing, RANLP'2007*, Borovest, Bulgaria, pp. 45-51.

Vidulin, V., M. Lustrek and M. Gams. 2009. Multi-Label Approaches to Web Genre Identification. *Journal of Language and Computational Linguistics*, 24(1): 97-114.

Yager, R. 1988. On ordered weighted averaging aggregation operators in multi-criteria decision making. *IEEE Transactions on Systems, Man, and Cybernetics*, 18:183-190.

Yang, Y., and J. O. Pedersen. 1997. A Comparative Study on Feature Selection in Text Categorization, *In proceedings of ICML'1997*.

Zhang, M. L., and Z. H. Zhou. 2007. Ml-KNN: A lazy learningapproach to multi-label learning. *Pattern Recognition*, 40(1):2038–2048.

Procesamiento del Lenguaje Natural, Revista nº 54, marzo de 2015, pp 21-28 recibido 24-10-14 revisado 27-01-15 aceptado 12-02-15

eSOLHotel: Generación de un lexicón de opinión en español adaptado al dominio turístico

eSOLHotel: Building an Spanish opinion lexicon adapted to the tourism domain

M. Dolores Molina González, Eugenio Martínez Cámara, M. Teresa Martín Valdivia, Salud M. Jiménez Zafra
Departamento de Informática, Escuela Politécnica Superior de Jaén
Universidad de Jaén, E-23071 - Jaén
{mdmolina, emcamara, maite, sjzafra}@ujaen.es

Resumen: Desde que la web 2.0 es el mayor contenedor de opiniones en todos los idiomas sobre distintos temas o asuntos, el estudio del Análisis de Sentimientos ha crecido exponencialmente. En este trabajo nos centramos en la clasificación de polaridad de opiniones en español y se presenta un nuevo recurso léxico adaptado al dominio turístico (eSOLHotel). Este nuevo lexicón usa el enfoque basado en corpus. Se han realizado varios experimentos usando una aproximación no supervisada para la clasificación de polaridad de las opiniones en la categoría de hoteles del corpus SFU. Los resultados obtenidos con el nuevo lexicón eSOLHotel superan los resultados obtenidos con otro lexicón de propósito general y nos animan a seguir trabajando en esta línea.
Palabras clave: Clasificación de polaridad, corpus de opiniones en español, lexicón dependiente del dominio, turismo.

Abstract: Since Web 2.0 is the largest container for subjective expressions about different topics or issues expressed in all languages, the study of Sentiment Analysis has grown exponentially. In this work, we focus on Spanish polarity classification of hotel reviews and a new domain-dependent lexical resource (eSOLHotel) is presented. This new lexicon has been compiled following a corpus-based approach. We have carried out several experiments using an unsupervised approach for the polarity classification over the category of hotels from corpus SFU. The results obtained with the new lexicon eSOLHotel outperform the results with other general purpose lexicon.
Keywords: Polarity classification, Spanish reviews corpus, dependent-domain lexicon, tourism.

1 Introducción

En los últimos años, el interés por el Análisis de Sentimientos (AS) (conocido en inglés como sentiment analysis u opinion mining) ha crecido significativamente debido a diferentes factores (Pang y Lee, 2008) (Liu, 2012) (Tsytsarau y Palpanas, 2012). Por una parte, el incremento de la creación y compartición de datos por parte de los usuarios de Internet haciendo uso de las nuevas plataformas y servicios que están emergiendo continua y expeditamente. Por otra parte, el consumo de datos online comienza a ser una tarea imprescindible y rutinaria para la toma de decisiones a nivel individual o colectivo.

Muchas son las tareas estudiadas en AS, siendo una de las más consolidadas la clasificación de la polaridad. En esta tarea se han seguido distintas aproximaciones, aunque son dos las líneas principales. Por una parte, la aproximación basada en técnicas de aprendizaje automático (Machine Learning ML), la cual se basa en entrenar unos modelos a partir de una colección de datos etiquetada a priori, con el objetivo de predecir el valor de salida correspondiente a cualquier dato de entrada válido. Los clasificadores pueden estar basados

ISSN 1135-5948 © 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

en distintos algoritmos, entre los más utilizados están las máquinas de soporte vectorial (conocido en inglés como Support Vector Machines, SVM) o máxima entropía (ME). Estos clasificadores tienen el inconveniente de necesitar gran cantidad de datos de entrada para un entrenamiento previo y poder obtener buenos resultados. Trabajos como el de Pang, Lee y Vaithyanathan (2002) usan este enfoque supervisado para resolver el problema de la clasificación de polaridad.

La segunda línea, conocida como aproximación basada en Orientación Semántica (OS), obtiene la polaridad de cada documento como la agregación de la inclinación positiva o negativa de sus palabras. La polaridad de las palabras puede ser determinada por diferentes métodos, por ejemplo usando una lista de palabras de opinión (Hu y Liu, 2004), utilizando búsquedas en la web (Hatzivassiloglou y Wiebe, 2000), consultando en una base de datos léxica como WordNet (Kamps et al., 2004) o considerando alguna característica lingüística para determinar el sentimiento a nivel de palabra (Ding y Liu, 2007) (Hatzivassiloglou y Mckeown, 1997) (Turney, 2002). Esta aproximación no necesita de una colección de datos etiquetada a priori para un entrenamiento previo, aunque sí de recursos léxicos normalmente dependientes del idioma para determinar la polaridad de las palabras. Aunque ambas aproximaciones tienen ventajas e inconvenientes, nuestro trabajo se engloba en la aproximación basada en OS. Muchos investigadores han guiado sus pasos intentando resolver estos problemas pero aún quedan otros retos que afrontar y abordar, como es la adaptación de la clasificación de opiniones al dominio tratado (Aue y Gamon, 2005). Es en este reto donde centraremos el esfuerzo de este artículo.

Por otra parte, la mayoría de los trabajos en AS tratan con documentos escritos en inglés a pesar de que cada vez es mayor la cantidad de información subjetiva que publican los usuarios de Internet en su propio idioma. Es por esta razón, que la generación y uso de recursos propios en el idioma de los documentos a tratar se esté convirtiendo en un tema crucial para realizar la clasificación de opiniones mediante orientación semántica. Así pues, nuestro artículo está enfocado al AS en español, de manera que los recursos que utilizaremos estarán en este idioma, tanto corpora como lexicones.

Resumiendo, el desarrollo de recursos lingüísticos nuevos es muy importante para seguir progresando en AS. Además, se hace necesario que esos nuevos recursos se implementen en otros idiomas distintos al inglés, como el español por ejemplo. Así, la descripción de un corpus nuevo de opiniones en el dominio turístico, la descripción de un lexicón de palabras con sentimientos dependiente del dominio y unos experimentos que certifiquen la validez de dichos recursos son la principal contribución de este artículo.

El presente artículo se estructura de la siguiente manera: en la sección 2 se describen brevemente otros trabajos relacionados con la clasificación de polaridad en opiniones escritas en español, trabajos que generan nuevos recursos léxicos y algunos trabajos relacionados con la adaptación al dominio en AS. En la sección 3 se explican los diferentes recursos utilizados, así como la metodología utilizada para la generación del nuevo lexicón adaptado al dominio. En la sección 4 se muestran los experimentos realizados y se discuten los resultados obtenidos. Por último, se exponen las conclusiones y el trabajo futuro.

2 Trabajos relacionados

Centrándonos en los trabajos realizados sobre AS, a continuación se presentan los más relevantes en un idioma distinto del inglés. Como primer trabajo se tiene el de Banea et al. (2008), el cual propone varios enfoques para el análisis de la subjetividad en varios idiomas mediante la aplicación directa de las traducciones de un corpus de opiniones etiquetadas en inglés para el entrenamiento de un clasificador de opiniones en rumano y español. Este trabajo muestra que la traducción automática es una alternativa viable para la construcción de recursos y herramientas para el análisis de la subjetividad en un idioma distinto al inglés. Brooke et al. (2009) presentan varios experimentos relacionados con recursos en español e inglés. Llegan a la conclusión de que, aunque las técnicas de aprendizaje automático pueden proporcionar un buen rendimiento, es necesario integrar el conocimiento y los recursos específicos del idioma con el fin de lograr una mejora notable. Se proponen tres enfoques: el primero utiliza los recursos de forma manual y automáticamente generados para el español. El segundo aplica aprendizaje automático sobre un corpus español y el último

traduce los corpus del español al inglés y luego aplica SO-CAL, (Semantic Orientation Calculator), una herramienta desarrollada por ellos mismos (Taboada et al., 2011). Martínez-Cámara et al. (2011) emplean un corpus de críticas de cine llamado MuchoCine (Cruz et al., 2008) para clasificar opiniones escritas en español usando un enfoque supervisado, y Martín-Valdivia et al. (2013) empleando el mismo corpus de cine en español y generando el corpus paralelo en inglés MCE realiza una combinación de la clasificación supervisada sobre ambos corpus y una clasificación no supervisada integrando SentiWordNet (Esuli and Sebastiani, 2006) sobre el corpus en inglés.

Para realizar la clasificación de la polaridad siguiendo un enfoque basado en orientación semántica, muchos autores usan o generan recursos léxicos en el idioma en el que están escritas las opiniones. Así, Taboada y Grieve (2004) ponen a disposición de los investigadores el corpus SFU en inglés, con 400 opiniones distribuidas en 8 categorías, con 25 opiniones positivas y otras 25 negativas cada categoría. Al poco tiempo, generan otro corpus en español siguiendo la misma filosofía, con 8 categorías similares, el corpus SFU en español. En Cruz et al. (2008) se describe la generación de un corpus MC de críticas de cine escritas en español a partir de la página web MuchoCine.com[1]. El corpus cuenta con 1.274 opiniones clasificadas como negativas y 1.351 opiniones clasificadas como positivas. Boldrini et al. (2009) presentan el corpus EmotiBlog que incluye comentarios sobre varios temas en tres idiomas: español, inglés e italiano. En Molina-González et al. (2013) se presenta un nuevo recurso para la comunidad investigadora en AS en español. El recurso llamado iSOL, el cual será utilizado en este artículo, es una lista de palabras de opinión generada a partir del conocido y ampliamente usado lexicón existente en inglés de Bing Liu (Hu and Liu, 2004). En Díaz-Rangel et al. (2014) se proporciona un lexicón de emociones en español compuesto de 2.036 palabras que llevan asociado un factor de probabilidad de uso afectivo (PFA) con respecto al menos una de las emociones básicas: alegría, enfado, tristeza, sorpresa y disgusto.

Por otra parte, como es bien sabido, la orientación semántica de muchas palabras es dependiente del dominio que se trate, existiendo diversos documentos que corroboran este hecho como son Engström (2004), Owsley, Sood y Hammond (2006) y Blitzer, Dredze y Pereira (2007). Existen trabajos más actuales como Dehkharghani et al. (2012), en el que se propone un método para construir un sistema de clasificación de la polaridad dependiente del dominio. El dominio seleccionado por los autores es sobre comentarios de los huéspedes de hoteles. Cada opinión se representa por un conjunto de características independientes del dominio y otro conjunto dependiente del dominio. En Demiroz et al. (2012) se propone un método para adaptar un recurso lingüístico de sentimientos independiente del dominio, como SentiWordNet, a un dominio específico. En Molina-González et al. (2013) se detalla la generación de un recurso léxico basado en listas de palabras de opinión adaptado al dominio de cine. Nuestra propuesta sigue un enfoque basado en corpus, pero en este caso el dominio utilizado es el turístico y concretamente, usaremos un corpus con opiniones extraídas de TripAdvisor para diferentes hoteles de Andalucía. Los buenos resultados obtenidos en los experimentos demuestran que nuestra propuesta es válida independientemente del dominio elegido.

3 Recursos: corpora y lexicones

En esta sección se describe, en primer lugar, el corpus de opiniones sobre hoteles. Este corpus se llama COAH (Corpus of Opinions about Andalusian Hotels) y está disponible libremente[2]. Los lexicones usados para la experimentación son el lexicón iSOL, independiente del dominio, usado en varios trabajos como Molina-González et al. (2013) y el nuevo lexicón eSOLHotel (iSOL enriquecido para el dominio de hoteles) generado a partir del corpus COAH. El corpus usado para probar la bondad del lexicón generado eSOLHotel es el corpus SFU en español[3], en particular, las opiniones pertenecientes a la categoría de hoteles.

3.1 Corpus COAH

Para compilar un corpus de opiniones es muy importante saber elegir la fuente de dichos

[1] http://www.muchocine.net/

[2] http://sinai.ujaen.es/coah

[3] https://www.sfu.ca/~mtaboada/download/downloadCorpusSpa.html

datos. En nuestro caso, hemos intentado satisfacer los siguientes requisitos:

- Debe haber gran cantidad de opiniones y éstas deben ser escritas por usuarios de los hoteles.
- Cada opinión debe estar valorada por el propietario de dicha opinión.
- El portal web debe ser un portal confiable en el dominio de hoteles.
- Debe ser un portal prestigioso internacionalmente en la búsqueda de información sobre hoteles.

Después de estudiar varios portales web, nuestra elección final fue TripAdvisor[4]. El corpus generado consiste en una colección de opiniones escritas por usuarios no necesariamente profesionales. Este hecho incrementa la dificultad de la tarea, porque los textos pueden no ser gramaticalmente correctos, incluso contener palabras mal escritas o expresiones informales. Se han seleccionado solo hoteles andaluces. Por cada provincia de Andalucía (Almería, Cádiz, Córdoba, Granada, Jaén, Huelva, Málaga and Sevilla), se han elegido 10 hoteles, siendo 5 de ellos de valoración muy alta y los otros 5 con las peores valoraciones, para obtener las mínimas opiniones neutras en el corpus. Todos los hoteles seleccionados deben tener al menos 20 opiniones escritas en español en los últimos años. Finalmente, se han obtenido 1.816 opiniones.

Las opiniones están valoradas en una escala de 1 a 5. El valor 1 significa que el autor manifiesta una opinión muy negativa sobre el hotel, mientras que una puntuación de 5 quiere decir que el autor tiene muy buena opinión sobre el hotel. Los hoteles con valor 3 se pueden catalogar como hoteles neutros, ni buenos ni malos, y por tanto, difíciles de clasificar. En la Tabla 1 se muestra el número de opiniones por valoración.

Valoración	Número de opiniones
1	312
2	199
3	285
4	489
5	531
Total	1.816

Tabla 1: Distribución por valoración

#Opiniones	1.816
#Hoteles	80
Media de opiniones por hotel	22,7
#Palabras	264.303
#Frases	9.952
#Adjetivos	17.800
#Adverbios	15.219
#Verbos	38.590
#Sustantivos	53.640
Media de palabras por frase	26,55
Media de palabras por opinión	145,54
Media de adjetivos por opinión	9,80
Media de adverbios por opinión	8,38
Media de verbos por opinión	21,25
Media de sustantivos por opinión	29,54

Tabla 2: Estadísticas de COAH

En la Tabla 2 se muestran algunas características del corpus. De los metadatos mostrados en la Tabla 2, se puede resaltar que las opiniones tienen una media de 145 palabras suficientes para dar la opinión subjetiva sin implicarse en descripciones objetivas fuera de nuestro estudio. Las páginas web extraídas fueron transformadas en ficheros xml (uno por hotel). Cada fichero xml tiene 20 opiniones. Cada opinión tiene dos tipos de información, una sobre el hotel y otra sobre la opinión del huésped del hotel.

A partir de los ficheros xml se genera un documento que solo alberga la valoración de un hotel específico, el título y la opinión. Para los experimentos se descartan aquellas opiniones neutras, es decir, con valoración 3. El resto de opiniones son catalogadas como positivas si su valoración es 4 ó 5, y negativas si su valoración es 1 ó 2. Por tanto, la clasificación binaria de las opiniones sobre hoteles del corpus COAH es la que se muestra en la Tabla 3.

Clases	Número de opiniones
Positiva	1.020
Negativa	511
Total	1.531

Tabla 3: Clasificación binaria del corpus COAH

En las Figuras 1 y 2 se muestra un ejemplo de un hotel, en XML y en formato texto.

[4] http://www.tripadvisor.es

<ID>1</ID>
<Nombre>Alcazaba Mar Hotel</Nombre>
<Categoria>4</Categoria>
<Dirección>Juegos del Argel, Urbanizacion E1 Toyo |Cabo de Gata </Dirección>
<CódigoPostal>04131</CódigoPostal>
<localidad>Retamar</localidad>
<Provincia>Almería</Provincia>
<País>España</País>
<Viajero>----------</Viajero>
<Localidad_Viajero>-------
</Localidad_Viajero>
<Valoración>3</Valoración>
<Título>"Adecuada la calidad al precio del hotel"</Título>
<Opinión>Acabamos de llegar del hotel. La verdad es que nos fuimos con mucho miedo por los comentarios escritos aquí. Nuestra opinion es que es un hotel comodo, tiene piscina buena, animacion excelente, y un personal muy amable. Quizas lo mas tenido en cuenta es el buffet..... </Opinión>
<Fecha_TipoViajero>Se alojó el Agosto de 2012, viajó con la familia</Fecha_TipoViajero>
<Relación_calidad-precio>3</Relación_calidad-precio>
<Ubicación>2</Ubicación>
<Calidad_del_sueño>3</Calidad_del_sueño>
<Habitaciones>3</Habitaciones>
<Limpieza>3</Limpieza>
<Servicio>4</Servicio>

Figura 1: Ejemplo de un hotel en el corpus COAH

Valoración	Título	Opinión
1	"Un hotel digno de mención!"	Como bien les comenté a los propietarios a la hora de abandonar el hotel, no dudaré un momento en recomendar una y otra vez el Hotel Albero de Granada. Su situación respecto del centro de Granada no es la mejor, pero para nuestros propósitos era perfecto (escapada de fin de semana con visita a la Alhambra). Se encuentra en la carretera de..... ..... Si vuelvo a Granada no dudaré en hospedarme en el mismo hotel. Muchas gracias por todo!!

Figura 2: Fragmento de una opinión del corpus COAH

3.2 Corpus SFU

Para realizar los experimentos, se elige parte del corpus SFU Corpus. El Corpus SFU se compone de opiniones de productos en inglés y español. La versión en inglés (Taboada y Grieve, 2004) tiene 400 opiniones (200 positivas y 200 negativas) de productos comerciales descargados de la web Epinions[5] en el año 2004. Se divide en ocho categorías: libros, coches, ordenadores, utensilios de cocina, hoteles, películas, música y teléfono. Cada categoría incluye 25 opiniones positivas y 25 opiniones negativas. Posteriormente, los autores de SFU Corpus hacen disponible la versión española del corpus[6], con el objetivo de ofrecer un corpus comparable para las siguientes investigaciones. Las opiniones en español se dividen en ocho categorías similares, y también cada categoría tiene 25 opiniones positivas y 25 opiniones negativas. En este caso, las opiniones se descargan desde la web Ciao.es[7]. Para realizar nuestros experimentos se eligen las opiniones de la categoría hoteles.

3.3 Lexicón iSOL

Este recurso fue generado a partir del lexicón en inglés de Bing Liu (Hu y Liu, 2004) traduciéndolo automáticamente al español, obteniendo el recurso SOL (Spanish Opinion Lexicon). Posteriormente, la lista fue revisada manualmente. La lista final de palabras de opinión se llama iSOL (improved SOL). El lexicón iSOL se compone de 2.509 palabras positivas y 5.626 palabras negativas, en total, el lexicón español tiene 8.135 palabras polarizadas. Este recurso fue evaluado satisfactoriamente en Molina-González et al. (2013) usando el corpus MuchoCine (Cruz et al., 2008). Los resultados mostraron que el uso de la lista mejorada de palabras polarizadas puede ser una buena estrategia para la clasificación de polaridad no supervisada.

3.4 Lexicón eSOLHotel

El lexicón iSOL es de propósito general, sin embargo, el AS es una tarea con un cierto grado de interrelación con el dominio tratado. Dentro de los enfoques seguidos para la compilación de un conjunto de palabras de opinión, el más adecuado para obtener términos con carga semántica dependientes del dominio es el que se conoce como el enfoque basado en corpus (Kanayama y Nasukawa, 2006).

Tomando como referencia el lexicón iSOL, se ha generado una lista de palabras de opinión para el dominio de hoteles. Para la generación

[5] http://www.epinions.com/
[6] https://www.sfu.ca/~mtaboada/download/downloadCorpusSpa.html
[7] http://www.ciao.es/

de la lista de palabras de opinión se ha seguido el enfoque basado en corpus. El elemento clave del enfoque basado en corpus es el uso de una colección de documentos etiquetados según su polaridad. El corpus español seleccionado para el proceso es COAH. Hemos seguido el mismo supuesto que Du et al. (2010), es decir, una palabra debe ser positiva (o negativa) si aparece en muchos documentos positivos (o negativos). Por lo tanto, hemos calculado la frecuencia de la palabra en cada clase de documentos (positivos y negativos). De una manera manual y subjetiva, se han seleccionado 166 palabras positivas y 131 palabras negativas, que cumplen los requisitos de aparecer en una clase más veces que en la otra y tener una orientación positiva o negativa. Por lo tanto, se añadieron las 297 palabras más frecuentes que aún no figuraban en la lista iSOL a la lista final obteniendo un total de 8.432 palabras indicadoras de opinión (2.675 positivas y 5.757 negativas). Esta nueva lista de integración de la información del corpus ha sido llamada eSOLHotel (SOL enriquecido y adaptado al dominio Hotel). En la siguiente Tabla 4 se muestran algunas de las palabras que han sido añadidas.

Palabras positivas	Palabras negativas
ensueño	asqueroso
luminoso	cucaracha
coqueto	desconchones
comodisima	humedades
intachable	mejorable
remodelado	reclamaciones
pasada	tugurio
supercentrico	zulo

Tabla 4: Palabras positivas y negativas añadidas al lexicón eSOLHotel

4 Experimentos y resultados

Antes de llevar a cabo los experimentos, a las opiniones de hoteles del corpus SFU se les ha realizado un *preprocesamiento* con el fin de tener en cuenta los mismos criterios que se han utilizado en la generación de los lexicones iSOL y eSOLHotel. Por ejemplo, las letras mayúsculas se han cambiado a minúsculas, a las vocales acentuadas se les ha quitado el acento y los caracteres especiales han sido separados de las palabras, para aislar dichas palabras.

Para decidir si una opinión se considera positiva o negativa, seguimos un simple método basado en la cuenta del número de palabras incluidas en las listas iSOL y eSOLHotel encontradas en las opiniones de hoteles del corpus SFU etiquetado en español. Así, nuestro método clasifica la opinión como positiva si el número de palabras positivas encontradas es igual o mayor que el número de palabras negativas encontradas, o como negativa en el resto de casos.

En la Tabla 5 se muestran los resultados obtenidos en la categoría de hoteles del corpus SFU en español usando los lexicones iSOL (independiente del dominio) y eSOLHotel (adaptado al dominio de hoteles).

Lexicón	Precisión	Valor F1	Exactitud
iSOL	77,41%	73,52%	70,0%
eSOLHotel	84,72%	81,22%	78,0%

Tabla 5: Resultados obtenidos en la clasificación binaria de corpus SFU usando iSOL y eSOLHotel

Los resultados que se muestran en la Tabla 5 confirman nuestra hipótesis de partida, es decir, que la inclusión de información del dominio en una lista de palabras de opinión genérica mejora los resultados de la clasificación de la polaridad. El porcentaje de mejora en la exactitud que se ha obtenido con la inclusión de información del dominio ha sido de un 11,43%. Siguiendo una metodología muy simple, como la que se ha descrito, se ha obtenido una mejora muy importante.

Con el fin de profundizar en el estudio de la bondad de la metodología seguida para la inclusión de información del dominio, se ha construido un clasificador supervisado. Para ello, se ha aplicado a los documentos un algoritmo de normalización morfológica basado en la eliminación de prefijos y sufijos, lo que en el ámbito del Procesamiento del Lenguaje Natural se conoce como *stemmer*. El algoritmo de *stemming* empleado ha sido el de Porter para español. Tras este proceso, los documentos se han representado como vectores de *unigramas* ponderados por el índice de relevancia TF-IDF. Por tanto, las características que recibirá como entrada el algoritmo de aprendizaje automático serán únicamente el valor TF-IDF de los *unigramas* de los documentos. Por último, se ha realizado una validación cruzada con el

algoritmo SVM. Los resultados que se han obtenido son un 82% y un 82,71% de exactitud y valor F1 respectivamente. De nuevo, los resultados de la Tabla 5 indican la bondad de la metodología presentada en el artículo, dado que la diferencia de valor F1 entre SVM y eSOLHotel es solo de un 1,83%. Por lo tanto, la pérdida de exactitud es tan mínima que puede considerarse aconsejable el uso de la lista en lugar del método supervisado, ya que en este caso no se necesitaría de un modelo de aprendizaje automático previamente entrenado.

5 Conclusiones y trabajos futuros

En este artículo se ha presentado una metodología de adaptación de un lexicón de palabras de opinión a un dominio concreto. Para ello se ha tomado un corpus de opiniones de hoteles como referencia (COAH), se han calculado la frecuencia de los términos que componen el corpus y se han seleccionado las palabras de opinión más representativas del corpus. La metodología se ha evaluado con las opiniones de hoteles del corpus SFU en español. Los resultados que se han obtenido (Tabla 5) ponen de manifiesto la bondad de la metodología y nos animan a seguir perfeccionando la metodología de adaptación al dominio.

El sistema de clasificación se puede todavía mejorar aún más. Como trabajo futuro se va a incluir un tratamiento de la negación basado en reglas lingüísticas específico para español. Este nuevo elemento del sistema nos va a permitir clasificar correctamente las opiniones negativas expresadas con términos positivos negados.

Agradecimientos

Esta investigación ha sido parcialmente financiada por el Fondo Europeo de Desarrollo Regional (FEDER), el proyecto ATTOS (TIN2012-38536-C03-0) del Gobierno de España y el proyecto AORESCU (P11-TIC-7684 MO) del gobierno autonómico de la Junta de Andalucía. Por último, el proyecto CEATIC (CEATIC-2013-01) de la Universidad de Jaén también ha financiado parcialmente este artículo.

Bibliografía

Aue, A. y M. Gamon. 2005. Customizing sentiment classifiers to new domains: A case study. En *Proceedings of Recent Advances in Natural Language Processing (RANLP)*.

Banea, C., R. Mihalcea, J. Wiebe, y S. Hassan, S. 2008. Multilingual subjectivity analysis using machine translation. En *Proc. of the conference on empirical methods in natural language processing*, páginas 127–135. ACL.

Blitzer, J., M. Dredze, y F. Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. En *Proceedings of the Association for Computational Linguistics (ACL)*.

Boldrini, E., A. Balahur, P. Martínez-Barco, y A. Montoyo. 2009. Emotiblog: an annotation scheme for emotion detection and analysis in non-traditional textual genres. En *DMIN*, páginas 491–497. CSREA Press.

Brooke, J., M. Tofiloski, y M. Taboada. 2009. Cross-linguistic sentiment analysis: From English to Spanish. En *Proceedings of the International Conference RANLP-2009*, páginas 50–54. ACL.

Cruz, F.L., J.A. Troyano, F. Enriquez, y J. Ortega. 2008. Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. *Procesamiento de Lenguaje Natural*, Volumen 41, páginas 73-80.

Dehkharghani, R., B. Yanikoglu, D. Tapucu, y Y. Saygin. 2012. Adaptation and use of subjectivity lexicons for domain dependent sentiment classification. En *Data Mining Workshops, 2012 IEEE 12th International Conference on*, páginas 669-673.

Demiroz, G., B. Yanikoglu, D. Tapucu, y Y. Saygin. 2012. Learning domain-specific polarity lexicons. En *Data Mining Workshops, 2012 IEEE 12th International Conference on*, páginas 674-679.

Díaz Rangel, I., G. Sidorov y S. Suárez-Guerra. 2014. Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. *Onomazein, 29, 23 p*

Ding, X. y B. Liu. 2007. The utility of linguistic rules in opinion mining. En *Proc. of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, páginas 811–812.

Du, W.T., S. Cheng, y X. Yun. 2010. Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon. En *Proc. of ACM International Conference on Web search and data mining.*

Engström, C. 2004. Topic dependence in sentiment classification. *Master's thesis*, University of Cambridge.

Esuli, A. y F. Sebastiani. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. En *Proceedings of Language Resources and Evaluation (LREC).*

Hatzivassiloglou, V. y K. McKeown. 1997. Predicting the semantic orientation of adjectives. En *Proceedings of the eighth conference on European chapter of the association for computational linguistics*, páginas 174–181.

Hatzivassiloglou, V. y J. Wiebe, J. 2000. Effects of adjective orientation and gradability on sentence subjectivity. En *Proceedings of the international conference on computational linguistics (COLING)*, páginas 299–305.

Hu, M. y B. Liu. 2004. Mining and summarizing customer reviews. En *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, Washington, USA, páginas 168-177.

Kamps, J., M. Marx, R.J. Mokken, y M. de Rijke. 2004. Using WordNet to measure semantic orientations of adjectives. En *LREC, European Language Resources Association.*

Kanayama, H. y T. Nasukawa. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, páginas 355–363. ACL.

Liu, B. 2012. Sentiment analysis and opinion mining. synthesis lectures on human language technologies. *Morgan and Claypool Publishers.*

Martín-Valdivia, M.T., E. Martínez-Cámara, J.M. Perea-Ortega, y L.A. Ureña-López. 2013. Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. *Expert Systems with Applications*, 40 (10), páginas 3934–3942.

Martínez-Cámara, E., M.T. Martín-Valdivia, y L.A. Ureña-López. 2011. Opinion classification techniques applied to a Spanish corpus. *Proceedings of the 16th international conference on Natural language processing and information systems, NLDB'11*, páginas 169–176.

Molina-González, M. D., E. Martínez-Cámara, M.T. Martín-Valdivia, y J.M. Perea-Ortega. 2013. Semantic Orientation for Polarity Classification in Spanish Reviews. *Expert Systems with Applications;* 40(18), páginas 7250-7257.

Owsley, S., S. Sood, y K.J. Hammond. 2006. Domain specific affective classification of documents. En *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, páginas 181–183.

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. Thums up?: Sentiment Analysis classification using machine learning techniques. En *Proceedings of the ACL-02 Conference on Empirical methods in natural language processing. Stroudsburg, PA, USA: Association for Computational Linguistics;* 10:79-86.

Taboada, M., J. Brooke, M. Tofiloski, K.D. Voll, y M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2), 267–307.

Taboada, M. y J. Grieve 2004. Analyzing appraisal automatically. *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text,* páginas 158 - 161. Stanford University, CA.

Tsytsarau, M. y T. Palpanas. 2012. Survey on mining subjective data on the web. *Data Mining and Knowledge. Discovery*. 24, 3 478-514.

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, Stroudsburg, PA, USA*, páginas 417-424.

Procesamiento del Lenguaje Natural, Revista nº 54, marzo de 2015, pp 29-36 recibido 10-11-14 revisado 27-01-15 aceptado 12-02-15

Choosing a Spanish Part-of-Speech tagger for a lexically sensitive task*

Selección de un etiquetador morfosintáctico primando la precisión en las categorías léxicas

Carla Parra Escartín
University of Bergen
Bergen, Norway
carla.parra@uib.no

Héctor Martínez Alonso
University of Copenhagen
Copenhagen, Dennmark
alonso@hum.ku.dk

Resumen: En este artículo se comparan cuatro etiquetadores morfosintácticos para el español. La evaluación se ha realizado sin entrenamiento ni adaptación previa de los etiquetadores. Para poder realizar la comparación, los etiquetarios se han convertido al etiquetario universal (Petrov, Das, and McDonald, 2012). También se han comparado los etiquetadores en cuanto a la información que facilitan y cómo tratan características intrínsecas del idioma español como los clíticos verbales y las contracciones.
Palabras clave: Etiquetadores morfosintácticos, evaluación de herramientas, lingüística de corpus

Abstract: In this article, four Part-of-Speech (PoS) taggers for Spanish are compared. The evaluation has been carried out without prior training or tuning of the PoS taggers. To allow for a comparison across PoS taggers, their tagsets have been mapped to the universal PoS tagset (Petrov, Das, and McDonald, 2012). The PoS taggers have also been compared as regards the information they provide and how they treat special features of the Spanish language such as verbal clitics and portmanteaux.
Keywords: Part-of-Speech taggers, tool evaluation, corpus linguistics

1 Introduction

Part-of-Speech (PoS) taggers are among the most commonly used tools for the annotation of language resources. They are often a key preprocessing step in many Natural Language Processing (NLP) systems. When using a PoS tagger in a workflow, it is important to know the impact of its error rate on any modules that use its output (Manning, 2011).

In this paper, we compare four different PoS taggers for Spanish. Our goal is to build an NLP system for knowledge extraction from technical text, and this requires very good performance on lemmatisation and right PoS assignment. Inflectional information is not relevant for our purposes. Instead of choosing a PoS system based solely on its reported performance, we benchmark the output of the 4 candidate systems against a series of metrics that profile their behaviour when predicting lexical and overall PoS tags, as well as the final quality of the resulting lemmatisation.

As the taggers had different tagsets, and we were only interested in retrieving the coarse PoS tag (*buys*_verb vs. *buys*_verb_3ps), we have mapped the tagsets to the universal PoS tagset proposed by Petrov, Das, and McDonald (2012).

The remainder of this paper is organised as follows: Section 2 discusses available PoS taggers for Spanish and describes them briefly. In Section 3 the different challenges encountered when comparing the four PoS taggers are discussed. Section 4 discusses the differences across the tagsets used by each PoS tagger and Section 5 shows the evaluation of each PoS tagger compared with the other three. Section 6 offers a summary.

2 Part-of-Speech taggers available for Spanish

For Spanish, several PoS tagger initiatives have been reported (Moreno and Goñi, 1995; Márquez, Padró, and Rodríguez, 2000; Car-

* The authors thank the anonymous reviewers for their valuable comments and the developers of the taggers we have covered in this article for making them available.

ISSN 1135-5948 © 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

reras et al., 2004; Padró and Stanilovsky, 2012, i.a.). The reported accuracies for these PoS taggers are reported to be 96–98%.

However, not all of these tools were available for our tests. One of the existing PoS taggers, the GRAMPAL PoS tagger (Moreno and Goñi, 1995) was not included in this comparison because it is not downloadable and it does not seem accessible via a web service[1].

To our knowledge, four PoS taggers for Spanish are the TreeTagger (*TT*) (Schmid, 1994), the IULA TreeTagger (*IULA*) (Martínez, Vivaldi, and Villegas, 2010), the FreeLing PoS tagger (*FL*) (Padró and Stanilovsky, 2012), and the IXA PoS tagger (*IXA*) (Agerri, Bermudez, and Rigau, 2014). Two of them (IULA and FL) are also available as web services developed during the PANACEA project. The fourth PoS tagger was recently released within the IXA pipes. The study reported in this paper compares these four PoS taggers.

2.1 Default TreeTagger TT

TT provides 22 already trained models for 18 languages and new models can be created by retraining the tagger with new data.

We used the already trained model for Spanish which is available on its website. This model was trained on the Spanish CRATER corpus and uses the Spanish lexicon of the CALLHOME corpus of the LDC.

Prior to tagging the text, the tool tokenises it. The tokeniser does not have specific rules for Spanish.

In the output of this tagger, every line contains one wordform with its PoS tag and its lemma (i.e. citation form). TT PoS tags do not contain inflectional information (e.g. tense, number, gender, etc.). This tagset is the most similar to the universal PoS tagset.

When the tagger fails to assign a lemma to a specific wordform, the value for the lemma is *<unknown>*. Nevertheless, a PoS tag is assigned to an unknown word. Examples 1-3 show words with unknown lemmas and their assigned PoS tags.

(1) NÖ NC <unknown>

(2) WFG NP <unknown>

(3) plurifamiliares ADJ <unknown>

TT concatenates Multiword Expressions (MWEs). Their wordforms are listed with whitespaces as they occur in the text, while their lemmas are joined by means of tildes. Examples 4 and 5 show MWEs tagged by TT.

(4) de conformidad con PREP de∼conformidad∼con

(5) junto con PREP junto∼con

2.2 IULA TreeTagger (IULA)

The IULA is an instance of the TT trained on the IULA technical corpus (IULACT). Additionally, each file undergoes a preprocessing step prior to tagging. This preprocessing step is described in detail in Martínez, Vivaldi, and Villegas (2010). It comprises the following tasks:

1. Sentence-boundary detection;
2. General structure-marking;
3. Non-analyzable element recognition;
4. Phrase and loanword recognition;
5. Date recognition;
6. Number recognition;
7. Named Entity (NE) recognition.

These tasks were introduced in what Martínez, Vivaldi, and Villegas (2010) call a *text handling* module. This module was developed in order to solve potential sources of errors prior to tagging with the aim of improving the overall quality of the PoS tagging process. The whole toolset is available through a web service where one can upload the corpus to be tagged and download the tagged corpus upon completion of the task.

Unlike the TT instance discussed in Subsection 2.1, the IULA PoS tagset provides inflectional information for the relevant PoS. The tagset is partially based on the EAGLES tagset for Spanish, and includes more fine-grained information.

When the tagger fails to assign a lemma to a specific wordform, instead of assigning to it the value *<unknown>* as TT does, IULA assigns wordforms as lemmas. Example 6 shows the tagging of the unknown word *plurifamiliares*. *Plurifamiliares* is in plural, and its lemma should be the singular wordform *plurifamiliar* but instead, the plural wordform is used.

(6) plurifamiliares JQ–6P plurifamiliares

Special elements such as MWEs are treated in a different way. The previous MWE examples 4 and 5 do not appear concatenated, but are tagged as separate words. MWEs such as dates or names are lemmatised with underscores. Examples 7-9 show some

[1] There is an available online demo limited to 5000 words.

tagged MWEs. Furthermore, as a result of the preprocessing step, the IULA adds additional xml-style tags to such elements.

(7) 18 de diciembre del 2001 T 18_de_diciembre_del_2001

(8) Baja Austria N4666 Baja_Austria

(9) Promoción MH-NEU N4666 Promoción_MH-NEU

2.3 FreeLing (FL)

FreeLing is an open source suite of language analysers. It offers a wide variety of services for several languages. Among these analysers are tokenisers, sentence splitters, morphological analysers, PoS taggers, etc. The PoS tagger has two different flavours, a hybrid tagger called *relax*, which combines statistical and manually crafted grammatical rules, and a model based on the Hidden Markov Model (HMM) similar to the TnT proposed by Brants (2000). In both cases, the tagger was trained with the *LexEsp* corpus (Sebastián, Martí, and Carreiras, 2000). Again, the web service offered by the PANACEA project was used. Since in the web service only the HMM model was deployed, this is the tagging model we have used in this paper.

FL displays first the wordform, followed by the lemma and then the PoS tag. It uses the EAGLES tagset for Spanish, which, similarly to the IULA tagset, also includes more fine-grained inflectional information.

Whenever FL analyses a sequence as a MWE, it displays both the wordform and the lemma joined with underscores. Examples 10 and 11 show some tagged MWEs.

(10) Baja_Austria baja_austria NP00G00

(11) Promoción_MH-NEU promoción_mh-neu NP00V00

Another peculiarity is that all lemmas are lowercased regardless whether they correspond to a proper noun, an abbreviation or something else. This can be observed in Example 10, where the lemma for the proper noun *Baja Austria* ([EN]: Lower Austria) is lowercased to *baja_austria*.

Finally, dates are also treated differently in FL. Their wordform is the date itself joined by underscores, and their "lemma" is the same date converted to a numerical format, where month names are converted to their corresponding number. Examples 12 and 13 show this.

(12) 18_de_diciembre_del_2001 [??:18/12/2001:??.??:??] W

(13) 15_de_octubre_del_2002 [??:15/10/2002:??.??:??] W

2.4 IXA pipes (IXA)

The IXA pipes are "ready to use NLP tools" (Agerri, Bermudez, and Rigau, 2014) developed by the IXA NLP Group. Among the available tools there are a tokeniser and a PoS tagger for Spanish. The PoS tagger requires not only that the text is previously tokenised, but also, that it is in the NLP Annotation Format (NAF) (Fokkens et al., 2014). These requirements are met by using the provided tokeniser and piping it to the PoS tagger.

IXA has been trained and evaluated using the Ancora corpus (Taulé, Martí, and Recasens, 2008). Two PoS tagging models are available: one based on the Perceptron algorithm (Collins, 2002), and another one based on Maximum Entropy (Ratnaparkhi, 1999). We use the default model for Spanish, which is the Perceptron.

Its output format is xml-based and thereby differs from the taggers previously discussed in this paper. The resulting document is tagged in NAF and has a header specifying the tools that have been used, when they were used, how long the process has taken and the version of the tool used. Next, the tokenised text appears.

For each tokenised wordform, the tool provides its NAF required attributes for word *id* and the sentence where it appears (*sent*), as well as the optional attributes *para*, *offset*, and *length*, which correspondingly refer to the paragraph the word belongs to, the offset in number of characters and its length in number of characters.

Where the tokenised text ends, a new section starts with PoS and lemma information about each word as identified by its *id*. For each term, the following attributes are provided:

1. *id*: The term id. This is the only required attribute, all the other attributes are optional.
2. *type*: Whether the term belongs to an open PoS (e.g. nouns), or a closed one (e.g. prepositions).
3. *lemma*: The wordform lemma.
4. *pos*: The Part of Speech of the wordform.
5. *morphofeat*: the PoS tag assigned to the form, containing inflectional information. IXA uses the same tagset as FL:

the EAGLES tagset for Spanish.

MWEs are signalled by the sub-element *...*. When the PoS tagger identifies a MWE, the sub-element ** will have several *<target>* subelements. *<target>* subelements refer to the wordform ids assigned in the text part of the document. In our test, IXA failed to identify any MWE.

3 Challenges

Comparing and evaluating four different PoS taggers is not a straightforward task. Differences in their tagsets, output formats and tokenisation processes have to be addressed. Prior to the PoS tagging process, the text has to be tokenised. As pointed out by Dridan and Oepen (2012), tokenisation is often regarded as a solved problem in NLP. However, the conventions used in the tokenisation task have a direct impact in the systems which subsequently use the tokenised text. In recent years, several authors have highlighted the importance of tokenisation and researched new tokenisation strategies (Dridan and Oepen, 2012; Fares, Oepen, and Zhang, 2013; Orosz, Novák, and Prószéky, 2013, i.a.).

In our study, each PoS tagger had its own tokeniser either as an integrated component (TT, IULA and FL), or available as a separate tool to be piped to the PoS tagger (IXA). Each tagger subsequently tagged this text, following its internal tokenisation.

At this stage, there are also two particular features of the Spanish language that may be handled differently by the tokenisers and/or the PoS taggers:

- The portmanteau (contracted) wordforms *al* and *del* (cf. 3.1);
- Verbal clitics, which are attached to verbs without hyphenation (cf. 3.2).

Finally, an additional challenge is the way in which each tagger detects and tags MWEs, such as named entities, proper names, dates, and complex prepositions.

3.1 Portmanteaux in Spanish

To a certain extent, the portmanteaux *al* and *del* are not difficult to tackle. They are the result of contracting the Spanish prepositions *a* and *de* with the determined masculine singular article *el*. Thus, *a* + *el* results in *al* and *de* + *el* results in *del*. Additionally, *al* can also be used to introduce subordinated infinite clauses in Spanish (eg. *al pasar*, [EN]: when/while passing). Each PoS tagger however, handles this phenomenon differently.

(a) **TT**: TT has a special tag for each of these wordforms: *PAL* for *al* and *PDEL* for *del*. TT treats the subordinated conjunction reading using a third tag: *CSUBI*. Thus, TT does not split these wordforms and handles them by using specific tags.

(b) **IULA**: The IULA assigns to these wordforms a double tag *P_AMS*, thus providing information from each of the components but joining the tags with an underscore. The lemmas are assigned correspondingly, retrieving the preposition and the article as separate items but joining these lemmas with an underscore: *a_el* and *de_el*.

(c) **FL**: FL retrieves the preposition and the article undoing the contraction completely. Thus, *al* and *del* become *a el* and *de el* and each word is analysed and tagged separately.

(d) **IXA**: IXA uses a strategy similar to that of the TT and uses one special tag *SPCMS* for contracted prepositions available in EAGLES for both *al* and *del*.

Each PoS tagger takes a different stance on this phenomenon. Two taggers tag the contracted prepositions with special tags (TT and IXA). The other two treat them as separate words and retrieve the underlying non-contracted wordforms (IULA and FL), although they represent them differently.

3.2 Verbal clitics

Verbal clitics are morphemes with the syntactic characteristics of a word which depend phonologically on another word. In Spanish, clitics are attached directly to the verb without hyphenation and fulfil a pronominal role. Moreover, it is possible to have two clitic pronouns, one referring to the direct object and the other to the indirect object. As a result of this agglutinative process, some wordforms will require an acute accent to comply with the Spanish orthographic rules. For instance, the wordform *enviármelo* ([EN]: 'send it to me') is composed of the verb *enviar* ([EN]: 'send') to which the clitics *me* ([EN]: 'me') and *lo* ([EN]: 'it') are attached.

Clitic handling is a challenge for PoS taggers. As Martínez, Vivaldi, and Villegas (2010) point out, "the appearance of an acute accent in some wordforms makes a brute-

force stemming difficult". They account for 32 wordforms with pronominal clitics for an infinite verb like *dar* ([EN]: 'give') and explain that as per now the verbal wordforms with clitics are kept in the lexicon of their application to determine if they belong to the language or not. In the four PoS taggers compared in the present paper, different strategies are used.

(a) **TT**: TT assigns special tags to verbs in which clitics are incorporated. It has three different tags available, depending on the verb wordform to which the clitics are attached: *VCLIinf* (infinitive), *VCLIger* (gerund), and *VCLIfin* (finite wordform). For instance, *utilizarla* ([EN]: use it) is assigned the tag *VCLIinf*.

(b) **IULA**: The IULA uses a different strategy. It separates the verb from the clitic, and analyses them as two different words. However, both the wordform and the lemma have additional information attached to them by means of underscores.

(c) **FL**: Like the IULA FL separates the verb from the clitic and analyses them separately. However, no additional marking is used to indicate that the verb and the clitic are one word.

(d) **IXA**: IXA ignores clitics and assigns to verbs with clitics the same tag that the verb would have without the clitic. Thus, *enviármelo* is assigned the tag *VMN0000*, which corresponds to the infinitival form of a main verb.

In conclusion, verbs with clitics are handled in different ways by the different taggers under investigation. While IXA seems to ignore the clitics, all the other taggers handle them differently. The TT has its own tag for this kind of phenomena, FL splits the verb and the clitic, and the IULA uses a mixed approach by splitting the verb but adding additional information to the wordform and the lemma.

4 *Tagset comparison*

The use of different tagsets, together with the other challenges previously discussed in section 3 makes the comparison of PoS taggers a challenging task. Of the four PoS taggers which we have investigated, only FL and IXA share the same tagset, while the others use different tagsets. Each tagger does not only provide a different granularity of categorial distinctions and features, but also treats intrinsic linguistic phenomena differently.

PoS Tagger	Tagset	Tags	Granularity
TT	TT ES	77	low
IULA TT	IULA tagset	435	high
FL	EAGLES	577	high
IXA	EAGLES	577	high

Table 1: PoS taggers tagset comparison.

Table 1 shows the differences across the tagsets used by each of the PoS taggers. A full comparison of the output of four different tools with such big differences regarding the number of tags and the information encoded in such tags could be a very tedious and inaccurate task. We were only interested in retrieving the coarse PoS, the best approach to map the tagsets seemed to be to map each tagset to the universal PoS tagset. Petrov, Das, and McDonald (2012) distinguish twelve PoS tags: "NOUN (nouns), VERB (verbs), ADJ (adjectives), ADV (adverbs), PRON (pronouns), DET (determiners and articles), ADP (prepositions and postpositions), NUM (numerals), CONJ (conjunctions), PRT (particles), '.' (punctuation marks) and X (a catch-all for other categories such as abbreviations or foreign words".

We have developed a mapping from each Spanish tagset to the universal PoS tagset[2]. When projecting to the universal PoS tags we loose the inflectional information. Furthermore, past participles have been mapped to the PoS *VERB*, regardless of whether they function as part of a periphrastic verb tense or function as modifiers or predicates in the same way as adjectives. The adjective-participle ambiguity is addressed differently in the annotation guidelines of all training data, and the behaviour of the tagger in this aspect is a consequence of the data it is trained on. This simple approach was chosen in order to avoid manual revision. A similar approach was taken in other cases, such as the verbs with attached clitics and the portmanteaux when these had not been previously preprocessed and split. In these cases, the categories *VERB* and *ADP*, respectively, were used as defaults.

[2]The mappings are available at the Universal-Part-of-Speech-Tags repository.

5 POS tagger performance

Bearing in mind all the discrepancies described in sections 3 and 4, a completely fair comparison of all the PoS taggers against a unique Gold Standard (GS) is not feasible. A measure like accuracy requires the input data to be linearly comparable—i.e. be tokenized using the same convention—and the output data to rely on the same number of labels—i.e. the datasets should be the same.

The ideal downstream evaluation method would be parsing, but it is too sensitive to tokenisation changes to be applicable when comparing taggers that tokenize differently, besides using different tagsets.

Nevertheless, the four different systems rely on the same sentence boundaries, and are thus sentence-aligned. Given the aforementioned limitations, we use a series of metrics that aim to assess the PoS tagging and lemmatisation performance of the systems by measuring matches—i.e. set intersections—at the sentence level.

In addition, we use metrics that assess the similarity between tag distributions (KL divergence) in order to assess whether the bias of each tagger is closer to the desired performance determined by a reference GS, and metrics that evaluate how many of the predicted lemma+PoS appear Spanish wordnet. The wordnet check serves as a downstream evaluation of the joint predicted values and is one of the instrumental parts of the knowledge-extraction system mentioned in Section 1.

Table 2 summarises the main features of each tagger. Only TT fails to provide inflectional information. Portmanteaux, verbal clitics and reflexive verbs are treated in different ways. While IULA and FL split them, TT and IXA do not. Finally, the tagsets differ also greatly in terms of their overall tag number and the number of non-lexical tags. IULA offers the greatest absolute number and greatest proportion of tags dedicated to non-lexical elements.

The discrepancies between the different tagsets can be addressed by mapping them to the universal PoS tagset. However, then we are losing some of the inflectional analyses produced by some of the taggers. Furthermore, a comparison against one Gold Standard might be biased against one of the various choices of handling MWEs and other special features.

	TT	IULA	FL	IXA
Morphosyntactic info	-	✓	✓	✓
Splits portmanteaux	-	✓	✓	-
Splits verbal clitics	-	✓	✓	-
Joins dates	-	✓	✓	-
Reflexive verb lemmatisation	-	✓	✓	-
Tagset size	77	435	577	
Number of non-lexical tags	42	241	173	

Table 2: PoS taggers features.

In order to allow a comparison, we created a GS following a procedure that attempts to minimise the starting bias in favour of a certain system. We chose two short texts from the freely available technical corpus TRIS (Parra, 2012). We processed this material with FL, because this tagger tokenizes most aggressively (cf. 2.1-2.4). MWEs detected by FL where manually split and tagged, aiming at facilitating the evaluation of this GS against the outputs of the other PoS taggers. Then, we converted the FL tagset to the universal PoS tagset (Petrov, Das, and McDonald, 2012) and manually corrected the tagged text. Each tagger was then compared against this GS.

Table 3 summarises the results of the automatic evaluation we carried out for all taggers against the GS. Three of the metrics have two variants; in an *overall* metric ($_o$), we compare the performance across all parts of speech, whereas in a *lexical* metric ($_l$), we only take into consideration the PoS that are tagged as ADJ, NOUN or VERB after projecting onto Petrov's tagset.

The metrics are the following:

- *Matching lemmas ($_o$/$_l$)*: Proportion of lemmas that match with the GS;
- *Matching PoS ($_o$/$_l$)*: Proportion of PoS tags that match. A match is defined as the minimum count of a given PoS tag matching that in the GS;
- *KL total ($_o$/$_l$)*: Kullback-Leibler divergence (Kullback and Leibler, 1951) between the PoS distribution of the system and that of the GS;
- *GS token ratio*: The relation between the amount of tokens and that of the GS. As explained earlier, we have chosen a GS with the most fragmentary tokenisation, so the ratio will always be equal or less to one. The higher the number, the most similar the tokenisation convention of the system is to the GS tokenisation.

- *WN GS matches*: Number of Spanish wordnet (Gonzalez-Agirre, Laparra, and Rigau, 2012) hits for all predicted lemma-PoS combinations in the GS;
- *WN sys matches*: Number of Spanish wordnet hits for all predicted lemma-PoS combinations matching those of the GS; and
- *WN intersection*: Number of Spanish wordnet hits that also appear in the GS.

	TT	IXA	IULA	FL
Matching lemmas$_o$	0.77	0.85	0.7	**0.89**
Matching lemmas$_l$	0.63	0.63	0.66	**0.78**
Matching PoS$_o$	0.86	0.87	0.85	**0.88**
Matching PoS$_l$	0.93	0.93	**0.94**	0.92
KL_o	**0.041**	0.053	0.042	0.063
KL_l	0.0029	0.0095	**0.0005**	0.001
GS token ratio	0.97	**0.99**	0.94	0.93
WN GS matches	402	402	402	402
WN sys matched	378	367	**386**	384
WN intersection	361	350	366	**373**

Table 3: PoS taggers evaluation.

As Table 3 illustrates, FL has the highest proportion of both overall and lexical matching lemmas (cf. *Matching lemmas o* and *Matching lemmas*$_l$ in Table 3). Its highest proportion of lexical matches is remarkable, as the lemmatisation of lexical words is more important (and error-prone) than that of closed grammatical classes such as articles and prepositions.

As previously stated, FL is also the system which uses a more fragmented tokenisation, and this makes it the most similar to the GS in that respect. However, it gets the lowest *GS token ratio* score. This might be because FL joins the lemmas of MWEs with underscores (cf. 2.3). IXA is the tagger which achieves a better score as regards the ratio of tokens in the GS (cf. *GS token ratio* in Table 3). This may be due to the fact that we split all MWEs in our GS. As mentioned earlier, IXA failed at identifying and tagging them in our test files.

KL divergence offers a different perspective on the different systems. The lower the KL divergence between PoS distribution is, the more similar to the GS is the expected prior for a certain PoS. Interestingly, FL has the highest overall KL divergence in spite of it having the highest performance on lemma retrieval, and the second lowest lexical KL divergence. This difference is due to FL having different conventions on the way that some function words are annotated and thus later converted to the universal PoS tags.

With regard to the overall results, FL has the highest PoS overlap (cf. *Matching PoS*$_o$ in Table 3) with the GS, followed by IXA by a close second. The better accuracy on lemma retrieval, paired with the lowest KL divergence on lexical PoS is also reflected in the highest wordnet hit count for FL (cf. *WN intersection* in Table 3). However, the IULA had a better performance when tagging lexical words (cf. *Matching PoS*$_l$ in Table 3). In fact, it manages to correctly tag lexical words, despite not necessarily achieving to lemmatise them correctly. This explains also, why it has the highest WordNet hit count of lemma-PoS combinations (cf. *WN sys matched* in Table 3).

Since our current research focuses on lexical words, the most important measures for our purposes are *Matching PoS*$_l$ and *Matching lemmas*$_l$. IULA may be our best choice when focusing on the assignment of right PoS tags. It performs better than the TT system which we have previously been using. FL, on the other hand, seems to be better for general lemmatisation tasks, regardless of the PoS tag.

6 Conclusions and future work

We have compared four PoS taggers for Spanish and evaluated their outputs against a common GS. For our purposes, we concluded that IULA would be the best choice, followed by FL. It is possible that a combined output of these two taggers may outperform each single tagger. Given the difficulties of combining different tokenisations in a voting scheme for PoS tagging, we leave this question for future work.

Evaluating on a technical text makes the task more difficult for most of the PoS taggers. The better performance of IULA may also be due it being the only tagger trained on a technical corpus. It is not impossible that for more general (i.e. less technical) texts the differences across the different taggers may be smaller.

We have also proposed a method to compare the output of different PoS taggers by mapping their respective tagsets to the universal PoS tagset and evaluating matches at the sentence level. As indicated earlier in Section 1, for our particular purposes no inflectional information was needed, and thus this method was enough. In case a more fine-grained tagging is needed, the universal PoS

tagset may need to be expanded.

This non-linear evaluation method is useful for tasks that depend on lemmatisation and coarse PoS tagging. Nevertheless, the method would have to be expanded for tasks that require inflectional information.

References

Agerri, R., J. Bermudez, and G. Rigau. 2014. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In *LREC'14*. ELRA.

Brants, T. 2000. TnT: A Statistical Part-of-speech Tagger. In *ANLP'00*, pages 224–231, Sheattle, Washington.

Carreras, X., I. Chao, L. Padró, and M. Padró. 2004. Freeling: An Open-Source Suite of Language Analyzers. In *LREC'04*. ELRA.

Collins, M. 2002. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *EMNLP'02*, volume 10, pages 1–8. ACL.

Dridan, R. and S. Oepen. 2012. Tokenization: Returning to a Long Solved Problem a Survey, Contrastive experiment, Recommendations, and Toolkit. In *ACL'12*, volume 2, pages 378–382. ACL.

Fares, M., S. Oepen, and Y. Zhang. 2013. Machine Learning for High-Quality Tokenization Replicating Variable Tokenization Schemes. In *Computational Linguistics and Intelligent Text Processing*, volume 7816. Springer Berlin Heidelberg, pages 231–244.

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*, pages 9–17.

Gonzalez-Agirre, A., E. Laparra, and G. Rigau. 2012. Multilingual Central Repository version 3.0. In *LREC'12*, pages 2525–2529. ELRA.

Kullback, S. and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Manning, C. D. 2011. Part-of-speech Tagging from 97Linguistics? In *CICLing'11*, pages 171–189. Springer-Verlag.

Márquez, L., L. Padró, and H. Rodríguez. 2000. A machine learning approach to POS tagging. *Machine Learning*, (39):59–91.

Martínez, H., J. Vivaldi, and M. Villegas. 2010. Text handling as a web service for the IULA processing pipeline. In *LREC'10*, pages 22–29. ELRA.

Moreno, A. and J. M. Goñi. 1995. GRAMPAL: A Morphological Processor for Spanish implemented in Prolog. In *GULP-PRODE'95*.

Orosz, G., A. Novák, and G. Prószéky. 2013. Hybrid Text Segmentation for Hungarian Clinical Records. In *MICAI (1)*, volume 8265 of *Lecture Notes in Computer Science*, pages 306–317. Springer-Verlag.

Padró, L. and E. Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *LREC'12*. ELRA.

Parra, C. 2012. Design and compilation of a specialized Spanish-German parallel corpus. In *LREC'12*, pages 2199–2206. ELRA.

Petrov, S., D. Das, and R. McDonald. 2012. A Universal Part-of-Speech Tagset. In *LREC'12*. ELRA.

Ratnaparkhi, A. 1999. Learning to Parse Natural Language with Maximum Entropy Models. *Machine Learning*, 34(1–3):151–175.

Schmid, H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, pages 44–49.

Sebastián, N., M. A. Martí, and M. F. Carreiras. 2000. *Léxico informatizado del español*. Edicions Universitat de Barcelona.

Taulé, M., M. A. Martí, and M. Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *LREC'08*. ELRA.

Procesamiento del Lenguaje Natural, Revista nº 54, marzo de 2015, pp 37-44 recibido 14-11-14 revisado 27-01-15 aceptado 10-02-15

Tratamiento de la Negación en el Análisis de Opiniones en Español*

Negation Scope Identification in Spanish Reviews

Salud M. Jiménez Zafra, Eugenio Martínez Cámara, M. Teresa Martín Valdivia, M. Dolores Molina González
Departamento de Informática, Escuela Politécnica Superior de Jaén
Universidad de Jaén, E-23071 - Jaén
{sjzafra, emcamara, maite, mdmolina}@ujaen.es

Resumen: El análisis de opiniones es una tarea a la que le quedan muchos frentes abiertos aún para que se pueda considerar resuelta. Entre ellos destaca el tratamiento de la negación, dado que una opinión negativa puede ser expresada con términos positivos negados. La negación es una característica particular de cada idioma, por lo que su tratamiento debe ajustarse a las singularidades del idioma en cuestión. En este artículo se presenta una aproximación lingüística para la identificación del ámbito de la negación en español, que se ha aplicado en un sistema de clasificación de la polaridad de opiniones sobre películas de cine.
Palabras clave: Análisis de la opinión, clasificación de la polaridad, identificación del ámbito de la negación

Abstract: Sentiment Analysis is a task that still has several opened challenges. One of those challenges is the treatment of the negation, because a negative opinion can be built using negated positive words. Negation is a particular feature of each language, thus it must be considered differently per each language. In this article is shown a linguistic approach for the negation scope identification with the aim of integrating it in a polarity classification system in the domain of movie reviews.
Keywords: Sentiment analysis, polarity classification, negation scope identification

1 Introducción

La Minería de Opiniones (MO) o Análisis de Sentimientos (AS) es una disciplina que combina técnicas de Procesamiento del Lenguaje Natural (PLN) y de Inteligencia Artificial y que está recibiendo bastante atención por parte de la comunidad científica seguramente por el amplio abanico de aplicaciones en las que se está empleando. La MO estudia el tratamiento de textos que incluyen información cargada de subjetividad. Muchos son los trabajos que tratan las distintas tareas y aplicaciones de la MO, pero a pesar de ello al AS todavía le queda un largo camino que recorrer para resolver muchos de los problemas que tiene aún abiertos (Liu, 2012). El presente trabajo se centra en uno de estos retos, la identificación del ámbito de la negación. La negación es un elemento clave en el análisis de opiniones, dado que son muy abundantes las opiniones negativas expresadas con términos positivos negados y viceversa. La oración *"No me gusta la carcasa del teléfono"* es un claro ejemplo de una opinión negativa con un término positivo (*gusta*) negado.

En el ámbito del AS, la mayor parte de la investigación se ha realizado para opiniones escritas en inglés. Sin embargo, la presencia cada vez mayor en Internet de otros idiomas pone de manifiesto la necesidad de desarrollar sistemas que traten lenguas diferentes al inglés. Así han aparecido algunos trabajos de investigación que tratan con textos escritos en chino (Zhang et al., 2009), en árabe (Rushdi-Saleh et al., 2011) o en español (Martín-Valdivia et al., 2013). En este artículo nos centraremos en un problema concreto del AS sobre textos en español. Nuestra aproximación consiste en la inclusión de un módulo de identificación del ámbito de la negación

* Este trabajo ha sido parcialmente financiado por el Fondo Europeo de Desarrollo Regional (FEDER), el proyecto ATTOS (TIN2012-38536-C03-0) del Gobierno de España, el proyecto AORESCU (TIC - 07684) del Gobierno regional de la Junta de Andalucía y el proyecto CEATIC-2013-01 de la Universidad de Jaén.

ISSN 1135-5948 © 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

en un sistema de clasificación de la polaridad en español, con el fin de mejorar la capacidad de predicción de la polaridad de textos de opinión.

Por otra parte, el tratamiento de la negación es un problema abierto dentro del PLN en general, y dentro de la MO en particular. Se trata de un fenómeno lingüístico que no ha sido suficientemente estudiado y que consideramos que requiere un análisis profundo. A pesar de las similitudes entre idiomas, la negación es un elemento lingüístico muy particular de cada lengua, por lo que para un tratamiento efectivo se debe realizar un estudio pormenorizado de los distintos elementos lingüísticos que intervienen en el proceso de negación de un enunciado. La investigación existente sobre la influencia de la negación es escasa, no estudia en profundidad estos aspectos lingüísticos y se centra casi en exclusividad en textos escritos en inglés. Este artículo tiene como una de sus principales aportaciones un estudio detallado del comportamiento de un conjunto partículas negativas, con el fin de determinar el conjunto de palabras que se ven afectadas para así decidir si se debe realizar una modificación de la orientación semántica de la palabra.

La segunda aportación de este trabajo es el uso de la identificación del ámbito de la negación en un sistema de clasificación de la polaridad. Para ello se ha seguido un enfoque no supervisado, en el que el clasificador de polaridad necesita de un recurso semántico que indique la orientación semántica de los términos que aparecen en los textos. Como se verá en las siguientes secciones se han evaluado tres recursos, SentiWordNet, iSOL y eSOL. Los resultados que se han obtenido confirman nuestra hipótesis de que la identificación del ámbito de la negación es un elemento de relevancia para un clasificador de la polaridad.

El resto del artículo se organiza como sigue: la siguiente sección presenta una sucinta exposición de la investigación relacionada con el AS y el tratamiento de la negación. La sección tres se centra en el estudio lingüístico llevado a cabo del ámbito de las diferentes partículas negativas existentes en español. Posteriormente, la sección cuatro detalla el marco experimental realizado para comprobar la efectividad de nuestra hipótesis, y por último, se exponen las conclusiones y la línea de trabajo a seguir.

2 Trabajos relacionados

Hasta ahora, la mayoría de las investigaciones relacionadas con el tratamiento de la negación se han realizado sobre opiniones escritas en inglés. Las primeras aproximaciones comenzaron en el año 2001 y sugieren métodos relativamente sencillos. Das y Chen (2001) proponen añadir "NOT" a las palabras de la oración que se encuentren cerca de términos negativos, como por ejemplo "no" o "don't". Pang, Lee, y Vaithyanathan (2002) utilizan la técnica de Das y Chen asumiendo que las palabras afectadas por una palabra negativa ("not", "isn't", "didn't", etc.) son todas aquellas que se encuentran entre dicha palabra y el primer signo de puntuación. Estos autores realizan experimentos empleando algoritmos de aprendizaje automático para la clasificación de las opiniones teniendo en cuenta el tratamiento de la negación (característica "NOT") y sin tenerlo en cuenta, llegando a la conclusión de que con este método se produce una mejora insignificante. Polanyi y Zaenen (2004) dan un paso más allá, considerando además de la negación, intensificadores y atenuantes, introduciendo de esta manera el concepto de modificadores de valencia contextuales. Además, se trata del primer modelo computacional que asigna puntuaciones a expresiones polares, invirtiendo la polaridad de las expresiones negadas. Desafortunadamente este modelo no llegó a implementarse, por lo que sólo se puede especular sobre su efectividad. Kennedy e Inkpen utilizan un modelo de negación muy similar al de Polanyi y Zaenen (2004), definiendo como ámbito de una palabra negativa/intensificador/atenuante aquella palabra inmediatamente posterior. En el caso de las palabras afectadas por la negación, el enfoque que siguen es el de invertir la polaridad y en el caso de las palabras que se encuentran en el ámbito de los intensificadores/atenuantes, lo que hacen es incrementar/disminuir el grado de positividad/negatividad según sea el caso. Para clasificar las opiniones utilizan dos métodos, en el primero de ellos clasifican un comentario en base al número de términos positivos y negativos que contiene y en el segundo emplean el algoritmo de aprendizaje automático SVM, llegando a la conclusión de que el modelado de la negación es un hecho importante. Por otro lado, Wilson, Wiebe, y Hoffmann (2005) proponen utilizar una ventana fija de tamaño 4 para determinar el

ámbito de la negación, es decir, una palabra se ve afectada por la negación si existe una expresión de negación entre las 4 palabras anteriores a ella. En este trabajo también emplean un método de aprendizaje no supervisado para clasificar las opiniones en el que utilizan características para modelar las palabras afectadas por la negación, por intensificadores/atenuantes y por otras expresiones polares.

Los enfoques presentados en el párrafo anterior son considerados los trabajos pioneros en el modelado de la negación en el AS en inglés. Sin embargo, estas soluciones no son suficientemente precisas. Por ello, se ha seguido trabajando en este tema para tratar de ofrecer mejores soluciones. Entre los últimos trabajos presentados cabe destacar los de Jia, Yu, y Meng (2009), Taboada et al. (2011) y Cruz Díaz (2014). Jia, Yu, y Meng (2009) proponen un sistema basado en reglas que utiliza información derivada de los árboles de dependencias de las oraciones de estudio, mejorando los enfoques existentes hasta el momento. En el completo y detallado trabajo de Taboada et al. (2011) se presenta una versión mejorada de su sistema SO-CAL (Taboada, Voll, y Brooke, 2008), para calcular la orientación semántica de una opinión, en el que marcan como negadas todas aquellas palabras que se encuentren después de una partícula negativa hasta llegar a un signo de puntuación, un conector o ciertas palabras pertenecientes a una determinada categoría gramatical (ej. "it" (pronombre)). Los autores introducen una nueva forma de tratar la negación que consiste en reducir el valor de polaridad de las palabras negadas en lugar de invertirlo. Cruz Díaz (2014) presenta un sistema para el tratamiento de la negación y de la especulación en textos médicos y opiniones, y el primer corpus de opiniones etiquetado a nivel de negación y especulación (Konstantinova et al., 2012). Por último, mencionar el trabajo de Wiegand et al. (2010) en el que se realiza una excelente revisión del estado del arte de la negación en inglés hasta ese momento.

Como se puede ver, la investigación sobre este tema en inglés es bastante amplia si la comparamos con la existente en español. Con respecto al tratamiento de la negación en documentos en español, el primer trabajo que conocemos es el de Brooke, Tofiloski, y Taboada (2009) en el que adaptan su primera versión del sistema SO-CAL para análisis de opiniones en inglés (Taboada, Voll, y Brooke, 2008), a un nuevo idioma, el español. Con respecto al tratamiento de la negación siguen el mismo enfoque que el empleado en su versión en inglés, teniendo en cuenta que en español los adjetivos pueden aparecer antes y después de los sustantivos. Finalmente, Vilares Calvo, Alonso Pardo, y Gómez Rodríguez (2013) tienen en cuenta la estructura sintáctica del texto para el tratamiento de la negación, de la intensificación y de las oraciones subordinadas. Los resultados de los autores muestran una mejora en el rendimiento con respecto a los sistemas puramente léxicos.

Nuestra propuesta se basa en la estructura sintáctica del texto, al igual que el último trabajo mencionado pero difiere del mismo en las reglas definidas y en el hecho de que Vilares Calvo, Alonso Pardo, y Gómez Rodríguez (2013) limitan el tratamiento de la negación a los términos "no", "nunca" y "sin", mientras que en nuestro trabajo se incluyen además nuevos términos, como son: "tampoco", "nadie", "jamás", "ninguno", "ni" y "nada".

3 Descripción de la arquitectura

En este trabajo se propone un sistema no supervisado para la clasificación de opiniones teniendo en cuenta la influencia de la negación. Así en primer lugar se describirá el enfoque utilizado para la identificación del ámbito de la negación, y posteriormente el sistema de clasificación de la polaridad.

3.1 Identificación del ámbito de la negación

La negación es un fenómeno lingüístico que debería ser tenido en cuenta en un amplio abanico de tareas dentro del PLN. El AS es una de esas tareas, ya que una correcta identificación del ámbito de la negación permitirá clasificar correctamente opiniones negativas construidas con términos positivos negados y viceversa. Para ello hay que comenzar con un estudio pormenorizado de las distintas partículas negativas que existen en español. En este trabajo se ha comenzado con las partículas negativas más importantes señaladas por la RAE (Española, 2009):"no", "tampoco", "nadie", "jamás", "ni", "sin", "nada", "nunca" y "ninguno". Cada una de las partículas indicadas puede tener un comportamiento diferente y afectar a distintos elementos dentro de una oración, por lo que la

manera más adecuada de identificar el ámbito es definir reglas para cada una de las partículas negativas.

Para la construcción de las reglas se ha utilizado el analizador de dependencias de Freeling[1], que permite generar el árbol de dependencias de una oración en base a su estructura sintáctica. Así, para cada partícula se han generado los árboles de dependencias correspondientes a distintas oraciones extraídas tras una revisión de distintos sitios web, en las que se hace uso de alguna de las partículas negativas abordadas, y se ha realizado un estudio de los mismos llegando a la conclusión de que es posible generalizar su tratamiento. En la Tabla 1 se muestran las reglas obtenidas tras realizar el estudio.

Partícula	**Regla**
no, tampoco, nadie, jamás, ninguno	Afecta al nodo padre y al árbol formado por el hermano de la derecha (incluido).
ni, sin	Afecta a todos los hijos y a todos los árboles formados por ellos hasta llegar a nodos hoja.
nada, nunca	Afecta al nodo padre.

Tabla 1: Reglas ámbito de la negación.

Con el fin de clarificar las reglas que se han definido, a continuación se muestran algunos ejemplos (Figuras 1, 2 y 3) utilizados para determinar el ámbito de cada partícula negativa. En cada figura se señala con una elipse la partícula analizada y con un recuadro su ámbito.

3.2 Sistema de clasificación de la polaridad

Una vez determinado el fragmento de la oración que se ve afectado por la negación, el siguiente paso es modificar su polaridad. Nuestra propuesta consiste en invertir la polaridad de las palabras pertenecientes a dicho fragmento que expresen opinión. Por ejemplo, en la oración *"La película no empieza mal con algunas escenas interesantes para introducirte en el drama vivido por el personaje principal."* (Figura 1), la partícula negativa "no" afecta a las palabras "empieza" y "mal" pero sólo "mal" expresa opinión (-1, negativa)

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

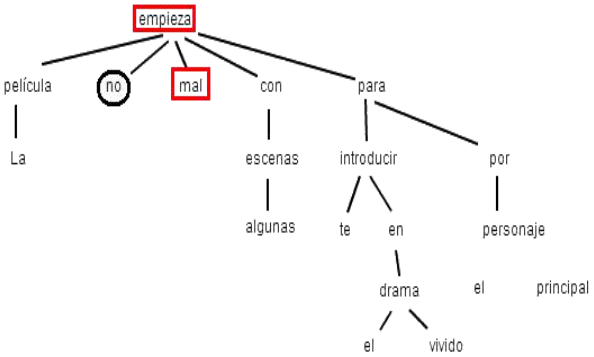


Figura 1: Árbol de dependencias en el que se analiza la partícula "no" en la oración: *La película no empieza mal con algunas escenas para introducirte en el drama vivido por el personaje principal.*

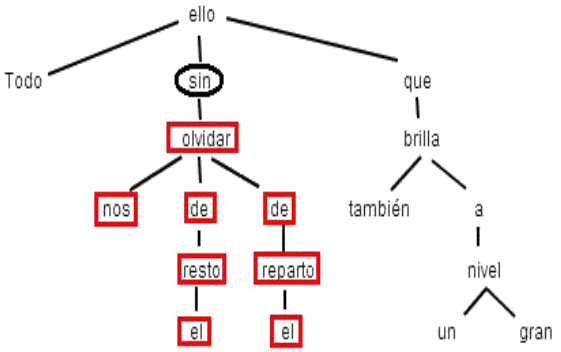


Figura 2: Árbol de dependencias en el que se analiza la partícula "sin" en la oración: *Todo ello sin olvidarnos del resto del reparto que brilla también a un gran nivel.*

por lo que su polaridad se verá invertida (1, positiva).

Para determinar la polaridad de una opinión se ha utilizado el enfoque basado en el cálculo de la orientación semántica de las palabras, que consiste en asignar a cada término que expresa opinión un valor que represente su polaridad, y en asignar a la opinión el valor correspondiente a la suma de estos valores. La Figura 4 muestra un esquema del sistema desarrollado.

El sistema propuesto permite identificar si una opinión es positiva o negativa sin llevar a cabo un entrenamiento previo, es decir, siguiendo un enfoque no supervisado. Para ello, cada opinión se divide en tokens y en oraciones para posteriormente realizar un análisis morfológico de cada una de las oraciones que forman la opinión. En este análisis se obtiene la categoría gramatical de cada una de las palabras de la oración. Tras realizar el análisis

Figura 3: Árbol de dependencias en el que se analiza la partícula "nunca" en la oración: *Nunca he visto una película peor que esta.*

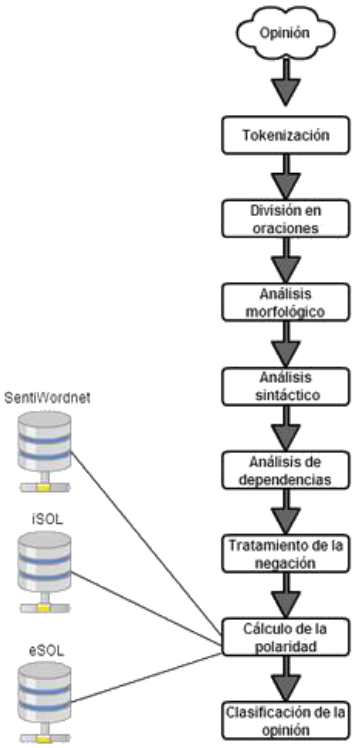


Figura 4: Arquitectura del sistema.

morfológico se procede a determinar las funciones de las palabras dentro de la oración mediante un analizador sintáctico. Una vez obtenida toda esta información sobre las palabras se realiza un análisis de dependencias que será de gran utilidad para determinar el ámbito de la negación.

El siguiente paso, fundamental en nuestro sistema, es el tratamiento de la negación. Aquí es donde se emplean las reglas obtenidas en el estudio realizado (Tabla 1). Cada palabra afectada por la negación se marca, de manera que a la hora de calcular la orientación semántica de la opinión se tenga en cuenta que su polaridad se tiene que invertir. Tras obtener el sentido de cada una de las palabras de la opinión y tras determinar cuáles forman parte del ámbito de la negación, el siguiente paso es calcular la polaridad de la opinión para poder clasificarla. Para ello se han utilizado tres recursos diferentes (SentiWordNet, iSOL, eSOL) que serán explicados posteriormente.

4 Marco experimental

A continuación se presentan los recursos que han hecho posible la consecución de este trabajo, así como los experimentos realizados y los resultados obtenidos.

Para la realización de las tareas de tokenización, análisis morfológico, análisis sintáctico, desambiguación y análisis de dependencias se ha utilizado Freeling (Padró, 2012). Se trata de una librería de código abierto que proporciona una amplia gama de herramientas para PLN en varios idiomas, entre ellos el español, razón por la cual hemos decidido emplearla en este trabajo.

Por otra parte, para probar la funcionalidad del sistema desarrollado se ha utilizado un corpus en español formado por 3878 críticas de cine recogidas de la web MuchoCine[2] (Cruz et al., 2008). Las críticas que componen el corpus están escritas por usuarios de la web. Esto aumenta la dificultad de la tarea, ya que los textos pueden contener errores gramaticales e incluso expresiones informales. Es preciso señalar que las críticas están puntuadas en un rango de 1 a 5, donde 1 indica que la película es muy mala y 5 que es muy buena. Para la realización de los experimentos se han considerado dos clases, "positiva" y "negativa". Las críticas con una valoración inferior a 3 se han clasificado como "negativas", mientras que las puntuadas con un valor superior a 3 se han etiquetado como "positivas". Las películas valoradas con un 3 (ni buenas ni malas) no se han tenido en cuenta en este estudio. Por tanto, la experimentación se ha llevado a cabo sobre un total de 2625 opiniones, de las cuales 1351 se corresponden con críticas "positivas" y 1274 con "negativas".

Debido al hecho de que el objetivo de nuestra propuesta es comprobar si la predicción del grado de subjetividad de un conjunto de opiniones mejora con la inclusión de un módulo para el tratamiento de la negación, resulta de gran utilidad analizar el número de opiniones de este corpus que utilizan alguna partícula de negación. De las 2625 opi-

[2] `http://www.lsi.us.es/~fermin/corpusCine.zip`

niones utilizadas en la experimentación, 2616 utilizan algún marcador negativo de los analizados (Tabla 2). En la Tabla 3 se puede ver la frecuencia de aparición de las diferentes partículas abordadas en este estudio reafirmando la importancia de su tratamiento.

Opiniones	Con partículas negativas	% opiniones con negación
Positivas	1345	99,56
Negativas	1271	99,76

Tabla 2: Opiniones del corpus MuchoCine con partículas negativas.

Partícula	Frecuencia
no	15517
sin	3816
ni	2107
nada	1583
nunca	671
tampoco	515
nadie	458
jamás	123
ninguno	97

Tabla 3: Frecuencia de aparición de las partículas negativas estudiadas en el corpus MuchoCine.

Por último, para determinar la polaridad de las opiniones del corpus se han utilizado diferentes recursos lingüísticos: SentiWordNet, iSOL y eSOL.

SentiWordNet es un recurso léxico que asigna a cada uno de los sentidos de las palabras de WordNet tres valores que reflejan su positivad, negatividad y objetividad. WordNet es una base de datos léxica en inglés que agrupa las palabras en base a su significado. Como es ampliamente conocido por la comunidad investigadora, SentiWordNet es un recurso para AS sobre textos en inglés. Para poder aplicarlo sobre textos en español se ha recurrido a una versión de WordNet en español con el fin de desambiguar cada término y así obtener su synset. Una vez que se tiene el synset del término, se puede acudir a SentiWordNet para obtener los valores de polaridad. La versión de WordNet en español empleada es la conocida como Multilingual Central Repository (MCR) (Gonzalez-Agirre, Laparra, y Rigau, 2012). La versión de WordNet en español de MCR contiene aproximadamente 38.000 synsets, que quedan todavía lejos de los 117.000 de WordNet.

Por su parte, iSOL es una lista de palabras indicadoras de opinión en español que es independiente del dominio (Molina-González et al., 2013). Está formada por 2509 términos positivos y 5626 negativos. Se trata de un lexicón generado a partir de la traducción automática de las palabras de la lista de Bing Liu, corregido manualmente y ampliado con nuevos términos. El último recurso empleado es eSOL. Se trata de una ampliación de la lista de palabras iSOL con términos adaptados al dominio de las críticas de cine (Molina-González et al., 2013).

4.1 Experimentos

Para evaluar el sistema desarrollado se han llevado a cabo 3 experimentos. En cada uno de los experimentos se realiza un proceso común. Se analizan las frases que componen cada opinión con el objetivo de obtener el significado de cada una de las palabras y de marcar aquellas que se vean afectadas por alguna de las partículas negativas abordadas en este estudio. La diferencia entre estos experimentos radica en el recurso utilizado para calcular la polaridad de las opiniones (Figura 4).

En el primer experimento se ha utilizado SentiWordNet, de manera que la polaridad de una opinión se ha calculado como la suma de la diferencia entre los valores de positividad y negatividad de cada palabra en base a su sentido más frecuente. En el segundo experimento se ha empleado la lista de palabras iSOL. En este caso, la polaridad de la opinión se ha calculado a partir de la suma de los términos de la opinión que pertenecen a la lista. Si la palabra pertenece a la lista de términos positivos se suma con valor 1 y si la palabra pertenece a la lista de términos negativos se suma con valor -1. En el último experimento se ha utilizado la lista de palabras eSOL y se ha procedido de la misma manera que se ha descrito en el segundo experimento.

Es preciso señalar que, en todos los experimentos, para las palabras que se han detectado como pertenecientes al ámbito de la negación se ha invertido su valor, es decir, se ha multiplicado por -1. Finalmente, la opinión se ha clasificado como "positiva" si el valor de su polaridad ha sido mayor que 0 y "negativa" en caso contrario.

Como medidas de evaluación del sistema se han tomado las más utilizadas en el ámbito del AS, es decir, Precisión (Prec.), Recall, F1

y Accuracy (Acc.).

4.2 Resultados

Para una mejor evaluación del impacto de la identificación del ámbito de la negación en el marco de un sistema de clasificación de la polaridad, se ha ejecutado el clasificador tanto sin el módulo de identificación de la negación como con él. En las Tablas 4 y 5 se muestran los resultados obtenidos. Además, en la Tabla 5 se ha incluido una columna que refleja la mejora con respecto al valor de Accuracy.

Recurso	Prec.	Recall	F1	Acc.
SWN	0,5896	0,5314	0,5590	0,5432
iSOL	0,6283	0,6246	0,6265	0,6270
eSOL	0,6365	0,6276	0,6320	0,6312

Tabla 4: Clasificación de la polaridad sin detección de la negación.

Recurso	Prec.	Recall	F1	Acc.	Mejora
SWN	0,6057	0,5489	0,5759	0,5596	3,02 %
iSOL	0,6416	0,6394	0,6405	0,6411	2,25 %
eSOL	0,6519	0,6430	0,6474	0,6465	2,42 %

Tabla 5: Clasificación de la polaridad con detección de la negación.

En la dos tablas se puede ver la diferencia entre los tres recursos lingüísticos empleados para la clasificación de la polaridad. Los bajos resultados que obtiene SentiWordNet consideramos que están justificados porque el recurso empleado para la identificación del synset (MCR) solo cubre aproximadamente 38000 synsets de los 117000 que contiene SentiWordNet. Evidentemente, muchos sentidos no pueden ser identificados por lo que la capacidad de identificación de la orientación semántica con SentiWordNet queda bastante mermada. La diferencia entre iSOL y eSOL también es explicable. Tal y como señalan Molina-González et al. (2013) en su trabajo, la lista eSOL contiene información del dominio del cine, por lo que como es de esperar, el resultado de la clasificación es mejor.

En cuanto al tratamiento de la negación, la Tabla 5 evidencia que se ha producido una pequeña mejora. Esto se ha debido a una reducción del número de Falsos Positivos y Falsos Negativos, ya que se han podido tratar correctamente los términos positivos o negativos que se encuentran negados. Estos resultados nos animan a seguir investigando en el tratamiento de la negación, y en consecuencia, en el estudio del resto de partículas que todavía quedan por incluir en nuestro sistema.

5 Conclusiones y trabajo futuro

En este artículo se ha presentado una primera aproximación a la identificación del ámbito de la negación que se ha aplicado sobre un sistema de clasificación de opiniones en español. La identificación de la negación se ha realizado siguiendo un enfoque basado en reglas lingüísticas, que se han definido después de realizar un estudio de cómo afectan algunas partículas negativas a los elementos que las circundan en una oración. Posteriormente, el módulo generado se ha incluido en un sistema de clasificación de la polaridad basado en el uso de recursos lingüísticos de opinión. Los resultados que se han obtenido, aunque las mejoras han sido pequeñas, certifican nuestra hipótesis inicial de que la toma en consideración del ámbito de la negación es de relevancia para la clasificación de la polaridad, lo que nos anima a seguir investigando.

Nuestro trabajo futuro se va a centrar en llevar a cabo un análisis de la significancia estadística y en el estudio del resto de partículas de negación que no se han incluido en este primer trabajo. Además, se va a realizar un análisis exhaustivo del corpus para determinar las oraciones en las que las partículas negativas afectan a expresiones polares, para así poder comprobar cómo han funcionado las reglas definidas, con objeto de poder refinarlas. También, se va a iniciar un estudio sobre la importancia de las partículas de intensificación a la hora de determinar la polaridad de una oración.

Bibliografía

Brooke, Julian, Milan Tofiloski, y Maite Taboada. 2009. Cross-linguistic sentiment analysis: From english to spanish. En *International Conference RANLP, 50-54*.

Cruz, Fermín L, Jose A Troyano, Fernando Enriquez, y Javier Ortega. 2008. Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. *Procesamiento de Lenguaje Natural*, 41.

Cruz Díaz, Noa Patricia. 2014. *Detección de la Negación y la Especulación en Textos*

Médicos y de Opinión. Ph.D. tesis, Universidad de Huelva.

Das, Sanjiv y Mike Chen. 2001. Yahoo! for amazon: Extracting market sentiment from stock message boards. En *Proc. of the Asia Pacific finance association annual conference (APFA)*, volumen 35, página 43. Bangkok, Thailand.

Española, Real Academia. 2009. *Nueva gramática de la lengua española*, volumen 1. Espasa Calpe Madrid, Spain.

Gonzalez-Agirre, Aitor, Egoitz Laparra, y German Rigau. 2012. Multilingual central repository version 3.0. En *LREC, 2525-2529*.

Jia, Lifeng, Clement Yu, y Weiyi Meng. 2009. The effect of negation on sentiment analysis and retrieval effectiveness. En *Proc. of the 18th ACM conference on Information and knowledge management, 1827-1830*.

Konstantinova, Natalia, Sheila CM de Sousa, Noa P Cruz Díaz, Manuel J Maña López, Maite Taboada, y Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. En *LREC, 3190-3195*.

Liu, Bing. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Martín-Valdivia, María-Teresa, Eugenio Martínez-Cámara, Jose-M Perea-Ortega, y L Alfonso Ureña-López. 2013. Sentiment polarity detection in spanish reviews combining supervised and unsupervised approaches. *Expert Systems with Applications*, 40(10):3934–3942.

Molina-González, M Dolores, Eugenio Martínez-Cámara, María-Teresa Martín-Valdivia, y José M Perea-Ortega. 2013. Semantic orientation for polarity classification in spanish reviews. *Expert Systems with Applications*, 40(18):7250–7257.

Padró, Lluís. 2012. Analizadores multilingües en freeling. *Linguamática*, 3(2):13–20.

Pang, Bo, Lillian Lee, y Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. En *Proc. of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, páginas 79–86. ACL.

Polanyi, L y A Zaenen. 2004. Context valence shifters. En *Proc. of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*.

Rushdi-Saleh, Mohammed, M Teresa Martín-Valdivia, L Alfonso Ureña-López, y José M Perea-Ortega. 2011. Oca: Opinion corpus for arabic. *Journal of the American Society for Information Science and Technology*, 62(10):2045–2054.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, y Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Taboada, Maite, Kimberly Voll, y Julian Brooke. 2008. Extracting sentiment as a function of discourse structure and topicality. *Simon Fraser Univeristy School of Computing Science Technical Report*.

Vilares Calvo, David, Miguel Ángel Alonso Pardo, y Carlos Gómez Rodríguez. 2013. Clasificación de polaridad en textos con opiniones en español mediante análisis sintáctico de dependencias. *Procesamiento de Lenguaje Natural*, 50:13–20.

Wiegand, Michael, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, y Andrés Montoyo. 2010. A survey on the role of negation in sentiment analysis. En *Proc. of the workshop on negation and speculation in natural language processing*, páginas 60–68. ACL.

Wilson, Theresa, Janyce Wiebe, y Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. En *Proc. of the conference on human language technology and empirical methods in natural language processing*, páginas 347–354. ACL.

Zhang, Changli, Daniel Zeng, Jiexun Li, Fei-Yue Wang, y Wanli Zuo. 2009. Sentiment analysis of chinese documents: From sentence to document level. *Journal of the American Society for Information Science and Technology*, 60(12):2474–2487.

Procesamiento del Lenguaje Natural, Revista nº 54, marzo de 2015, pp 45-52 recibido 23-11-14 revisado 27-01-15 aceptado 16-02-15

Esquema de anotación para categorización de citas en bibliografía científica

Annotation scheme for citation classification in scientific literature

Myriam Hernández Álvarez
Escuela Politécnica Nacional
Facultad de Ingeniería de Sistemas
Quito, Ecuador
myriam.hernandez@epn.edu.ec

José Gómez Soriano
Universidad de Alicante
Dpto. de Lenguajes y Sistemas Informáticos
Alicante, España
jmgomez@ua.es

Resumen: El análisis de citas bibliográficas que usa variaciones de métodos de conteo provoca deformaciones en la evaluación del impacto. Para enriquecer el cálculo de los factores de impacto se necesita entender el tipo de influencia de los aportes de un investigador sobre el autor que los menciona. Para ello, se requiere realizar análisis de contenido del contexto de las citas que permita obtener su función, polaridad e influencia. El presente artículo trata sobre la definición de un esquema de anotación tendiente a la creación de un corpus de acceso público que sea la base de trabajo colaborativo en este campo, con miras al desarrollo de sistemas que permitan llevar adelante tareas de análisis de contenido con el objetivo planteado.
Palabras clave: Análisis de citas bibliográficas, análisis de contenido, esquemas de clasificación de citas, anotación de corpus, función, polaridad, influencia.

Abstract: Citation analysis that uses counting methods causes deformations in impact factor assessment. To enrich impact factor calculation is necessary to understand the kind of influence that the contributions of an author have over another´s work. For this purpose, it is required to perform citation content analysis to obtain its function, polarity and influence in a context within an article that mentioned it. In this paper, we focus in the definition of an annotation scheme aimed at creating a public access corpus that be the basis of collaborative work in this field, in order to develop citation content analysis to obtain criteria for impact evaluation.
Keywords: Citation analysis, content analysis, citation classification schemes, corpus annotation, function, polarity, influence.

1 Introducción

El análisis de citas bibliográficas en artículos científicos sirve para evaluar el impacto de un autor, de su obra o, incluso, de la revista en la que publica. El método actual de medición utiliza, básicamente, métodos cuantitativos relacionados con el conteo de las citas (Garfield, 1972), aunque también se utilizan ciertas variaciones como el PageRank, que es un algoritmo que tiene en cuenta la propia relevancia del que cita, (Page, et al., 1999) o la cocitación (Small, 1973), que añade como medida de similitud entre dos trabajos el número de documentos comunes que los citan.

Está bien documentado el hecho de que los métodos que usan criterios puramente cuantitativos provocan deformaciones en la percepción del impacto y la relevancia de los artículos citados. Un artículo con un alto número de citas no necesariamente es un artículo correcto y sus resultados tampoco tienen por qué estar confirmados por los investigadores que lo citan, puesto que muchas de las citas pueden ser críticas o manifestar algún tipo de rechazo. Radicchi (2012) mostró

ISSN 1135-5948 © 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

que artículos incompletos, erróneos o controvertidos tienden a tener un número de citas mayor. Los investigadores pueden estar tentados a publicar este tipo de obras para recibir un mayor número de referencias (Marder, Kettenmann y Grillner, 2010) o utilizar prácticas poco éticas para adquirir relevancia. Por ejemplo, Van Noorden (2013) explicó el caso de cinco revistas brasileñas que usaron auto citas y citas cruzadas para sesgar el índice del Journal Citation Reports. El premio Nobel Randy Schekman, en una publicación en The Guardian[1] (Sample, 2013), denunció estas prácticas por parte de prestigiosas revistas científicas que prefieren novedad y polémica antes que trabajo científico serio. En la misma publicación[1], el editor en jefe de Nature declaró que muchas veces a lo largo del tiempo ha manifestado su preocupación respecto a los peligros que conlleva un exceso de confianza en los factores de impacto basados en conteos de citas.

Estos análisis basados en contar el número de citas o variaciones de esta técnica, no toman en cuenta el tipo de influencia de un autor sobre otro (Zhang, Ding, y Milojević, 2013). Para entender esta influencia se requiere conocer la disposición del autor hacia el artículo citado y la función de la cita en el artículo que la menciona. No todas las citas tienen el mismo efecto en el artículo que las cita. El impacto de un artículo citado puede variar considerablemente si se toma en cuenta que la referencia contiene una crítica, es el punto de partida de un trabajo o si, simplemente, reconoce el trabajo de otros autores. Por esta razón, se vuelve importante identificar métricas más completas que tomen en cuenta el contenido de lo que se dice sobre la obra citada para evaluar su impacto y relevancia. Para ello, se hace necesario realizar un análisis de contenido del texto que contiene las citas para obtener ciertas características que puedan ser aplicadas a las actuales métricas para mejorar el cálculo bibliométrico de los investigadores y revistas. Se requiere la construcción de un índice más complejo que tenga en cuenta la intención del autor y su disposición hacia el trabajo que cita para determinar el impacto de éste de forma más precisa y analizando más factores que, únicamente, el número de citas recibidas. Intención y disposición son criterios que se relacionan con la función y la polaridad de la cita que forman parte del análisis de contenido.

Athar (2014) demostró que, para determinar tanto la función como la polaridad de una cita bibliográfica, se tienen mejores resultados cuando se analiza no solo la oración que contiene la cita sino también oraciones anteriores y/o posteriores que forman parte de un contexto. Este contexto debe ser definido de forma dinámica detectando las oraciones adyacentes a la cita que tienen algún argumento sobre ella. Sin embargo, el reconocimiento automático de argumentos en textos es todavía una tarea que presenta grandes retos lo que obstaculiza la detección automática del contexto de una cita. Por otra parte, la estructura del artículo y el sitio en el que se encuentra la cita también pueden servir para definir su función. Un artículo referenciado en la introducción probablemente sea una cita superficial, mientras que un artículo nombrado en la sección en la que se describe la metodología o los resultados, tiene una mayor probabilidad de cumplir con una función clara dentro del artículo que cita (White, 2004).

Dentro de este marco se ha definido que, para poder desarrollar la investigación en este campo, se requiere la generación de un corpus que clasifique las citas en forma estándar, que esté públicamente disponible y que permita el trabajo colaborativo en el área de Análisis de Contenido de referencias bibliográficas (Hernández y Gómez, 2014). Esta información permitirá obtener nuevos factores que podrán ser incorporados en el cálculo de un índice de impacto más completo, preciso y justo que permita evaluar de mejor forma la influencia de los artículos citados y que evite incentivos perniciosos. Para llevar a cabo esta ambiciosa tarea, es necesario que existan ciertos recursos para que los investigadores puedan avanzar y

[1]http://www.theguardian.com/science/2013/dec/09/nobel-winner-boycott-science-journals

poner a prueba sus sistemas. Uno de los más cruciales es la construcción de un corpus etiquetado que sirva de gold standard y que sea lo suficientemente grande para que las evaluaciones puedan dar resultados estadísticamente significativos.

Con este objetivo en mente, se ha definido un esquema de anotación para este corpus con las siguientes consideraciones: que contemple factores que se requiere incluir en un sistema de evaluación de citas bibliográficas: la clasificación de la influencia y la función de la cita y el análisis de la polaridad; que sirva como estándar de anotación en esta área de investigación; que pueda ser etiquetado por personas que no sean especialistas en el tema del artículo que se anota ni relacionadas con las Tecnologías del Lenguaje Humano; que clasifique de la manera más clara posible todas las citas del artículo, tomando en cuenta su función, influencia y polaridad; que elimine la máxima ambigüedad posible evitando el solapamiento de las categorías; que permita una construcción ágil de un corpus de citas bibliográficas anotadas que facilite una población lo suficientemente grande para que sea representativo y que sirva como gold standard; que se encuentre en el punto óptimo de granularidad y utilidad; que permita un etiquetado semi- supervisado y que realice un marcado lingüístico del texto para obtener reglas que faciliten trabajos posteriores.

2 Esquema de anotación

Para delimitar un esquema de anotación se hace necesario considerar que hay, al menos, dos enfoques a la categorización de funciones de las citas. El primero define que un esquema es útil en la medida en que es exhaustivo en su profundidad y granularidad. Un ejemplo de este enfoque es el esquema de 35 categorías de Garzone (1996). El segundo enfoque es el tomado por Teufel y Moens (1999) que cuestiona esta categorización de grano fino pues asevera que la mayoría de instancias son difíciles de detectar porque no se encuentran pistas lingüísticas evidentes que puedan servir para clasificarlas y que aún si están presentes claves explícitas, el problema de detectarlas es formidable. Teufel, y Moens (1999) también expresan que juzgar la naturaleza de la cita conlleva un alto nivel de subjetividad y que hay una ausencia de medios para mapear esa naturaleza a los propósitos de la cita.

En el punto medio de granularidad, nuestro esquema está diseñado para especificar características que pueden definir impacto de una cita tomando en cuenta su función, polaridad e influencia dentro del texto que la referencia. Con los resultados preliminares de la anotación veremos la relación entre la clasificación de acuerdo a los tres criterios. Tratamos que cada categoría sea fácilmente diferenciable de las demás para que los anotadores humanos puedan distinguirlas y para facilitar la generación de un corpus anotado que permita que se lo siga aumentando en forma manual o automática.

Para aplicar este diseño es necesario definir el contexto de la cita que debe incluir lo más posible lo que se dice sobre la cita. Debido a la complejidad que presenta la búsqueda de argumentos para definir un contexto, se resolvió establecer el contexto con la longitud de un párrafo. Para tomar esta decisión partimos del hecho de que, por definición, un párrafo es el conjunto de oraciones que expresan una idea o argumento completo, por lo tanto, un párrafo tiene una buena posibilidad de incluir la argumentación relevante sobre una cita. Esto funciona bien, y mantiene el criterio de que el contexto varíe de acuerdo a la presencia de argumentos en torno a la referencia.

Para construir un corpus gold-standard con información extra que facilite la obtención de reglas que sirvan de guía en un proceso de anotación manual o automático, proponemos una nueva metodología de anotación. El codificador detecta estructuras sintácticas fijas (palabras clave) y variables (patrones en forma de etiquetas XML) en las oraciones y forma patrones que le ayudan a reconocer las categorías del esquema. Se ha comprobado experimentalmente que este tipo de anotación aclara dudas en los codificadores y permite una mayor coherencia en las anotaciones y un porcentaje de acuerdo más alto. Los datos de patrones y palabras clave (skip-grams) se guardan y ofrecen información adicional para anotaciones sucesivas. Esta información

facilita la labor de los codificadores, puesto que les sirve como ejemplos adicionales que aclaran los casos que se van presentando y permiten una definición objetiva de las diferencias entre funciones. Esta información de patrones y palabras clave también será la base para un sistema de aprendizaje que automáticamente genere un corpus más extenso, a través de su uso como reglas en un modelo de aprendizaje semi-supervisado.

Con el objetivo mencionado, se delinean dos pasos para la anotación. En el primero se pide al anotador que lea el párrafo y establezca las partes variables y fijas que detecta; marcando solamente lo que le parezca indispensable para definir la estructura de la oración y lo que se dice sobre la cita. Estos patrones deben ser lo más simple posible, por lo que se le pide al anotador que solamente marque las partes básicas. Cuando se han establecido estos patrones, se le sugiere al anotador que revise la información disponible respecto a palabras clave y partes variables. Con estos datos, resulta mucho más fácil, tomar la decisión respecto a la clase a la que pertenece la cita.

El esquema de clasificación usado, se presenta en las Tablas 1, 2, 3 y en la Figura 1.

Función de la cita	Descripción
Based on, supply	El artículo que cita se construye sobre material del artículo citado que puede ser un concepto o herramienta. El artículo citado es usado en el artículo que cita.
Useful, Standard	El material del artículo citado (concepto o herramienta) se aplica en algún otro trabajo, no en el propio. El trabajo citado se relaciona con una idea usada como medida, norma o modelo.
Acknowledge, Corroboration, Positive contrast	El artículo citado se menciona para reconocer algún trabajo previo. El artículo que cita confirma o soporta algún aspecto de la cita. Se hace contraste positivo.
Weakness, correct, negative contrast	Se nota o corrige un error o debilidad del trabajo citado. Se establece un contraste negativo.
Weakness, Hedges	Uso de lenguaje cuidadoso para ocultar una disposición negativa hacia el trabajo citado.

Tabla 1: Esquema de anotación para funciones

Influencia	Descripción
Significant	La cita es importante para el trabajo que la referencia. El trabajo citado, por alguna razón merece atención.
Perfunctory	La cita no cumple un rol importante en el artículo que la menciona. La cita se hace para reconocer trabajo previo.

Tabla 2: Esquema de anotación para influencia

Polaridad	Descripción
Positive	El autor tiene una disposición favorable hacia el trabajo citado.
Negative	El autor tiene una disposición no favorable hacia el trabajo citado.

Tabla 3: Esquema de anotación para polaridad

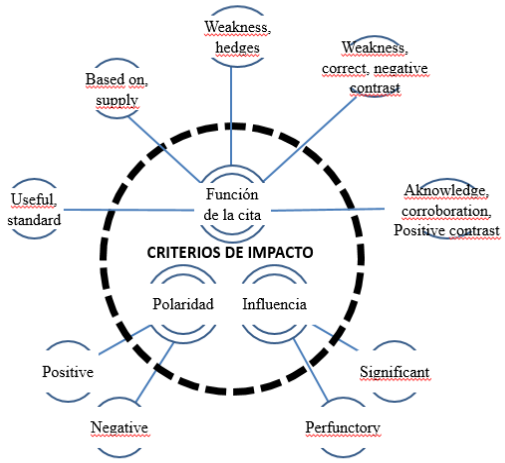


Figura 1: Criterios para evaluar el impacto de una cita de acuerdo a su función, polaridad e influencia

La lógica de la clasificación tiene que ver con funciones con definición de clases

claramente separadas que cubren las posibilidades existentes y que además se van a relacionar con medidas de impacto. En orden de evaluación de impacto, considerando la profundidad de la relación entre el trabajo que cita y el trabajo citado, tendremos los grupos de funciones: "Based on, supply", "Useful, Standard", "Acknowledge, Corroboration, Positive Contrast"; valores negativos de impacto podrían tener las funciones "Weakness" y "Weakness, Hedges". Juntamos categorías que reflejan similar importancia de la cita para el artículo que la referencia; así disminuimos la complejidad del esquema y logramos que continúe sirviendo a nuestro objetivo que es el análisis de funciones en relación con la evaluación del impacto.

"Weakness, Hedges" es una categoría interesante. Esta clase es un caso especial de "Weakness" en el que se presenta una forma de lenguaje cauteloso para ocultar una disposición negativa respecto al artículo citado, con el fin de evitar reacciones no deseadas por parte de los afectados. Detectar la presencia de "hedges" permitirá descubrir citas que implican posiciones y polaridades negativas encubiertas de diversas maneras. Para ello nos basamos en un análisis realizado por Hyland (1996).

2.1 Ejemplos

Usando los patrones para clasificación de citas, se pueden obtener los ejemplos para cada función. Al momento hemos considerado solamente referencias bibliográficas en idioma inglés, puesto que es el lenguaje común de la ciencia; sin embargo, opinamos que los patrones, las etiquetas y la clasificación, podrían ser aplicados también en otros idiomas; por supuesto sería necesario tomar en cuenta características especiales de los distintos idiomas y multilingüismo. Veamos un ejemplo de patrones en inglés:

Texto: "Argumentative Zoning (Teufel, 1999; Teufel and Moens, 2002) attempts to solve this problem by representing the structure of a text using a rhetorically-based schema. We used another technique". Patrón: METHOD (CITE) attempts to solve * METHOD. AUTHOR used another technique. Clasificación de la cita, función: Weakness, Hedges; polaridad: Negativa. En este último caso, al usar la secuencia de palabras "attempts to solve" se podría indicar que intentaron resolver un problema, pero no lo consiguieron y por ello se prefirió usar otro método. Muestra velada negatividad hacia la cita.

2.2 Validez del esquema

Un requerimiento indispensable para respaldar la calidad de un modelo como el discutido, requiere demostrar la confiabilidad de los datos obtenidos. La confiabilidad de los datos se relaciona con la confiabilidad del proceso de anotación. Para evaluar este parámetro con varios codificadores, es necesario medir el acuerdo obtenido en una pequeña muestra del corpus que ha sido revisado por los mismos anotadores. Para poder generalizar los resultados obtenidos en esta muestra a todo el proceso en el que probablemente van a intervenir nuevos anotadores y no solo los que codificaron la muestra, se necesita que el proceso sea confiable (Artstein y Poesio, 2008).

Según Krippendorff (2004) la confiabilidad o reproducibilidad de la anotación se asegura cumpliendo tres requerimientos: un esquema claro con instrucciones detalladas y criterios específicos para escoger codificadores. Para medir reproducibilidad se deben tener al menos tres anotadores que deben trabajar independientemente entre sí.

En nuestro experimento se cumplieron estos tres requerimientos. Se propuso un esquema claro, detallado y con suficientes ejemplos de aplicación; los anotadores son personas que lo han revisado cuidadosamente y tienen conocimientos del área de lingüística computacional; y, por último, para anotar la muestra se tuvieron tres codificadores que trabajaron en forma separada. Se pidió a los anotadores que sigan en forma consistente un procedimiento claramente establecido, en el cual se realiza primeramente una pre anotación con los patrones y palabras claves, de la forma como se ha explicado.

3 Organización de los experimentos y datos

Se usaron artículos del archivo de la Asociación de Lingüística Computacional (ACL por sus siglas en inglés), escogidos de forma aleatoria. Los textos se pre procesaron para marcar párrafos para detectar contexto. Para validar el modelo y calcular el acuerdo entre anotadores se utilizaron citas en artículos que fueron etiquetados de forma independiente por tres personas con conocimientos en el campo de lingüística computacional. Los datos usados para el cálculo del acuerdo entre anotadores contienen 101 citas, una variable para función, una variable para polaridad y una variable para influencia con 303 decisiones cada una.

El proceso de anotación se realizó en varias etapas. La primera consistió en un proceso de pre anotación para reconocer y numerar las citas en el texto. Para ello, se utilizó un programa desarrollado por el grupo de investigación de PNL de la Facultad de Ingeniería de Sistemas de la Escuela Politécnica Nacional de Quito. Este programa reconoce expresiones regulares asociadas a las referencias bibliográficas en el formato oficial de la Antología de la ACL[2]. La segunda etapa consistió en reconocer patrones en el texto. La información de las etiquetas de los patrones en el texto, guían al anotador en la clasificación de la función y la polaridad. Por último, se realizó un procesamiento de cada uno de los artículos, para definir el número de veces que la referencia fue nombrada y el sitio en la cual se la mencionó: introducción, método, resultados, discusión o sus equivalentes. Estos datos se usan para definir el tipo de influencia de la cita. Para estos últimos pasos se usó el editor de NetBeans IDE 8.02 para formato XML. Las anotaciones de cada codificador se guardaron en archivos de texto separados, cuya primera línea correspondía a los nombres de las variables, es decir la función, la polaridad y la influencia, separadas por una tabulación; y, las siguientes líneas fueron los resultados de las anotaciones para cada cita. Estos archivos fueron cargados en el programa desarrollado por Geertzen, J., 2012, para obtener el cálculo del nivel de acuerdo entre anotadores.

4 Resultados y discusión

Artstein y Poesio, 2008 dicen que los datos son confiables si se muestra que los anotadores entendieron las categorías asignadas y por tanto producen en forma consistente resultados similares. De este modo, la confiabilidad es un requisito para demostrar la validez de un esquema. Si los codificadores no muestran acuerdo entre sí, puede deberse a que algunos de ellos están equivocados o a que el esquema de anotación no es apropiado para los datos. Adicionalmente, la confiabilidad implica la habilidad de distinguir entre las clases, la que se posibilita si el esquema es claro. El acuerdo entre anotadores puede evaluarse utilizando coeficientes que toman en cuenta la corrección por la probabilidad de que los codificadores estén de acuerdo en un ítem simplemente por azar. Artstein y Poesio (2008) sugieren que los coeficientes se escojan de acuerdo a la tarea. En el caso de que la anotación tenga un sesgo hacia algunas(s) categorías, de acuerdo a la recomendación de los autores mencionados, para la evaluación de confiabilidad se debe usar el coeficiente de Krippendorff. Obtenemos valores para varios coeficientes incluido el mencionado. Las Tablas 4, 5 y 6 se muestran los resultados.

Fleiss	Krippendorff	Pairwise avg.
A_obs=0.845 A_esp=0.365 Kappa=0.756	D_obs = 0.155 D_esp = 0.637 Alpha = 0.756	% agr = 84.5 Kappa=0.756

Tabla 4: Acuerdo entre anotadores con pre anotación correspondiente a la Función

Fleiss	Krippendorff	Pairwise avg.
A_obs = 0.96 A_exp = 0.72 Kappa=0.859	D_obs = 0.04 D_exp = 0.281 Alpha = 0.859	% agr = 96 Kappa=0.86

Tabla 5: Acuerdo entre anotadores con pre anotación correspondiente a la Influencia

[2] http://www.aclweb.org/anthology/

Fleiss	Krippendorff	Pairwise avg.
A_obs = 0.98 A_exp=0.776 Kappa=0.912	D_obs = 0.02 D_exp = 0.225 Alpha = 0.912	% agr = 98 Kappa=0.913

Tabla 6: Acuerdo entre anotadores con pre anotación correspondiente a la Polaridad

Se obtienen los coeficientes: Kappa de Fleiss (Fleiss, 1971), Alpha de Krippendorf (Krippendorff, 2004) y Kappa para el promedio por pares. Se utilizó el software de Geertzen, J. (2012) para el cálculo de los coeficientes. A_obs es el acuerdo observado, A_exp es el acuerdo esperado, D_obs es el desacuerdo observado y D_exp es el desacuerdo esperado. Los valores bajos en los coeficientes nos indicarían que los anotadores tuvieron problemas para distinguir entre categorías y lo contrario validaría la claridad del esquema, como sucede en nuestro caso.

De acuerdo a los resultados, los valores para la clasificación de la polaridad pueden ser mapeados directamente de las funciones. Son funciones positivas: "Based on, Supply", "Useful, Standard", "Acknowledge, Corroboration, Positive Contrast". Son negativas: "Weakness" y "Weakness Hedges".

Los resultados demostraron que la influencia de la cita tiene que ver con el número de veces que se la menciona en una sección del artículo diferente a la Introducción. Los artículos que se califican como superficiales aparecen principalmente en la Introducción, son fundamentales las citas que tienen una función "Based on, Supply" o se citan más de dos veces. Con estos criterios, una buena aproximación para la clasificación de la influencia, podría realizarse en forma automática con los datos de las funciones y de la ubicación de la cita y esta clasificación no requeriría ser anotada manualmente.

5 Conclusiones y trabajo futuro

Un requerimiento indispensable para respaldar y probar un modelo como el discutido, es la demostración de confiabilidad en los resultados obtenidos. La confiabilidad de los datos tiene que ver con la confiabilidad del proceso de anotación. Para medir la confiabilidad en la anotación del esquema con varios codificadores, es necesario medir el acuerdo obtenido en una pequeña muestra del corpus que ha sido revisado por las mismas personas. De esta manera, el modelo del esquema se valida a través del resultado que lo califica como reproducible.

En la clasificación de funciones de las citas, el porcentaje de acuerdo, sin tomar en cuenta una corrección por acuerdos al azar, es de 84.5%. Con la respectiva corrección incluida en el cálculo de Kappa, se tiene un K = 0,756. De acuerdo a Landis y Koch (1977), se puede concluir que esta magnitud de Kappa corresponde a un substancial acuerdo entre anotadores, nivel que se considera para Kappa entre 0,6 y 0,8. El acuerdo entre anotadores es incluso mayor para la clasificación de polaridad e influencia. La polaridad tiene un K = 0,912 y la influencia un K = 0,859. Los mismos autores (Landis y Koch, 1977) califican a los resultados para Kappa, que van entre 0,8 y 1,0, como perfectos. Se esperaban valores mayores de Kappa para clasificaciones binarias, como las realizadas para Influencia y Polaridad, porque puede relacionarse directamente la precisión de los anotadores con el número de clases entre las que tienen que decidir. A partir de estos resultados, el esquema de clasificación propuesto y la metodología de anotación fueron validados a satisfacción.

Una de las contribuciones de este trabajo, es la de presentar un esquema y metodología de anotación que permitirán el desarrollo de un corpus base, que podrá ser extendido a través de aprendizaje automático o incluso de métodos manuales.

Como trabajo futuro nos planteamos usar este esquema de clasificación para construir un corpus anotado de acceso público, que pueda servir a la comunidad científica para el desarrollo de sistemas enfocados a evaluar el factor de impacto en bibliografía científica utilizando criterios adicionales obtenidos a partir del análisis de contenido del contexto de las citas.

La experimentación que confirma que el

esquema planteado es relevante para tareas de aprendizaje automático está en etapa de desarrollo y los resultados serán presentados en un artículo posterior. Se están usando dos técnicas: aprendizaje supervisado con reglas desarrolladas como expresiones regulares formadas por los patrones en las anotaciones; y, SVM utilizando como características las etiquetas de los patrones.

Bibliografía

Artstein, R., y Poesio, M. 2008. Inter-coder agreement for computational linguistics. Computational Linguistics, 34(4), 555-596.

Athar, A. 2014. Sentiment analysis of scientific citations. Technical Report, University of Cambridge.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5):378–382.

Garfield, E. 1972. Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. Science 178: 471–9.

Garzone, M. A. 1996. Automated classification of citations using linguistic semantic grammars. Master's thesis, The University of Western Ontario.

Geertzen, J. 2012. Inter-Rater Agreement with multiple raters and variables. Retrieved October 8, 2014, from https://mlnl.net/jg/software/ira/.

Hernández, M., y Gómez, J. M. 2014. Survey in sentiment, polarity and function analysis of citation. In Proceedings of the First Workshop on Argumentation Mining ACL 2014, Baltimore, MD, pp. 102–3.

Hyland, K. 1996. Writing without conviction? Hedging in science research articles. Applied linguistics, 17(4), 433-454.

Krippendorff, Klaus. 2004. Reliability in content analysis: Some common misconceptions and recommendations. Human Communication Research, 30(3):411–433.

Landis, J. R., y Koch, G. G. 1977. The measurement of observer agreement for categorical data. Biometrics, 159-174.

Marder, E., Kettenmann, H., y Grillner, S. 2010. Impacting our young. Proceedings of the National Academy of Sciences of the United States of America 107: 21233.

Page, L., Brin, S., Motwani, R., y Winograd, T. 1999. The PageRank citation ranking: Bringing order to the web. Technical Report, Stanford InfoLab.

Radicchi, F. 2012. In science "there is no bad publicity": Papers criticized in comments have high scientific impact. Scientific Reports 2: 815.

Sample, I. 2013. Nobel winner declares boycott of top science journals. The Guardian. Available at http://www.theguardian.com/science/2013/dec/09/nobel-winner-boycott-science-journals

Small, H. 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science 24: 265–9.

Teufel, S., y Moens, M. 1999. Discourse-level argumentation in scientific articles: Human and automatic annotation. In M. Walker (ed.), Towards Standards and Tools for Discourse Tagging: Proceedings of the Workshop, pp. 84–93. Somerset, NJ: Association for Computational Linguistics.

Van Noorden, R. 2013. Brazilian citation scheme outed. Nature, 500(7464), 510–1.

White, H. D. 2004. Citation analysis and discourse analysis revisited. Applied linguistics, 25(1), 89-116.

Zhang, G., Ding, Y., y Milojević, S. 2013. Citation content analysis (cca): A framework for syntactic and semantic analysis of citation content. Journal of the American Society for Information Science and Technology, 64(7), 1490-1503.

Procesamiento del Lenguaje Natural, Revista nº 54, marzo de 2015, pp 53-60 recibido 24-11-14 revisado 27-01-15 aceptado 13-02-15

Anotación y representación temporal de *tweets* multilingües

Temporal annotation and representation of multilingual tweets

Asunción Vázquez-Méndez, Ana García-Serrano
ETSI Informática UNED
C/Juan del Rosal, 16
28040 Madrid
avazquez254@alumno.uned.es, agarcia@lsi.uned.es

Resumen: El tiempo es un elemento de importancia capital en todo espacio de información y Twitter no es una excepción. La explotación de la información temporal en tareas de recuperación y organización de información, tiene una larga tradición. Sin embargo, esta clase de enfoques, basados en contenido, no han sido muy explorados para el dominio de Twitter, y en consecuencia escasean los Corpus de *tweets* anotados con información temporal. En este artículo, se propone un modelo de anotación de la información temporal en el dominio de Twitter, basado en el Análisis de Conceptos Formales, en el que los atributos del contexto serán las expresiones temporales, eventos y tipos de eventos presentes en los *tweets*. Se define un Calendario especialmente adecuado a los fenómenos de conmemoración de aniversarios y fechas señaladas en Twitter, el Calendario Imaginario-Colectivo. El Corpus de estudio ha sido extraido de la colección de RepLab2013. Se incluye un completo análisis del mismo desde una perspectiva temporal.
Palabras clave: Información temporal, Anotación temporal de tweets, Representación de información basada en contenido

Abstract: Time is a crucial element in any space of information and Twitter is no an exception. Although the exploitation of temporal information in retrieval and organization tasks has a long tradition, content-based approaches have not been fully explored for Twitter and researchers lack of sufficient Corpus annotated with temporal information. In this paper, we propose a temporal document annotation model based on Formal Concept Analysis theory for Twitter domain. The tweets attributes defining the temporal context are the temporal expressions, the events and their types. It is also proposed a calendar especially suited to the phenomena of commemoration of anniversaries and dates in Twitter: The Social-Imaginary Calendar. The Corpus used to the experiments is a subset of the RepLab2013 collection. A detailed description of its temporal aspects is provided.
Keywords: Temporal information, Temporal annotation of tweets, Content-based information representation

1 Introducción

El tiempo juega un papel fundamental en todo espacio de información y Twitter[1] no es una excepción. A caballo entre red social y red de noticias, millones de personas comparten a diario, en forma de *tweet*, datos y opiniones relevantes para la reputación de personajes públicos, compañías y Gobiernos. Dos de los aspectos que mejor caracterizan Twitter son la actualidad de su contenido y el fenómeno de las *"tendencias"* (los temas más comentados en un momento dado) y ambos se miden en términos temporales.

En los últimos años Twitter ha acaparado un gran esfuerzo investigador tanto en lo que respecta a la detección de temas y tendencias como al análisis de sentimiento en los *tweets*, esto es, la polaridad de la opinión que reflejan sobre las entidades aludidas. Sin embargo, la información temporal de los *tweets*, fuera de la fecha de creación, esto es, la que se extrae del contenido, no ha recibido excesiva atención.

La explotación de esta información temporal "latente", la que se refiere a las expresiones temporales y eventos, presenta grandes desafíos y oportunidades (Alonso et al.,

[1] http://twitter.com

ISSN 1135-5948
© 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

2011)(Vicente-Díez y Martínez, 2009) y abre la puerta a la construcción de un contexto temporal complejo de los *tweets*, en el que se trata de poner en relación el momento en que son compartidos (determinado por su fecha de creación o *timestamp*) con el *momento en que sucede lo que se comparte*, para de esta manera definir qué significa la *"actualidad"* en los distintos temas que se tratan en Twitter, así como establecer eventuales relaciones de causalidad entre ellos.

Por otro lado, en un año cualquiera se dan multitud de eventos que, asociados a una fecha, permanecen en el *imaginario colectivo*, convirtiéndose en efemérides compartidas por grandes grupos de población, como es el caso del *"11 de septiembre"* a nivel mundial, del *"15 de marzo"* en España o del *"25 de junio"* para los seguidores de Michael Jackson. Con frecuencia estos eventos se reflejan en un aluvión de comentarios en Twitter, por lo que con un adecuado procesamiento de la semántica temporal, puede aprovecharse ese *"conocimiento colectivo"* para la confección de un calendario anotado con las fechas clave para una determinada entidad.

En este artículo, proponemos una aproximación a la representación de *tweets* de acuerdo a sus aspectos temporales (expresiones temporales y eventos). Las principales características de la propuesta son: utilizar el paradigma del Análisis de Conceptos Formales, como una línea de tiempo con múltiples granularidades, y articularse entorno a un calendario *dual*, basado en el *Calendario Gregoriano* y en uno definido por nosotros con el objeto de reflejar fechas señaladas, aniversarios y cualquier tipo de efeméride, el *Calendario "Imaginario-Colectivo"*.

El resto del artículo se organiza de la siguiente manera: en la Sección 2 se abordan los trabajos preliminares en anotación y representación de la información temporal asociada a un documento; en la Sección 3 se detalla la propuesta de caracterización temporal de un conjunto de documentos y se describe el desarrollo computacional para la integración y uso de herramientas y recursos de anotación, así como la explotación de estas anotaciones para la formación de los retículos temporales; en la Sección 4 se desarrollan los experimentos, se describe el corpus anotado, se extraen los descriptores y se construyen diferentes retículos en base a éstos; por último, en la Sección 5 se valoran los resultados obtenidos, se recapitula el trabajo presentado y se plantean algunas líneas de investigación abiertas que pueden ser abordadas en los próximos años.

2 Trabajos relacionados

La explotación de la información temporal contenida en un documento, cuenta con una larga tradición en la investigación, que experimentó el impulso definitivo en los años 90 con la celebración de la sexta conferencia MUC[2]. Fruto de esta investigación, que ha buscado mejorar, mediante conocimiento temporal, todo tipo de tareas en los sistemas de Recuperación de Información (detección y seguimiento de temas, búsqueda automática de respuestas, extracción automática de resúmenes, etc.) surgieron los esquemas de anotación temporal cuyo máximo exponente es TimeML (Pustejovsky et al., 2003), lenguaje que hoy es estándar. A la par que dichos esquemas, se fueron diseñando anotadores automáticos cada vez más complejos, como Tarsqi (Verhagen et al., 2005), que anota eventos y representa mediante grafos las relaciones temporales entre ellos, HeidelTime (Strötgen y Gertz, 2010), que es multilingüe y multidominio o Tipsem (Llorens, Saquete, y Navarro, 2010), también multilingüe (inglés).

Los paradigmas de visualización de la información temporal que se han utilizado van desde las líneas de tiempo y los grafos, a los mapas espacio-temporales y los grafos animados. Alonso, Gertz, y Baeza-Yates (2009) usan líneas de tiempo para agrupar resultados de búsqueda en base a las características temporales de los documentos. En un proceso similar al nuestro, extraen las expresiones temporales explícitas, implícitas y relativas de cada documento y las normalizan para crear un "perfil temporal". Este perfil se adapta a la granularidad que mejor define a la colección, según el calendario Gregoriano (Goralwalla et al., 2001) y se procede a su representacion en una línea de tiempo. Cada elemento de la línea de tiempo representará un *cluster*; obviamente puede haber *clusters* vacíos y también puede haber documentos que corresponden a más de un *cluster*, cuando tienen varias expresiones temporales; en este caso, se tratará de determinar el "*cluster principal*", atendiendo a la expresión

[2]Message Understanding Conference: www.cs.nyu.edu/cs/faculty/grishman/muc6.html

temporal predominante, esto es, que aparezca más en el documento.

En nuestra propuesta se extraen los eventos, además de las expresiones temporales, y se construye con todo ello un retículo basado en el Análisis de Conceptos Formales (en adelante, FCA, del inglés *"Formal Concept Analysis"*), que presenta dos ventajas principales: no requiere seleccionar un descriptor temporal principal, y permite el uso simultáneo de varias granularidades. En un enfoque similar, Ritter et al. (2012) extraen eventos de Twitter y los representan, de acuerdo a su contenido temporal, en un calendario que se actualiza en tiempo real[3].

Hasta donde sabemos, ningún otro trabajo ha utilizado retículos de conceptos para representación de información desde el punto de vista temporal. No obstante, el enfoque basado en FCA para modelar el contenido de conjuntos de *tweets* fue explorado por Castellanos, Cigarrán, y García-Serrano (2013) en tareas de detección de temas. Dichos autores modelan los *tweets* tomando como descriptores sus términos. Esta selección de los descriptores presenta un problema dual, por un lado hay una alta dependencia del dominio, por otro, el número de conceptos o posibles temas es potencialmente muy alto. Nuestra selección de descriptores trata de solventar estos problemas; el número de eventos anotados es sensiblemente menor que el conjunto de rasgos del vocabulario, además, las expresiones temporales son normalizables, y por tanto, independientes del dominio.

3 *Propuesta y desarrollo computacional*

3.1 Contexto temporal de una colección de documentos

Sea $\Delta = \{d1, ..., d_n\}$ una colección de documentos, su información temporal puede ser de los siguientes tipos (ver ejemplo en la Figura 1):

- **fechas de creación** o *timestamps* de los documentos,
- **expresiones temporales** contenidas en los documentos,
- **eventos** contenidos en los documentos, que se relacionan con las expresiones temporales y entre ellos.

[3]Twitter Calendar: http://ec2-54-170-89-29.eu-west-1.compute.amazonaws.com:8000/

Figura 1: Información temporal de un *tweet*

Construimos el conjunto $\tau = \Phi \bigcup T \bigcup E$ donde:

- $\Phi = \{f_1, ..., f_m\}$ es el conjunto de fechas de creación (distintas) de los documentos
- $T = \{t_1, ..., t_p\}$ es el conjunto de expresiones temporales normalizadas presentes en Δ
- $E = \{e_1, ..., e_q\}$ es el conjunto de los eventos presentes en Δ, lematizados

Se define el **contexto temporal** de Δ, $C_T := (\Delta, \tau, I)$ donde Δ es el conjunto de documentos, τ es el conjunto de atributos temporales e I es la relación binaria de incidencia que relaciona cada documento con los atributos que posee. Así construido, C_T es un **contexto formal**.

Consideramos que dos documentos pueden presentar similitud temporal, bien porque hayan sido creados en momentos temporales próximos bien porque el contenido descrito pertenezca al mismo evento o describa eventos que suceden en momentos cercanos. El retículo de conceptos formales $\beta(C_T)$ realizará un agrupamiento de los documentos que tenga en cuenta esta doble dimensión: creación/contenido.

3.2 Calendario *Imaginario-Colectivo*

La normalización de las expresiones temporales, requiere de la definición de un *"calendario"*. Lo tradicional es utilizar el calendario **Gregoriano**, que presenta las granularidades: *año, mes, día, hora, minuto* y *segundo*, con las relaciones "$\gg$" (más *gruesa*) y "$\ll$" (más *fina*) (Goralwalla et al., 2001):

$G_{\text{año}} \gg G_{mes} \gg ... \gg G_{segundo}$

La elección de una granularidad concreta no es necesaria para FCA, por el contrario, podemos representar una expresión en múltiples sistemas, con el objeto de no perder ninguna información. Por ejemplo, dadas las expresiones *"1969"*, *"1967"* y *"1967-06-01"*, si fuéramos a representar los docu-

mentos en una línea de tiempo, podríamos considerar que la granularidad más adecuada es $G_{\text{año}}$, ya que solo la última expresión se puede expresar en granularidades más finas. Sin embargo en FCA, podemos elegir el conjunto de descriptores sin perder información y a la vez mantener la relación entre dos documentos que hacen referencia al mismo año: *"1969"*, *"1967"*, *"1967-06"* y *"1967-06-01"*. El documento que posee la expresión temporal *"1967-06-01"* tendrá 3 descriptores.

Hay un tipo de expresiones que, por no estar completamente determinadas, no se pueden representar en el Calendario Gregoriano, pero tienen dimensión temporal, aunque su carácter puede ser estacional o periódico. Hablamos de expresiones del tipo *"día de Navidad"*, *"Diciembre"* o *"invierno"*, cuando **no se refieren a un año concreto**; su valor en lenguaje TimeML sería, respectivamente: "XXXX-12-25", "XXXX-12" y "XXXX-XX-XXWI". Estas expresiones no se pueden representar en una línea de tiempo al uso, sin embargo tienen cabida natural en FCA, y son relevantes en ciertos dominios, como las redes sociales, donde es frecuente hacer alusiones a todo tipo de Aniversarios o fechas señaladas.

Definimos el ***Calendario Imaginario-Colectivo*** como la terna:

$$C_{IC} = (A, \varrho, \varphi)$$

donde A representa un año natural cualquiera, $\varrho = (A_m, A_d)$ es el conjunto de granularidades (mes y día) y φ la función de conversión obvia:

$\varphi(XXXX - 12 - 25) = XXXX - 12$

Con esta definición no pretendemos capturar el significado de cada fecha para cada persona, sino ese conjunto de efemérides compartidas por un conjunto concreto de la sociedad que puede ser los seguidores de los Beatles, los habitantes de un país o la población mundial.

3.3 Desarrollo computacional

Para la construcción del retículo temporal asociado a una colección de documentos, es necesario localizar y extraer las expresiones temporales y eventos presentes en ellos y procesarlas adecuadamente para obtener el conjunto de atributos. Se ha desarrollado un entorno computacional que integra herramientas y recursos Web de anotación y FCA; su arquitectura se representa en la Figura 2 y consta de las siguientes fases:

- **Preprocesado** Se preparan los documentos para ser anotados, mediante la eliminación de los caracteres no permitidos por XML. Se eliminan también las *urls*, para evitar la anotación de fechas en las rutas de carpetas y los *emoticonos* con símbolos numéricos (como "<3"), etc.
- **Anotación** Se anotan los documentos con HeidelTime. El subconjunto de documentos en inglés se anota también con Tarsqi. No es necesario utilizar un reconocedor de idioma pues los *tweets* de RepLab están etiquetados con esta información.
- **Descriptores** Se parsean los archivos de salida de Tarsqi y HeidelTime, extrayendo, para cada documento, las fechas de creación, las expresiones temporales, los eventos y su tipo. Tras descartar las expresiones poco frecuentes o indeseadas, se enriquece el conjunto de expresiones, añadiendo su equivalencia en todas las posibles granularidades de los Calendarios definidos en la propuesta. Los eventos se lematizan usando la librería **nltk.stem.wordnet** [4]. Finalmente se genera la tabla que representa al contexto formal, constituido por los documentos (objetos) y sus descriptores temporales (atributos).
- **Retículo de conceptos** Una vez contruido el contexto formal, recurrimos al entorno de FCA Concept Explorer (Yevtushenko et al.,), para el cálculo del conjunto de conceptos formales y la representación del retículo.

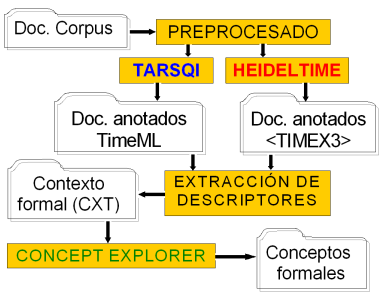


Figura 2: Diagrama funcional implementado

[4] http://www.nltk.org/

4 *Experimentación*

4.1 Anotación y estudio lingüístico del Corpus

El corpus de experimentación es un subconjunto de la colección de *tweets* de **RepLab2013**.

RepLab (Amigó et al., 2013) es un Foro de Evaluación de sistemas de gestión de reputación online. Para la edición de 2013 se seleccionaron 61 entidades de cuatro temáticas diferentes (música, universidades, banca y automóviles) y por cada una, se recogieron varias decenas de miles de *tweets*, en inglés y en español, de un periodo comprendido entre el 1 de junio y el 31 de diciembre de 2012.

Se formaron para cada entidad un conjunto de entrenamiento (unos 700 *tweets*) y uno de validación (unos 1500 *tweets*) y se anotaron los conjuntos de entrenamiento con información relativa a temática, relación con la entidad y posibles implicaciones que pudiera tener el contenido del *tweet* para la reputación de la misma. Se pretendía que los conjuntos de datos de entrenamiento y validación estuvieran formados por *tweets* distantes en el tiempo, esto es, con una brecha temporal entre ellos de varios meses. Para ello, se asignaron los primeros *tweets* al conjunto de entrenamiento y los últimos al de validación.

Para llevar a cabo la experimentación de esta propuesta, se ha elegido la entidad **"Beatles"**. El **"Corpus Beatles"**, descrito con detalle en (Vázquez-Méndez, 2014), está formado por un conjunto de entrenamiento de 701 *tweets* (538 en inglés, 163 en español) y por uno de validación de 1531 *tweets* (1130 en inglés, 401 en español).

4.1.1 Temporalizacion y temática

El periodo temporal abarcado por ambos conjuntos es bastante pequeño. El 98 % de los *tweets* de entrenamiento fueron publicados entre el 1 y el 5 de junio, mientras que el mismo porcentaje de los *tweets* de validación, lo fue entre el 22 y el 31 de Diciembre. Ubicar la colección temporalmente es fundamental, tanto para elegir la granularidad más adecuada, como para extraer conclusiones respecto a eventos cuyo periodo de vigencia coincida con el de los datos disponibles.

En cuanto a la temática, atendiendo a las anotaciones del conjunto de entrenamiento, se puede ver que los temas más habitualmente tratados están relacionados con comentarios de fans (22 %), letras y vídeos de canciones (23 %) y referencias varias a productos (ediciones remasterizadas de discos, etc.) (26 %). También se detectan varios temas que se presumen de actualidad por referirse a un evento concreto que se produce en una ventana temporal de unos pocos días respecto a la publicación del *tweet*, como son el Concierto del Jubileo (5 %) y el 45º aniversario del lanzamiento del álbum Sgt Pepper (3 %).

4.1.2 Expresiones temporales

Se anota el corpus Beatles con HeidelTime, lo que da como resultado un total de 287 *tweets*, el 16 %, con presencia de expresiones temporales (etiquetadas con TIMEX3). Se trata pues de un porcentaje pequeño pero significativo. La tipología de expresiones se resume en la Figura 3.

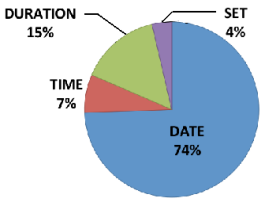


Figura 3: Tipos de TIMEX3

Predominan las expresiones de tipo "fecha" (DATE), en particular las que se refieren a tiempo presente:

- *Vamos a soñar imaginen q todos los integrantes d los beatles estuvieran vivos sería* <TIMEX3 type="DATE" value="2012-06-04">***hoy***</TIMEX3> *la locura en el concierto en honor a la reina*

Las referencias al pasado suelen aparecer de forma explícita, en *tweets* donde se data el contenido (una canción, un vídeo) o se menciona un evento de cierta relevancia, como la publicación de un álbum, un concierto, etc:

- <TIMEX3 type="DATE" value="1967-06-01">***June 1, 1967***</TIMEX3> - *The Beatles release Sgt. Pepper's Lonely Hearts Club Band. The album is certified gold its first day in stores. #TheBeatles*

El segundo tipo en importancia es "duración" (DURATION); suele tener mucho que ver con fechas señaladas en la historia de la entidad, esto es, con aniversarios o periodos en los que destaca algún aspecto de la entidad:

- *Hoy se cumplen*<TIMEX3 type="DURATION" value="P45Y"> ***45 años***</TIMEX3> *del estreno de Sgt. Pepper's Lonely Hearts Club Band, álbum de The Beatles.*

4.2 Eventos

El subcorpus de *tweets* en inglés, se anota también con Tarsqi (etiquetas TIMEX3, EVENT y LINK). Al contrario de lo que ocurría para las expresiones temporales, el porcentaje de anotación es bastante alto (62 %) y muchas veces se anotan varios eventos por *tweet* (1865 eventos anotados en 843 *tweets*). Tras el proceso de lematización, los eventos a considerar como descriptores se reducen considerablemente (646 lemas distintos).

Las palabras etiquetadas como eventos son mayoritariamente verbos, aunque también hay sustantivos o adjetivos. La influencia del dominio se hace notar en la presencia de verbos como *"listen"* o *"play"*:

- *This morning I have mostly been* <EVENT class="PERCEPTION"> ***listening***</EVENT> *to 'The Beatles - White Album'*
- *TONIGHT @thepeel *Beatles Tribute Band* Abbey Road LIVE! Sgt Pepper 45th Anniversary* <EVENT class="OCCURRENCE"> ***Show***</EVENT> *! $20! #avl #avlent #avlmusic #Asheville*
- *CHOON! The Beatles really were* <EVENT class="STATE" > ***brilliant***</EVENT> *! #jubileeconcert*

En cuanto a clase de evento, la inmensa mayoría son de ocurrencia (OCCURRENCE); también hay una presencia significativa de eventos de estado (STATE, I_STATE) y de percepción (PERCEPTION) (Figura 4).

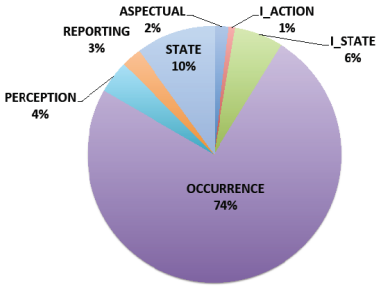


Figura 4: Clases de eventos (EVENT)

4.3 Elección de descriptores y retículos temporales

El conjunto de descriptores τ, tal como se define en la propuesta, constituye el contexto temporal más completo para un conjunto de *tweets*. Hemos realizado varios experimentos, para distintas configuraciones de τ, aumentando paulatinamente la complejidad del contexto (ver Tabla 1) y la información representada.

Experimento	**Descriptores**
I	$\tau = \Phi$
II	$\tau = \Phi \bigcup T$
III	$\tau = \Phi \bigcup T \bigcup E$
IV	$\tau = \Phi \bigcup T \bigcup clases(E)$

Tabla 1: Elección de descriptores

El retículo formado en el Experimento I (Figura 5) agrupa los *tweets* por día, mes y año, dando idea del grado de actividad de los usuarios en relación a la entidad. El tamaño de los nodos del retículo va en relación al número de objetos del concepto; por cada uno se indican los atributos (en color gris) y el número y porcentaje de *tweets* contenidos (en color blanco).

Aunque para extraer conclusiones sobre periodos de interés deberíamos contar con ventanas temporales más amplias, sí se pueden detectar alteraciones significativas. Por ejemplo, el día 4 de junio la actividad fue notablemente más alta que en el resto. La razón fue la celebración del concierto del Jubileo en honor a la Reina de Inglaterra, en el que participó Paul Mcartney y se cantaron varias canciones de los Beatles, lo que animó a los *twitteros* a comentar. Obviamente el retículo no da esta información tan concreta, pero nos alerta de que algún evento importante puede haber sucedido.

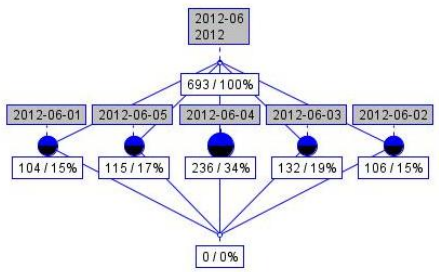


Figura 5: Diagrama de Hasse ExpI

Al introducir en el contexto las expresiones temporales presentes en el contenido (Experimento II), ya se crean agrupaciones de *tweets* que comparten ciertos atributos. Por ejemplo, el retículo ha aislado aceptablemente parte del tema *"Sgt Pepper"*. Este tema se corresponde con el nombre de un álbum de los Beatles lanzado el 1 de junio de 1967; al conmemorarse el aniversario de su lanzamiento en 2012, se convirtió en un tema comentado en Twitter. En la Figura 6 se muestra en detalle el concepto en que se basa la agrupación: prácticamente todos los *tweets* que hacen referencia al año 1967 lo hacen al álbum Sgt Pepper y todos los *tweets* escritos el 1 de junio que hacen referencia a 1967, también.

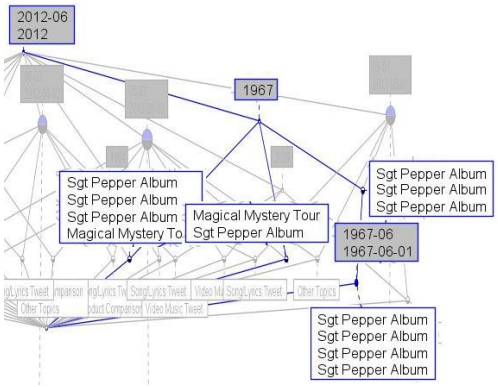


Figura 6: ExpII Tema *SgtPepper*

A continuación, agregamos las expresiones lematizadas de los eventos (Experimento III). Las reglas de asociación obtenidas por Concept Explorer permiten identificar agrupaciones de *tweets* informativas. Así se pone de manifiesto que el evento *"release"* en *tweets* con el atributo *"1967"*, se refiere exclusivamente al lanzamiento del álbum Sgt. Pepper y que todos los *tweets* escritos el 4 de junio, que contienen el evento *"sing"*, tratan del concierto del Jubileo (ver Figuras 7,8).

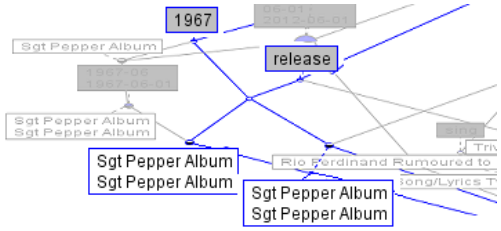


Figura 7: ExpIII Tema *SgtPepper*

Finalmente, en el Experimento IV, se decide sustituir en τ, los eventos por sus clases. De esta forma, se gana independencia respecto al dominio y se pueden detectar otro tipo de relaciones, como las que existen entre algunas clases de eventos y algunos tipos de *tweets*. Es el caso de los eventos de percepción (PERCEPTION), muy relacionados con la expresión de opiniones en distintas formas: comentarios de fans, comentarios sobre productos,etc. En la Figura 9 se muestra el retículo de conceptos para este tipo de eventos; como se ve, las percepciones sobre el concierto del Jubileo, corresponden todas al día de su celebración.

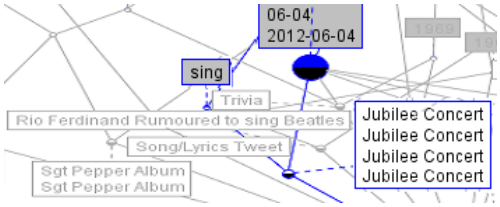


Figura 8: ExpIII Tema *Jubilee Concert*

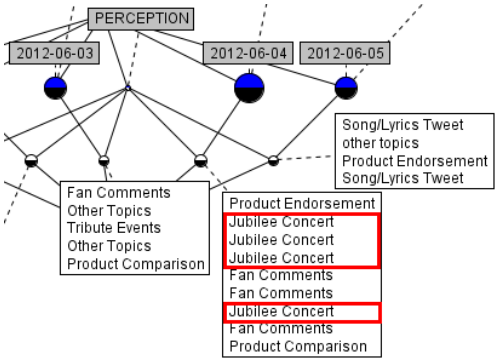


Figura 9: ExpIV Tema *Jubilee Concert*

5 Conclusiones y trabajos futuros

La explotación de la información temporal de los *tweets*, mediante su traducción en descriptores, ha permitido detectar eventos anclados a una fecha, como el aniversario del lanzamiento del album *Sgt. Pepper* y el concierto del Jubileo. Es un buen resultado, teniendo en cuenta, que el periodo de tiempo abarcado era muy reducido.

Por otro lado, la gran parte de *tweets* anotados temáticamente, lo era de un modo genérico: comentarios de fans, comentarios sobre productos, etc. Se consiguió cierto nivel de agrupación para tipos concretos de *tweets* asociados a eventos de PERCEPCIÓN; de todas formas, el modelo intentará desagregar los temas, buscando qué ha motivado cada *tweet, por qué se ha escrito en el momento en que se ha escrito*. Estas subdivisiones no

pueden ser evaluadas con las anotaciones del Corpus de las que disponemos.

La propuesta que aquí se detalla ha querido poner de manifiesto que en los *tweets*, pese a lo reducido de su extensión, hay información temporal latente que puede contribuir a mejorar el rendimiento de los sistemas en tareas como la detección y seguimiento de temas o la agrupación de documentos en Twitter.

La explotación de esta información temporal, requiere de una anotación acorde a las peculiaridades del dominio. Anotadores como Heideltime pueden configurarse para anotar textos en lenguaje *"coloquial"* para el idioma inglés, pero se necesitan Corpus de *tweets* anotados, tanto en inglés como en español, que puedan servir de *gold-standard*.

Otra línea de trabajo futuro es la anotación de *hashtags* con información temporal, como *#15m* o *#jubileeconcert*.

Bibliografía

Alonso, O., M. Gertz, y R. Baeza-Yates. 2009. Clustering and exploring search results using timeline constructions. En *Proceedings of the 18th ACM conference on Information and knowledge management*, páginas 97–106.

Alonso, O., J. Strötgen, R. Baeza-Yates, y M. Gertz. 2011. Temporal Information Retrieval: Challenges and Opportunities. *TWAW*, 11:1–8.

Amigó, E., J. Carrillo de Albornoz, I. Chugur, A. Corujo, J. Gonzalo, T. Martín-Wanton, E. Meij, M. de Rijke, y D. Spina. 2013. Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems. En *CLEF*, volumen 8138 de *LNCS*, páginas 333–352. Springer.

Castellanos, A., J. Cigarrán, y A. García-Serrano. 2013. Modelling Techniques for Twitter Contents: A Step beyond Classification based Approaches. En *Working Notes of the CLEF 2013*.

Goralwalla, I., Y. Leontiev, M.T. Özsu, D. Szafron, y C. Combi. 2001. Temporal granularity: Completing the puzzle. *Journal of Intelligent Information Systems*, 16(1):41–63.

Llorens, H., E. Saquete, y B. Navarro. 2010. TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2. En *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, páginas 284–291.

Pustejovsky, J., M.J. Castaño, R. Ingria, R. Saurí, R.J. Gaizauskas, A. Setzer, G. Katz, y D.R. Radev. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. *New directions in question answering*, 3:28–34.

Ritter, A., O. Etzioni, S. Clark, y others. 2012. Open domain event extraction from Twitter. En *Proceedings of the 18th ACM SIGKDD International conference on Knowledge discovery and data mining*, páginas 1104–1112. ACM.

Strötgen, J. y 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.

Vázquez-Méndez, A. 2014. *Explotación de la Información Temporal en Twitter para la organización de tweets*. Tesis de Máster, UNED.

Verhagen, M., I. Mani, R. Saurí, R. Knippen, S.B. Jang, J. Littman, A. Rumshisky, J. Phillips, y J. Pustejovsky. 2005. Automating temporal annotation with TARSQI. En *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, páginas 81–84.

Vicente-Díez, M.T y P. Martínez. 2009. Temporal semantics extraction for improving web search. En *20th International Workshop on Database and Expert Systems Application*, páginas 69–73. IEEE.

Yevtushenko, S., J. Tane, T.B. Kaiser, S. Obiedkov, J. Hereth, y H. Reppe. ConExp - The Concept Explorer. *URL: http://conexp.sourceforge.net*.

Procesamiento del Lenguaje Natural, Revista nº 54, marzo de 2015, pp 61-68 recibido 24-11-14 revisado 27-01-15 aceptado 10-02-15

TASS 2014 - The Challenge of Aspect-based Sentiment Analysis*

TASS 2014 - El Reto del Análisis de Opiniones a nivel de aspecto

Julio Villena Román,
Janine García Morera
Daedalus, S.A.
E-28031 Madrid, Spain
{jvillena, jgarcia}@daedalus.es

Eugenio Martínez Cámara,
Salud M. Jiménez Zafra
SINAI Research Group - University of Jaén
E-23071 Jaén, Spain
{emcamara, sjzafra}@ujaen.es

Resumen: El análisis de la reputación y el análisis de opiniones son dos tareas que están en boga actualmente. Pero esa moda viene justificada por la necesidad, cada vez más acuciante, de conocer la orientación de las opiniones que se publican diariamente en Internet. TASS es un taller de trabajo que tiene como fin fomentar la investigación en el descubrimiento de la orientación de la opinión de textos en español publicados en Internet. En este artículo se describe la tercera edición de TASS, en el que se han mantenido dos tareas propuestas en las dos ediciones anteriores, y se han planteado otras dos nuevas relacionadas con el análisis de opiniones a nivel de aspecto, y que se encuentran circunscritas en el fenómeno de la Televisión Social.
Palabras clave: TASS 2014, Análisis de Opiniones, Análisis de Opiniones a nivel de Aspecto, Televisión Social

Abstract: Currently, reputation and sentiment analysis are trendy tasks. However, the interest on these two tasks is growing by the need of knowing the polarity of the opinions published on the Internet. TASS is a workshop whose goal is to boost the research on sentiment analysis in Spanish. Hereinafter the third issue of TASS is described, in which four tasks have been proposed. Two of the proposed tasks are known by former participants and the other two ones are new. These new tasks are related to sentiment analysis at entity level, and they are circumscribed on the Social TV phenomenon.
Keywords: TASS 2014, Sentiment Analysis, Aspect Based Sentiment Analysis, Social TV.

1 Introduction

Workshop on Sentiment Analysis at SEPLN (TASS, in Spanish) is an experimental evaluation workshop on reputation and sentiment analysis (SA) focused on Spanish language, organized as a satellite event of the SEPLN Conference. After two successful editions in 2012 (Villena-Román et al., 2013) and 2013 (Villena-Román et al., 2014), TASS 2014[1] was held on September 16th, 2014 at University of Gerona, Spain.

The long-term objective of TASS is to foster the research on the field of reputation, i.e., the process of tracking, investigating and reporting an entity's actions and other entities' opinions about those actions, in Spanish language. As a first approach, reputation analysis encompasses at least two technological aspects: SA and text classification.

Nowadays, SA means the computational treatment of opinion, sentiment and subjectivity in text (Pang and Lee, 2008). It is a major technological challenge and the task is so hard that even humans often disagree on the sentiment of a given text, as issues that one individual may find acceptable or relevant may not be the same to others. And the shorter the text is (for instance, Twitter messages), the harder the task becomes.

On the other hand, automatic text classification (or categorization) is used to guess the topic of the text, among those of a predefined

* This work has been partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER), ATTOS project (TIN2012-38536-C03-0), Ciudad2020 (INNPRONTA IPT-20111006), MULTIMEDICA (TIN2010-20644-C03-01) from the Spanish Government, AORESCU project (P11-TIC-7684 MO) from the regional government of Junta de Andalucía, MA2VICMR (S2009/TIC-1542) from the regional government of Comunidad de Madrid and CEATIC-2013-01 project from the University of Jaén.

[1] www.daedalus.es/TASS2014

ISSN 1135-5948
© 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

set of categories, so as to be able to assign the reputation level into different axis or points of view of analysis. Text classification techniques, albeit studied for a long time, still need more research effort to be able to build complex models with many categories with less workload and increase the precision and recall of the results. In addition, these models should deal with specific text features in social media messages.

Up to now, TASS has proposed analyses at document level (tweet level), but the SA research community are beginning to go a step further, related to the fact that the society needs a fine-grained study of people attitude expressed on a tweet. Aspect-Based Sentiment Analysis (ABSA) is the task that is concerned with the extraction and classification of opinions on a specific entity. An entity can be decomposed into several parts or aspects, so that can be seen as a hierarchical structure whose head is the entity. ABSA is not only focused on opinions on entities, but also each of the aspects that are part of an entity. In a pragmatic way, an ABSA system does not take into account the hierarchical relation between the entity and the aspects, and both are considered in the same way. ABSA encompasses two subtasks, aspect extraction and aspect sentiment classification. The first one empathizes on the identification of the aspects presented on a text, and the second one comprehends the classification of the attitude of the opinion holder about the aspect.

The previous paragraphs described trendy, hard, and interesting tasks that are basic for a posterior study of reputation. Within this context, the aim of TASS is to provide a forum for discussion the latest research work in these fields. The setup is based on a series of challenge tasks intended to provide a benchmark forum for comparing different approaches. Moreover, the aim of TASS is to provide a common reference dataset for the research community, so it is generated and open-release the corpus fully tagged. Polarity classification and topic classification are two fixed tasks of TASS, but due to the relevance of ABSA, the 2014 edition of TASS has included two new tasks, aspect identification and aspect-based polarity classification, which is focused on the context of Social TV.

The rest of the paper is organized as follows. Section 2 describes the corpora provided to participants and used for the challenge tasks. The third section describes the different tasks proposed in the 2014 edition. Section 4 and 5 describes the participants and the analysis of the results, and the last section draws some conclusions and future directions.

2 Corpus

Experiments were based on two corpora. After the workshop, both were published only for research purposes.

2.1 General corpus

The general corpus, which is the same used in the previous two editions, contains over 68,000 tweets gathered between November 2011 and March 2012. The tweets are written in Spanish by about 150 well-known personalities and celebrities of the world of politics, economy, communication, mass media and culture.

The general corpus was divided into two sets: training (10%) and test (90%). Table 1 shows a summary of the training and test corpora provided to participants.

Attributes	**Value**
Tweets	68,017
Tweets (test)	60,798 (89%)
Tweets (train)	7,219 (11%)
Topics	10
Tweet languages	1
Users	154
User types	3
User languages	1
Date start (train)	2011-12-02 T00:47:55
Date end (train)	2012-04-10 T23:40:36
Date start (test)	2011-12-02 T00:03:32
Date end (test)	2012-04-10 T23:47:55

Table 1: Features of the General Corpus

Each message in both the training and test set was tagged with its global polarity, indicating whether the text expresses a positive, negative or neutral sentiment, or no sentiment at all. Five levels have been defined: strong positive (P+), positive (P), neutral (NEU), negative (N), strong negative (N+) and one additional no sentiment tag (NONE).

Furthermore, the level of agreement of the expressed sentiment within the text was also included, to clarify whether a neutral sentiment comes from neutral keywords (AGREE-

MENT) or else the text contains positive and negative sentiments at the same time (DISAGREEMENT).

On the other hand, a selection of a set of topics was made based on the thematic areas covered by the corpus, such as politics, literature or entertainment. Each message in both the training and test set was assigned to one or several of these topics.

All tagging was carried out semi automatically: a baseline machine learning model was first run (Villena-Román et al., 2011) and then all tags were manually checked by two human experts. For polarity at entity level, due to the high volume of data to check, this tagging was done just for the training set.

2.2 Social-TV corpus

Social-TV corpus was collected during the 2014 Final of Copa del Rey championship in Spain, between Real Madrid and F.C. Barcelona. It was played on 16 April 2014 at Mestalla Stadium in Valencia.

Over 1 million of tweets were collected from 15 minutes before to 15 minutes after the match. After filtering useless information, tweets in other languages than Spanish, a subset of 2773 was selected.

All tweets have been manually tagged with the aspects of the expressed messages and its sentiment polarity. Tweets may cover more than one aspect.

The general defined aspects were: *afición* (fans), *árbitro* (referee), *autoridades* (political authorities), *entrenador* (coach), *equipo* (team), *jugador* (player), *partido* (game) and *retransmisión* (broadcasting).

Some of the detailed aspects were: Equipo-Barcelona, Equipo-Real_Madrid, Jugador-Isco, Jugador-Dani_Álves, and the other players.

Sentiment polarity has been tagged from the point of view of the person who writes the tweet, using 3 levels: P (positive), NEU (neutral) and N (negative). No distinction was made in cases when the author does not express any sentiment or when he/she expresses a no-positive or no-negative sentiment.

The Social-TV corpus has been randomly divided into two sets: training (1773 tweets) and test (1000 tweets), with a similar distribution of both, aspects and sentiments. The training set was released with the aim of the participants could train and validate their models. The test corpus was provided without any tagging and was used to evaluate the results provided by the different systems. The list of the 31 aspects that have been defined can be read at the workshop webpage.

```
<tweet id="456544898791907328">
  <sentiment aspect="Equipo-Real_Madrid
      " polarity="P">#HalaMadrid</
      sentiment> ganamos sin <sentiment
       aspect="Jugador-
      Cristiano_Ronaldo" polarity="NEU"
      >Cristiano</sentiment>. .perdéis
      con <sentiment aspect="Jugador-
      Lionel_Messi" polarity="N">Messi<
      /sentiment>. Hala <sentiment
      aspect="Equipo-Real_Madrid"
      polarity="P">Madrid</sentiment>!
      !!!!!
</tweet>
<tweet id="456544898942906369">
  @nevermind2192 <sentiment aspect="
      Equipo-Barcelona" polarity="P">
      Barça</sentiment> por siempre!!
</tweet>
<tweet id="456544898951282688">
  <sentiment aspect="Partido" polarity=
      "NEU">#FinalCopa</sentiment> Hala
       <sentiment aspect="Equipo-
      Real_Madrid" polarity="P">Madrid<
      /sentiment>, hala <sentiment
      aspect="Equipo-Real_Madrid"
      polarity="P">Madrid</sentiment>,
      campeón de la <sentiment aspect="
      Partido" polarity="P">copa del
      rey</sentiment>
</tweet>
```

Figure 1: Sample tweet (Social TV corpus)

Figure 1 shows the information of three sample tweets in the training set.

3 Description of tasks

The main goal of TASS is to boost the research on reputation and SA in Spanish. With the aim of reaching it, the organization of TASS always proposed four tasks, two of them that have the purpose of analysing the evaluation of the investigation on SA and Topic Classification, and another two ones that are usually linked with needs of the society, which are usually voiced by a business demand. The two fixed tasks of TASS are Sentiment Analysis and Topic Classification at document level, which will be described hereinafter.

2014, in Spain, has been the year in which the TV channels have greatly taken advantage of the social networks with the objective of increasing the participation of the viewers in the TV shows. Last July, the CEO of Twitter Spain, José López de Ayala, asserted that the 66% of mobile phone users publish

tweets while they are watching TV. Also, he pointed out that the TV ads, whose last image is a Twitter hashtag, they achieve to increase by 60% the number of tweets related to that tag. Therefore, analyzing the sentiment related to the phenomenon of Social TV was a great candidate to be the target of a task. The level of analysis required by a reputation or SA system in the context of Social TV is deeper than the proposed one in the traditional tasks of TASS, in plain English, Social TV needs a fine-grained analysis. The level of analysis needed in a Social TV context is entity or aspect level. This is the reason why this year two new tasks were proposed. The new tasks required the development of SA systems at aspect level.

3.1 Task 1: SA at Document Level

The first of the two fixed tasks of TASS is the performing SA at document level. In the context of the workshop the task proposed the development of polarity classification systems at tweet level, in other words, build systems to classify tweets in several predefined polarity classes. Six is the number of polarity labels (P+, P, NEU, N, N+, NONE) in which the systems had to classify the tweets of the general corpus. But the systems had to be prepared to classify tweets in four labels (P, NEU, N, NONE), because the performance of the systems is evaluated in an environment of six classes and four classes.

Accuracy was used for ranking the systems. Precision, Recall and F1-measure will be used to evaluate each individual category.

Results were submitted in a plain text file with the following format:

```
tweetid \t polarity
```

where polarity could be: P+, P, NEU, N, N+ and NONE for the 6-labels case; P, NEU, N and NONE for the 4-labels case.

The same test corpus of previous years was used for the evaluation, to allow comparison between systems. Obviously, participants were not allowed to use any test data to train their systems. However, to deal with the problem reported last years of the imbalanced distribution of labels between the training and test set, a new selected test subset containing 1000 tweets with a similar distribution to the training corpus was extracted and used for an alternate evaluation of the performance of systems.

3.2 Task 2: Topic Classification

The challenge of this task is to automatically identify the topic of each message in the test set of the General corpus. Participants could use the training set of the General corpus to train and validate their models.

Participants were expected to submit up to 3 experiments, each one in a plain text file with the following format:

```
tweetid \t topic
```

A given tweet ID can be repeated in different lines if it is assigned more than one topic.

Micro averaged precision, Recall and F1-measure calculated over the full test set will be used to evaluate the systems. Systems were ranked by F1. To allow the comparison with previous years, the same test corpus will be used for the evaluation. Again, participants were not allowed to use any test data to train their systems.

3.3 Task 3: Aspect Detection

The main objective of this task is the automatic identification of the different aspects expressed by users, among a predefined list, in their opinions expressed in Twitter about a given topic. For example, in the following tweet:

```
CR7 jugó bien, Messi no, el Madrid
se mereció la victoria
(CR7 played well, Messi didn't, the
Madrid team deserved to win)
```

Three aspects can be identified CR7 as the player Cristiano Ronaldo, Messi as the player Lionel Messi and Madrid as the Real Madrid team. This task is a multi-label classification and tweets can have more than one aspect, as shown in the example.

A new Social-TV corpus was delivered and used for the training and evaluation of the systems (see description above).

Participants are expected to submit up to 3 experiments, each in a plain text file with the following format:

A given tweet id can be repeated in different lines if it is assigned more than one

```
tweetid \t aspect
```

aspect. We consider an aspect as the minimum set of words, not the detected terms or fragment in neither the text nor its offsets.

As evaluation measures micro averaged precision, recall and F1 were used, calculated over the full test set. The final list of participants was ranked by F1.

3.4 Task 4: Aspect-based SA

This task was similar to the first one, but sentiment polarity (using 3 levels) should be determined at aspect level of each tweet in the Social-TV corpus (fine-grained polarity detection). They worked with the aspects detected in the previous task. Again, participants were provided with this Social-TV corpus to train and evaluate their models. Aspects were tagged in the corpus to make participant focus on the polarity classification and not on aspect identification. The complex of the task arises from the fact that tweets can contain more than one sentence with more than one aspect per sentence, so more advance text processing techniques are needed.

Participants were expected to submit up 3 experiments, each in a plain text file with the following format:

```
tweetid \t aspect \t polarity
```

Allowed polarity values were P, NEU and N.

Accuracy, micro averaged Precision, Recall and F1-measure was used to evaluate the systems, considering a unique label combining aspect-polarity. Systems were ranked by F1.

4 Participants

This year 35 groups registered (as compared to 31 groups last year) and finally 7 groups (14 last year) sent their submissions. The list of active participant groups is shown in Table 2, including the tasks in which they have participated.

Along with the experiments, all participants were invited to submit a paper with the description of their experiments and the analysis of the results. These papers were reviewed by the program committee and were included in the workshop proceedings. References are listed in Table 3.

Group	**1**	**2**	**3**	**4**
LyS	✓	✓	✓	✓
SINAI-ESMA	✓			
Elhuyar	✓			
SINAIword2vec	✓			
JRC	✓			
ELiRF-UPV	✓	✓	✓	✓
IPN	✓	✓		
Total groups	7	3	2	2

Table 2: Participant groups

Vilares et al. (2014) used a machine learning approach, using several linguistic resources and other information extracted from the training corpus to feed to a supervised classifier. With respect to task 3, they developed a naive approach, collecting a set of representations to identify the predefined aspects requested by the organizers. Jiménez Zafra et al. (2014) developed an unsupervised classification system which is based on the use of an opinion lexicon, and on the application of a syntactic heuristic for identifying the scope of Spanish negation words. San Vicente Roncal and Saralegi Urizar (2014) implemented a SVM algorithm that combines the information extracted from polarity lexicons with linguistic features. Montejo Ráez, García Cumbreras, and Díaz-Galiano (2014) used supervised learning with SVM over the summatory of word vectors in a model generated from the Spanish Wikipedia. Perea-Ortega and Balahur (2014) focused on different feature replacements carried out for both the development and test data sets provided. The replacements performed were mainly based on repeated punctuation signs, emoticons and affect words, by using an in-house built dictionary for SA. Then, they applied a machine learning approach to get the polarity of the tweets. Hurtado and Pla (2014) adapted the tweet tokenizer Tweetmotif (Connor, Krieger and Ahn, 2010) and they used Freeling (Padro y Stanilovsky, 2012) as lemmatizer, entity recognizer and morphosyntactic tagger. Hernández Petlachi and Li (2014) proposal is based on semantic ap-

proaches with linguistic rules for classifying polarity texts in Spanish. Polarity classification in the words is done according to a dictionary of semantic orientation where each term is labeled with a use value and emotional value, along with linguistic rules to solve various constructions that could affect the polarity of text.

Group	Report
LyS	(Vilares et al., 2014)
SINAI-ESMA	(Jiménez Zafra et al., 2014)
Elhuyar	(San Vicente Roncal and Saralegi Urizar, 2014)
SINAIword2vec	(Montejo Ráez, García Cumbreras, and Díaz-Galiano, 2014)
JRC	(Perea-Ortega and Balahur, 2014)
ELiRF-UPV	(Hurtado and Pla, 2014)
IPN	(Hernández Petlachi and Li, 2014)

Table 3: Participant reports

5 Results

After the submission deadline, runs were collected and checked and results were evaluated and made available to the participants using an automated web page in the password protected area in the website. Results for each task are described hereinafter.

5.1 Task 1: Sentiment Analysis at Document Level

Task 1 includes the experiments using the full test set and using the selected 1k test set.

Thirty-two runs for 5-level evaluation were submitted by 7 different groups. Results for the best-ranked experiment from each group are listed in the tables below. All tables show the precision (P), recall (R) and F1 value achieved in each experiment. Table 4 considers 5 polarity levels, with the whole test corpus. Accuracy values range from 0.64 to 0.37.

As previously described, an alternate evaluation of the performance of systems was done using a new selected test subset containing 1000 tweets with a similar distribution to the training corpus. Results are shown also in next Table 4. Accuracy values range from 0.48 to 0.33 (1k test corpus). Figures are much lower as compared to the previous evaluation, thus showing a high bias in the semi-automatic tagging of the whole test corpus.

In order to perform a more in-depth evaluation, results are calculated considering the classification only in 3 levels (POS, NEU, NEG) and no sentiment (NONE) merging P and P+ in only one category, as well as N and N+ in another one. The same double evaluation using the whole test corpus and a new selected corpus has been carried out, shown in Table 5.

The distributions of successful tweets per groups and per sentiment, for 3-level evaluation, are shown in Table 6 and Table 7.

# of groups	Correct tweets
7	13112 (21.6%)
6	11215 (18.5%)
5	9898 (16.3%)
4	7512 (12.4%)
3	5536 (9.1%)
2	4716 (7.8%)
1	4595 (7.6%)
0	4214 (6.9%)

Table 6: Distribution of successful tweets per groups, for 3-level evaluation

Label	Correct tweets
P	22007 (36.2%)
N	15655 (25.7%)
NONE	18076 (29.7%)
NEU	846 (1.4%)

Table 7: Distribution of successful tweets per sentiment, for 3-level evaluation

5.2 Task 2: Topic Classification

Table 8 shows the results for Task 2. Precision ranges from 67% to 27%. As in Task 1, different submissions from the same group usually have similar values.

Run Id	P	R	F1
ELiRF-UPV-run3	0.67	0.75	0.70
ELiRF-UPV-run2	0.70	0.71	0.70
ELiRF-UPV-run1	0.68	0.69	0.69
LyS-1	0.68	0.60	0.64
LyS-2	0.68	0.59	0.63
IPN-2	0.27	0.33	0.30

Table 8: Results for Task 2

Run Id	**Acc.**	**1k-Run Id**	Acc.
ELiRF-UPV-run3	0.64	ELiRF-UPV-run1-1k	0.48
ELiRF-UPV-run1	0.63	ELiRF-UPV-run3-1k	0.48
ELiRF-UPV-run2	0.63	Elhuyar-Run2-1k	0.47
Elhuyar-Run1	0.61	Elhuyar-Run3-1k	0.47
Elhuyar-Run3	0.61	ELiRF-UPV-run2-1k	0.47
Elhuyar-Run2	0.61	Elhuyar-Run1-1k	0.47
LyS-1	0.58	SINAIword2vec-1-1k	0.46
LyS-2	0.56	LyS-2-1k	0.46
SINAIword2vec-1	0.51	LyS-1-1k	0.45
SINAI-ESMA-1	0.51	JRC-run3-baseline-stop-1k	0.42
SINAI-ESMA-without_negation	0.51	JRC-run1-ER-1k	0.41
JRC-run1-ER	0.48	JRC-run2-RPSN-ER-AWM-4-all-2-skipbigrams-1k	0.40
JRC-run2-RPSN-ER-AWM-4-all-2-skipbigrams	0.48	SINAI-ESMA-without_negation-1k	0.37
JRC-run3-baseline-stop	0.48	SINAI-ESMA-1-1k	0.37
IPN-Linguistic_2	0.37	IPN-Linguistic_2-1k	0.35
IPN-1	0.37	IPN-1-1k	0.33

Table 4: Task 1, classification on 5 levels

Run Id	**Acc.**	**1k-Run Id**	Acc.
ELiRF-UPV-run2	0.71	ELiRF-UPV-run3-1k	0.66
ELiRF-UPV-run1	0.71	ELiRF-UPV-run1-1k	0.65
ELiRF-UPV-run3	0.70	LyS-2-1k	0.64
Elhuyar-Run1	0.70	Elhuyar-Run3-1k	0.64
Elhuyar-Run2	0.70	SINAIword2vec-2-1k	0.63
Elhuyar-Run3	0.70	Elhuyar-Run2-1k	0.63
LyS-1	0.67	LyS-1-1k	0.63
LyS-2	0.67	Elhuyar-Run1-1k	0.62
SINAIword2vec-2	0.61	SINAIword2vec-1-1k	0.61
JRC-run2-RPSN-ER-AWM-4-all-2-skipbigrams	0.61	ELiRF-UPV-run2-1k	0.60
SINAI-ESMA-1	0.61	JRC-run1-ER-1k	0.56
JRC-run1-ER	0.61	JRC-run3-baseline-stop-1k	0.56
SINAI-ESMA-without_negation	0.60	JRC-run2-RPSN-ER-AWM-4-all-2-skipbigrams-1k	0.55
JRC-run3-baseline-stop	0.60	SINAI-ESMA-1-1k	0.52
SINAIword2vec-1	0.59	SINAI-ESMA-without_negation-1k	0.52
IPN-Linguistic_2	0.55	IPN-Linguistic_2-1k	0.52

Table 5: Task 1, classification on 3 levels

5.3 Task 3: Aspect Detection

The aim of task 3 is detecting the aspects from a predefined list, related with the football domain, and using the Social-TV corpus. Results for Task 3 are shown in Table 9.

Run Id	**P**	**R**	**F1**
ELiRF-UPV-run1	0.91	0.91	0.91
LyS-1	0.81	0.90	0.85

Table 9: Results for Task 3

5.4 Task 4: Aspect-Based SA

Last, results for Task 4 are shown in Table 10. Once we have identified a representation of an aspect, the next step consists of detecting its scope of influence, i.e. the fragment of the text, which is talking about the aspect that was referred to and its SA.

The results are similar and robust and do not drop too low compared to those obtained on the while test corpus.

6 Conclusions and Future Work

Each edition of TASS has a positive conclusion because every year the Spanish SA research community improves their systems

Run Id	P	R	F1
ELiRF-UPV-run2	0.58	0.60	0.59
ELiRF-UPV-run1	0.57	0.59	0.58
ELiRF-UPV-run3	0.56	0.58	0.57
LyS-2	0.52	0.58	0.55
LyS-1	0.51	0.57	0.54
LyS-3	0.46	0.51	0.48

Table 10: Results for Task 4

and results. It is very important to note that the general results obtained are comparable to those of the international community. Two aspects are noteworthy, the first one is related to the fact that each edition is increasing the number of unsupervised or at least semi-supervised systems. This is important because they do not need prior knowledge to perform the classification, which is scarce in the dynamic context of social networks. The other issue is related to the fact that the systems submitted try to use the last methods in the state of the art, like classifiers based on deep learning.

The main purpose for future editions is to continue increasing the number of participants and the visibility of the workshop in international forums.

References

Hernández Petlachi, Roberto and Xiaoou Li. 2014. Análisis de sentimiento sobre textos en español basado en aproximaciones semánticas con reglas lingüísticas. In *Proceedings of the TASS workshop at SEPLN*. Gerona, Spain, September.

Hurtado, Lluís and Ferrán Pla. 2014. Elirf-upv en tass 2014: Análisis de sentimientos, detección de tópicos y análisis de sentimientos de aspectos en twitter. In *Proceedings of the TASS workshop at SEPLN*. Gerona, Spain, September.

Jiménez Zafra, Salud M., Eugenio Martínez Cámara, M. Teresa Martín Valdivia, and L. Alfonso Ureña López. 2014. Sinai-esma: An unsupervised approach for sentiment analysis in twitter. In *Proceedings of the TASS workshop at SEPLN*. Gerona, Spain, September.

Montejo Ráez, Arturo, M. Ángel García Cumbreras, and M. Carlos Díaz-Galiano. 2014. Participación de sinai word2vec en tass 2014. In *Proceedings of the TASS workshop at SEPLN*. Gerona, Spain, September.

Pang, Bo and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January.

Perea-Ortega, José M. and Alexandra Balahur. 2014. Experiments on feature replacements for polarity classification of spanish tweets. In *Proceedings of the TASS workshop at SEPLN*. Gerona, Spain, September.

San Vicente Roncal, Iñaki and Xabier Saralegi Urizar. 2014. Looking for features for supervised tweet polarity classification. In *Proceedings of the TASS workshop at SEPLN*. Gerona, Spain, September.

Vilares, David, Yerai Doval, Miguel A. Alonaso, and Carlos Gómez-Rodríguez. 2014. Lys at tass 2014: A prototype for extracting and analysing aspects from spanish tweets. In *Proceedings of the TASS workshop at SEPLN*. Gerona, Spain, September.

Villena-Román, Julio, Sonia Collada-Pérez, Sara Lana-Serrano, and José Carlos González Cristóbal. 2011. Hybrid approach combining machine learning and a rule-based expert system for text categorization. In *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS)*. AAAI Press.

Villena-Román, Julio, Janine García-Morena, Sara Lana-Serrano, and José Carlos González-Cristóbal. 2014. Tass 2013 - a second step in reputation analysis in Spanish. *Procesamiento del Lenguaje Natural*, 52(0):37–44.

Villena-Román, Julio, Sara Lana-Serrano, Eugenio Martínez-Cámara, and José Carlos González-Cristóbal. 2013. Tass - workshop on sentiment analysis at SEPLN. *Procesamiento del Lenguaje Natural*, 50:37–44.

Procesamiento del Lenguaje Natural, Revista nº 54, marzo de 2015, pp 69-76 recibido 24-11-14 revisado 27-01-15 aceptado 10-02-15

Detección automática de chilenismos verbales a partir de reglas morfosintácticas. Resultados preliminares

Automatic detection of verbal chilenismos using morphosyntactic rules. First results

Walter A. Koza, Pedro Alfaro Faccio
Pontificia Universidad Católica de Valparaíso
Av. El Bosque 1290, Viña del Mar, Chile
walter.koza@ucv.cl

Ricardo Martínez Gamboa
U. Diego Portales/U. de La Frontera
Vergara 240, Santiago, Chile
ricardomartinezg@gmail.com

Resumen: En el presente artículo, se describen las tareas realizadas para el desarrollo de un extractor automático de verbos diferenciales del español chileno mediante la aplicación de reglas de lenguaje natural. A partir de este objetivo, se procedió a la modelización de características léxicas, morfológicas y sintácticas de estas expresiones, la cual permitió la posterior implantación en máquina. En primer lugar, se clasificaron los chilenismos verbales en cuatro tipos, según su uso registrado en los diccionarios y su naturaleza sintáctica: puros, puros pronominales, de sentido y de sentido pronominales. En segundo lugar, se establecieron reglas sintácticas para el reconocimiento automático. En el trabajo computacional se utilizaron las herramientas Smorph y Módulo Post Smorph, que trabajan en bloque a base de reglas del lenguaje natural. Este método se probó en un corpus compuesto por 5.194 *tweets*, producidos por usuarios chilenos, logrando 85,54% de precisión, 96.16% de cobertura y 90,53% de medida f. Los resultados permiten validar el método propuesto, aunque se detectaron algunas limitaciones y detecciones erróneas, lo que implicaría la necesidad de especificación de algunas reglas y la creación de nuevas, tanto para la detección automática como para el filtrado de etiquetados erróneos.
Palabras clave: Chilenismo verbal, detección automática, reglas morfosintácticas, Smorph, MPS.

Abstract: In this paper, the tasks made for obtaining an automatic extractor for verbal chilenismos using natural language rules are described. With this objective, a formalization of lexical, morphological and syntactic features was made, for a subsequent computational implementation. Firstly, verbal chilenismos were classified in four kinds, according to the use registered in the dictionaries and syntactic features: pure, pure-clitic, of sense, and of sense- clitic. Secondly, syntactic rules were established for the automatic recognition. Smorph and Post Smorph Module were used in the computational work, both use natural language rules. The method was tested in a corpus composed by 5194 *tweets* produced in Chile, obtaining 85.54% of precision, 96.16% of coverage, and 90.53% of F-measure. The results show that this method is able for this kind of work, all the same, some limitations and mistakes were detected and more specific and new rules are necessary for the recognition task and for filtering wrong tagged.
Keywords: Verbal chilenismo, automatic detection, morphosyntactic rules, Smorph, MPS.

ISSN 1135-5948 © 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

1 Introducción

Los diccionarios diferenciales son obras lexicográficas que buscan codificar una variedad nacional de una lengua a través de un proceso de estandarización. Particularmente en Chile se han elaborado varios de estos diccionarios. Más allá de la relevancia de estas obras, las metodologías que se han solido utilizar podrían considerarse insuficientes al momento de enfrentar grandes bases de datos de lenguaje natural, en la medida en que ocupan principalmente técnicas manuales en las que priman el contraste entre obras, las entrevistas a hablantes y el juicio de expertos. En este sentido, contar con una herramienta automática que permita la recopilación de voces y la construcción de lexicones sería de gran ayuda para los lexicógrafos.

En este marco, se propone una modelización del chilenismo verbal a través de una descripción morfosintáctica, que permita su implantación en máquina con el fin de establecer un método de detección automática basada en lenguaje natural. Mediante la concresión de este objetivo se pretende aportar a las tareas de extracción de información, aplicadas a la lexicografía diferencial, y proponer una modelización de estructuras morfosintácticas.

Así, el trabajo consistió en la elaboración de un diccionario electrónico con los lemas univerbales del DUECh, a los que se les asignó un modelo de acuerdo a sus especificidades morfológicas. Esto permitió contar con una herramienta que generara chilenismos verbales en las diversas formas flexivas y que los pudiera detectar en textos de lenguaje natural.

Desde una perspectiva teórica, se establecieron cuatro tipos de chilenismos verbales: (1) puros (‘achorar’), (2) puros pronominales (‘chacrearse’), (3) de sentido (‘afilar’) y (4) de sentido pronominales (‘hacerla’). Posteriormente, se elaboraron reglas de carácter sintáctico a partir de la combinación entre pronombres clíticos, verbos y otras estructuras, como, por ejemplo, sintagmas preposicionales.

Para el trabajo computacional, se recurrió a los *software* Smorph (Aït-Mokthar, 1998) y Módulo Post Smorph, MPS (Abacci, 1999), que trabajan en bloque. Smorph es un analizador y generador textual que, en una sola etapa, realiza la segmentación, lematización y análisis morfológico. MPS, por su parte, toma como entrada el *output* de Smorph y, a través de reglas de recomposición, descomposición y correspondencia, analiza la cadena de lemas resultante del análisis morfológico.

Este método fue probado en un corpus compuesto por 5.194 *tweets*, producidos en Chile, provistos por AnaliTIC[1].

2 Caracterización del chilenismo verbal

2.1 Acerca del chilenismo

Como se mencionó, este trabajo tiene por objetivo desarrollar un método de detección automática de chilenismos verbales -o verbos diferenciales chilenos-, a partir de reglas de lenguaje natural. Se ha adoptado una definición de chilenismo verbal que se corresponda con sus características lingüísticas y que, a la vez, sea acorde con el trabajo computacional de detección automática. En este sentido, se considera que un chilenismo es toda palabra (o expresión) de uso documentado en el habla de Chile, cuyo lema: (i) no está registrado en el DRAE (‘funar’); (ii) el DRAE lo registra como chilenismo (‘vitrinear’); y (iii) si bien el lema se encuentra en el DRAE, en Chile, se utiliza para expresar un significado distinto al descrito en dicho diccionario (‘pinchar’).

Otra de las características de estas expresiones es que tienden a estar registradas en diccionarios diferenciales. Este tipo de trabajos constituyen obras lexicográficas que buscan codificar los significados de una variedad nacional de una lengua (Chávez, 2010). La obra más importante de este tipo en Chile es el *Diccionario ejemplificado de chilenismos y de otros usos diferenciales del español de Chile*, *DECh*, primer diccionario diferencial chileno de corte científico (Morales, 1984). En esta obra Morales (1984: XXXI) procura incluir entre sus voces:

> *todo uso o acepción que, además de emplearse o de haberse empleado alguna vez en Chile, no perteneciera al empleo común o general, es decir, no dialectal, tal como lo registra la Real Academia en sus diccionarios oficiales.*

Con este propósito, el procedimiento para determinar las voces que no corresponden al español general consistió en contrastar ítems léxicos de uso documentado en Chile que no estuviesen incorporados en el DRAE (Sáez,

[1] www.analitic.cl

2011). Para ello el autor estableció diferentes tipos de diferencialidad, indicados por abreviaciones o signos específicos :

[N] no incorporada en el DRAE
[*] desplazamiento gramatical
[f] cambio fónico
[g] cambio gráfico
[l] cambio lexemático
[c] cambio de la extensión semántica
[C] cambio en la comprensión (metáforas, metonimias)

Gracias a esta caracterización, las diferencialidades fueron abordadas de manera más detallada que la simple no incorporación al DRAE, al tiempo en que se profundizó en la naturaleza de las voces diferenciales de la variedad del español de Chile.

En 2010 la Academia Chilena de la Lengua publica su propio diccionario diferencial de chilenismos, el *Diccionario de uso del español de Chile*, *DUECh*, cuya construcción sigue los principios que se utilizaron para el DECh. En efecto, Matus (2010: 4) ha señalado que en este diccionario

> *para verificar esta diferencialidad dialectal se ha empleado una batería de contrastividad constituida por un conjunto de diccionarios que contienen léxico general, corpus electrónicos, buscadores (como Google) y encuestas aplicadas a informantes. Esta batería ha sido aplicada rigurosamente a cada una de las unidades léxicas y para cada una de sus acepciones.*

Si bien, el DUECh no indica los tipos de diferencialidad y, en términos metodológicos, no existe información del modo en que los ítems léxicos fueron seleccionados (Sáez, 2011), esta obra constituye el referente más actualizado y exhaustivo de los usos dialectales en Chile.

Cabe destacar que es en este contexto en que surge la motivación para el presente trabajo, pues se busca contribuir desde la extracción automática de información a la elaboración de este tipo de lexicones. A tales efectos, se elaboró un diccionario electrónico para la detección automática a partir de la lista de lemas verbales presentes en el DUECh. Posteriormente, se le asignó un modelo específico a cada uno de ellos, de acuerdo con sus características morfosintácticas.

2.2 Morfología del chilenismo verbal

En el DUECh se registran 944 casos de chilenismos verbales. Estos se distribuyen en las terminaciones “ar” (629 casos, 66,6%), “er” (10 casos, 1,1%), “ir” (13 casos, 1,5%) y los restantes se distribuyen en modelos pronominales (289 casos, 30,6%), tales como “arse”, “earla” o “árselas”. La forma lexicogenética verbal más productiva del español de Chile es la terminación “ear” (329 casos, 34,85%), tal como ya documentara Morales, Quiroz y Mayorga (1969). Esta es, además, la menos documentada por el DRAE, dado que se incluyen en los diccionarios de la Academia solo el 24,3% de estos verbos. Otro aspecto relevante consiste en que una de las maneras más productivas del español de Chile para formar verbos nuevos es la que, de acuerdo con Morales y Quiroz (1983), corresponde a un desplazamiento de tipo gramatical, en este caso, la pronominalización de un verbo ya documentado en el español general, como, por ejemplo “agarrar(la)”. En general estos verbos forman un complejo transitivo interno en que se desplaza gramaticalmente la función verbal desde lo intransitivo o lo transitivo (externo) a formas en que se internaliza el Objeto Directo. En estos casos prácticamente la totalidad de las formas no clíticas están presentes en el DRAE.

A partir de esta descripción, se elaboraron modelos morfológicos y se establecieron cuatro categorías a partir de características morfosintácticas:

1. Puros: se trata de verbos cuyo uso se da casi de manera exclusiva en Chile. La mayoría de estos no está registrado en el DRAE (‘marquetear’) o este último los clasifica como chilenismos (‘lolear’).
2. Puros pronominales: además de tener un uso exclusivo en Chile, estos verbos van acompañados de clíticos (‘enyegüecerse’).
3. De sentido: verbos que, si bien sus lemas están incluidos en el DRAE, poseen un uso particular en Chile. Este es el caso de ‘pinchar’, que significa ‘tener a una relación sentimental sin compromiso de exclusividad ni vínculo legal o religioso’.
4. De sentido pronominales: al igual que los de sentido, su lema está registrado en el DRAE, no obstante, para conformar una expresión propia de Chile, se deben combinar con clíticos (‘podérsela’, ‘casarse’).

Esta clasificación obedece a dos motivos. Por un lado, se pretende establecer una clasificación que contemple la naturaleza de los chilenismos verbales y, por otro, busca ser análoga al trabajo computacional.

El método fue probado en un corpus textual compuesto por 5.194 *tweets*, producidos en Chile entre el 22 y el 28 de noviembre de 2013 con el hashtag #Falabella, provistos por AnaliTIC.

En la sección siguiente, se presenta la implantación en máquina realizada a partir de la categorización propuesta.

3 Metodología

A fin de corroborar la descripción morfológica de los verbos diferenciales chilenos, se llevó a cabo una modelización de dicha descripción para, posteriormente, realizar una implantación en máquina y, así, generar la conjugación; con ellos, finalmente, se pretende detectar estas expresiones en textos de lenguaje natural.

Para el trabajo informático, se recurrió a las herramientas, que trabajan en bloque, Smorph (Aït Mokthar, 1998) y Módulo Post Smorph, MPS (Abacci, 1999).

3.1 Smorph

Smorph es un analizador y generador textual que, en una única etapa, realiza segmentación, lematización y análisis morfológico. Se trata de una herramienta declarativa, en la cual la información lingüística está separada de la maquinaria algorítmica, lo que permite que se la pueda adaptar tanto a cualquier lengua como a cualquier variedad lingüística -por ejemplo, en este caso, al español chileno. En este programa se declaran cinco tipos de informaciones: (1) Códigos Ascii, (2) Entradas, (3) Modelos, (4) Terminaciones y (5) Rasgos.

Los códigos Ascii refieren a la notación específica de Smorph, por lo que no ha sido intervenida para este experimento. Describimos las demás informaciones a continuación.

3.1.1 Entradas

Las entradas constituyen el diccionario lingüístico en el que las expresiones (palabras) tienen la posibilidad de aparecer. En este archivo, la información se declara de tres manera posibles:

- A partir de los lemas con la indicación precisa del modelo morfológico que siguen (1)
- Directamente con la indicación de los rasgos morfológicos (2)
- Con la indicación de categoría gramatical y la información considerada pertinente por el usuario (3)

(1) penquear @vch1
(2) lo /clac .
(3) de /prepde .
con /prepcon .

En el caso de ‘penquear’ (‘reprender’) se presenta el lema que se expresa convencionalmente con la forma infinitiva, tal como ocurre en los diccionarios comunes. Es decir, ‘penquear’ es el lema que representa al grupo de verbos ‘penqueo’, ‘penqueas’, ‘penquea’, ‘penqueamos’, ‘penqueáis’, ‘penquean’, ‘penqueé’, ‘penqueaste’, etc. En el caso de (2), no se recurre a ningún modelo, sino que solo se señala el carácter de pronombre clítico acusativo mediante la expresión ‘clac’. En el caso de las preposiciones (3), fue necesario destacar cada una de ellas, por lo que, además de la etiqueta ‘prep’, de preposición, se le adicionó la preposición misma. En esta ocasión, se establecieron modelos morfológicos y morfosintácticos para los verbos chilenos puros, diferenciándose aquellos que eran pronominales. Así, por ejemplo, ‘penquear’ se considera un chilenismo puro cuando remite a ‘reprender’, pero cuando se combina con un clítico reflexivo (‘me penqueé’), el significado alude a embriagarse. Para diferenciar ambos significados, en el archivo entradas, el lema ‘penquear’ aparece dos veces. De este modo, se distingue el uso con pronombre clítico del que no lo requiere.

(4) penquear @vch1
penquear @vchpron1

Adicionalmente, es necesario señalar que se asignaron etiquetas especiales con información morfológica y sintáctica para los verbos de sentido pronominales, según el pronombre que requieren para convertirse en chilenismo y, en caso de corresponder, la preposición pertinente. Por ejemplo, ‘comer’, cuando alude a ‘tener relaciones sexuales’, se combina con un clítico reflexivo, más la preposición ‘a’, más un sintagma nominal (‘Juan se come a la vecina’). Similar

comportamiento tiene 'hacer' ('tener algo como objeto frecuente de acción', ejemplo, 'le hace al canto'). Para estos verbos, se creó la etiqueta 'vchpronrefa' (verbo chileno pronominal preposición 'a').

Para la presente investigación, se utilizó la lista de entradas correspondientes a verbos, nombres, adjetivos, adverbios, preposiciones, siglas y marcadores discursivos desarrollado por el equipo Infosur[2] de la Universidad Nacional de Rosario, Argentina. A este archivo se le adicionaron los verbos incluidos en el DUECh.

De este, se han extraído 960 unidades univerbales que corresponden a 483 verbos definidos como intransitivos y 477 verbos definidos como transitivos.

3.1.2 Modelos

En los modelos, se consigna la estructura morfológica. Los modelos se introducen a través del símbolo @, que indica el lugar en que va la forma básica o raíz a la que se concatenan las terminaciones. En el ejemplo, se muestra un fragmento para el modelo 1 de verbos chilenos regulares de la primera conjugación.

```
@vch1      -2
   +o      vch/pres/ind/1a/sg/c1/r
   +as     vch/pres/ind/2a/sg/c1/r
   +ás     vch/pres/ind/2a/sg/c1/r
   +ai     vch/pres/ind/2a/sg/c1/r/ch
   +a      vch/pres/ind/3a/sg/c1/r
   +amos   vch/pres/ind/1a/pl/c1/r
   +áis    vch/pres/ind/2a/pl/c1/r
   +an     vch/pres/ind/3a/pl/c1/r
   +aba    vch/imp/ind/1a/sg/c1/r
   +abas   vch/imp/ind/2a/sg/c1/r
   +abai   vch/imp/ind/2a/sg/c1/r/ch
   +aba    vch/imp/ind/3a/sg/c1/r
   (…)
```

Esto se lee de la siguiente manera, primero se indica el número de caracteres que se extrae al lema. Eso significa que a un verbo como 'lolear' se le quita 'ar' y se va combinando con las diferentes desinencias correspondientes, con las variaciones de persona, número, tiempo y modo.

A cada uno de ellos se le asignó el modelo correspondiente de acuerdo a sus particularidades de regularidad, por ejemplo, en el caso de 'huevear', al ser un verbo regular se le asignó el modelo de verbos 1. En cambio, a un verbo del tipo 'acolloncar', se le asignaron dos modelos, uno correspondiente a la forma regular 'acollonc-' (vch10) y otro, a la irregular 'acollonqu-' (vch11). Además, cabe destacar que en los modelos se incluyó la variación de segunda persona del singular del español chileno, para expresiones como 'penqueai'.

3.1.3 Terminaciones

Se trata de una serie de caracteres que expresan un rasgo o un conjunto de rasgos. En las terminaciones se incluyen, entre otros aspectos, las desinencias verbales. Vale aclarar que Smorph permite la inclusión de lo que se ha denominado 'terminaciones distinguidas' (Aït-Mokthar y Lázaro, 1995). Estas consisten en los finales de palabras que permiten determinar la categoría gramatical, son similares a la noción de sufijo aunque pueden diferir en algunos casos. Así, por ejemplo, se sabe que toda palabra terminada en –ción es un nombre femenino singular o que la terminación –ó es propia de un verbo en pretérito perfecto simple, de la tercera persona del modo indicativo. Las terminaciones distinguidas permiten detectar aquellas palabras que no estén incluidas en el archivo de entradas, tales como los neologismos. En esta ocasión, se cargó 'ó' como terminación distinguida de la siguiente manera:

```
ó          v/3/sg/perf/ind
```

3.1.4 Rasgos

Para construir los modelos, se recurre a rasgos morfológico-sintácticos y, en esta ocasión, a información léxica presente en los diccionarios diferenciales. Por ejemplo, se tienen: EMS (etiqueta morfosintáctica), que incluye los valores 'n' (nombre), 'adj' (adjetivo), 'v' (verbo), 'vch' (verbo chileno), 'vchpron' (verbo chileno pronominal), 'cl' (clítico), 'prep' (preposición), 'adv' (adverbio).

A partir de estas cuatro infomaciones, Smorph realiza su análisis. La figura 1 muestra un ejemplo de *tweet* que luego será analizado:

> RT @RadarInformador: quieres sapear a tu vecina cuando se saca la ropa? helicoptero con control a 15 lukas falabella

Figura 1: Ejemplo extraído del corpus

[2] www.infosurrevista.com.ar.

A partir del ejemplo de la figura 1, Smorph da como resultado un archivo con la información asignada a cada uno de sus constituyentes. La tabla 1 muestra este contenido de modo esquemático, destacándose en negrita el verbo:

'RT'.	['RT', mi].
'@RadarInformador'.	['@RadarInformador', mi].
':'.	['2p', 'EMS', 'dosp'].
'quieres'.	['querer', 'EMS', 'v','EMS', 'ind', 'PERS', '2a', 'NUM', 'sg', 'TPO', 'pres', 'TR', 'hi', 'TDIAL', 'est'].
'sapear'.	**['sapear', 'EMS', 'vch', 'EMS', 'infin', 'TR', 'r', 'TC', 'c1'].**
'a'.	['a', 'EMS', 'prep'].
'tu'.	['tu', 'EMS','det', 'TDET', 'pos'].
'vecina'.	['vecino', 'EMS', 'adj', 'GEN', 'fem', 'NUM', 'sg']. ['vecino', 'EMS', 'nom', 'GEN','fem', 'NUM','sg'].
'cuando'.	['cuando', 'EMS','rel'].
'se'.	['lo', 'EMS', 'cl', 'TPCRF', 'rflse'].
'saca'.	['sacar', 'EMS', 'v', 'EMS', 'ind', 'PERS', '3a', 'NUM', 'sg', 'TPO', 'pres', 'TR', 'r', 'TC', 'c1', 'TDIAL', 'estrpi'].
'la'.	['el', 'EMS', 'det', 'TDET', 'art']. ['lo', 'EMS', 'cl', 'TPCL', 'nrfl'].
'ropa'.	['ropa', 'EMS', 'nom', 'GEN', 'fem', 'NUM', 'sg'].
'?'.	['nif', 'EMS','pun'].
'helicoptero'.	['helicoptero', mi].
'con'.	['con', 'EMS', 'prep'].
'control'.	['control', 'EMS', 'nom', 'GEN','_', 'NUM', 'sg'].
'a'.	['a', 'EMS','prep'].
'15'.	['num', 'EMS', 'numer'].
'lukas'.	['lukas', mi].
'falabella'.	['falabella', mi].

Tabla 1: Esquema de datos de salida de Smoph. Ejemplo extraído del corpus

3.2 Módulo Post Smorph, MPS

MPS tiene como *input* la salida de Smorph y, a partir de reglas de recomposición, descomposición y correspondencia, declaradas por el usuario, analiza la cadena de lemas resultante del análisis morfológico. Con este programa, se elaboraron reglas sintácticas para la combinación de verbos chilenos con y sin pronombres.

Las fuentes declarativas de MPS están constituidas por un único tipo de archivo, rcm.txt, que incluye un listado de reglas que especifican cadenas posibles de lemas con una sintaxis informatizada. Las reglas pueden ser de tres tipos: (1) recomposición: D + N = SN; (2) descomposición: Contracc = P + D; y (3) correspondencia: Art = D.

En el presente trabajo se recurrió a reglas de reagrupamiento, de las etiquetas 'cl' (clítico: 'me', 'se', etc.), los verbos clasificados como vchr y ciertas preposiciones. Algunas de las combinaciones fueron las siguientes:

Reglas	Ejemplos
cl + cl + vchr = chilenismo	'me la rebusco'
clref + vchr + a + SN = chilenismo	'se come a la vecina'
cldat + vchr + a + SN = chilenismo	'le hace al canto'
vchr (en forma infinitiva) + cl = chilenismo	'enyegüecerse'

Tabla 2: Reglas para MPS y ejemplos de chilenismos

A partir de este tipo de reglas, MPS logró detectar chilenismos del modo en que muestra la figura 2.

'@fdoverdugo'.['@fdoverdugo', mi]. ':'.['2p', 'EMS', 'dosp']. '#Falabella'.['#Falabella', mi]. **'te jode'. ['te joder', 'EMS', 'chil'].** 'con'. ['con', 'EMS', 'prepcon']. 'CtaCte'.

Figura 2: Información de salida de MPS.

Este método fue aplicado a un corpus compuesto por 5.194 *tweets*, producidos en Chile con el *hashtag* #Falabella, entre el 22 y el 28 de noviembre de 2013.

4 Resultados

El corpus contenía 443 chilenismos verbales, de los cuales el método propuesto fue capaz de detectar 426. Asimismo, se detectaron de forma errónea 72 verbos. A partir de allí, se determinó una precisión de 85,54%, una cobertura de 96,16% y una medida f de 90,53%.

En la tabla 3, se muestran los chilenismos clasificados de acuerdo con la propuesta presentada, junto con las cantidades obtenidas.

Chilenismos	Total	Detectados	Omitidos	Errores
Puros	50	48	2	0
Puros pro.	11	11	0	0
De sentido	187	178	9	51
De sentido pro.	195	189	6	21

Tabla 3: Resultados generales

A continuación, en la tabla 4, se presentan algunos ejemplos de chilenismos verbales puros hallados en el corpus:

Type	Ejemplo
Funar	['Ayúdame' , 'EMS', 'v+cl']. 'a'. ['a', 'EMS', 'prep']. 'funar '. ['funar ', 'EMS', 'chil']. 'a'. ['a', 'EMS', 'prep']. '#Falabella'. ['#Falabella', mi].
Tincar	'Lo'. ['lo', 'EMS', 'art']. 'que'. ['que' , mi]. 'mas'. ['mas' , mi]. 'me'. ['me', 'EMS', 'cl']. 'tincó'. ['tincar', 'EMS', 'chil']. 'del'. ['del', mi]. 'cybermonday'. ['cybermonday', mi]. 'de'. ['de' , mi]. 'falabella'. ['falabella' , mi]. 'fue las'. ['ir las', 'EMS', 'chil']. 'space'. ['space' , mi]. 'bag'. ['bag', mi]. 'xdd' ['xdd', mi]
Huevear	'para'. ['parar', 'EMS', 'prep']. 'el'. ['el', mi]. ''. [''' , mi]. 'hueveo'. ['huevear', 'EMS', 'chil']. ''. [''' , mi].
Maraquear	'pacos'. ['pacos', mi]. 'maraqueando'. ['maraquear', 'EMS', 'chil'].
Agringar	'flaytes'. ['flaytes', mi]. 'agringados'. ['agringar', 'EMS', 'chil'].
Pitutear	'oooppss'. ['oooppss' , mi]. 'Pituteando'. ['pitutear', 'EMS', 'chil'].

Tabla 4: Ejemplos de reconocimiento automático

Un aspecto de interés consiste en que se detectaron dos neologismos a partir de las terminaciones distinguidas: ‘loguear’ y ‘clickear’. Estos candidatos a chilenismos verbales pueden ser sometidos a análisis por parte de los lexicógrafos.

5 Conclusiones

A partir de los resultados, puede señalarse que el método propuesto resulta útil y adecuado para la detección de lo que aquí se ha denominado chilenismos verbales puros, clíticos o no. Esto se debe a que se pudo implantar en máquina los lemas recogidos en el diccionario de chilenismos y se logró conjugarlos mediante la modelización de las estructuras morfológicas que presentan.

No obstante, se detectaron algunos problemas derivados de verbos cargados en el archivo de Smorph como chilenos pronominales, pero que presentan un uso no diferencial. Tal es el caso de ‘hacerla’, cuando el pronombre ‘la’ remitía a un Complemento Directo referenciado por este. Este fue uno de los errores de mayor importancia (superior al 50%), debido a la frecuencia de uso no diferencial de verbos como ‘hacer’ o ‘poder’. A fin de poder subsanar este inconveniente, se considera la posibilidad de adicionar métodos estadísticos que permitan reanalizar los datos.

Un segundo problema consiste en la variación grafemática de los verbos. Se observó que verbos como ‘huevear’, aparecen escritos en el corpus de distintas maneras por los usuarios: ‘webear’, ‘weviar’, ‘huear’, etc. Al respecto cabe señalar que los datos del corpus provienen de un modo de comunicación en el que las prácticas discursivas tienden a ser informales, lo que permite que los hablantes utilicen diversas opciones para escribir una misma palabra. Asimismo, se observa este hecho con mayor frecuencia en los verbos en la forma voseante: ‘comís’, ‘comíh’, ‘comí’’, etc. Esto se debe que el poco prestigio del uso de estas formas en la comunicación escrita, no ha permitido que en Chile se estandarice su escritura. A pesar de que se intentó normalizar este hecho a través de la modelización de verbos en diversas posibles formas de escritura, será necesario en trabajos futuros otorgar mayor importancia a este tipo de variaciones y declararlas en el archivo de modelos verbales de Smorph.

Tal como se demostró, es posible incorporar a Smorph un diccionario con información dialectal, en este caso, del habla chilena, específicamente a nivel de morfología verbal. Si bien, mediante este método, no es posible determinar la procedencia del autor (en este caso, el autor del *tweet*), la herramienta permite detectar palabras registradas como propias del español de Chile. Asimismo, al tratarse de una herramienta declarativa, existe la posibilidad de adaptarla para modelizar cualquier variedad

lingüística del español, cargando los diccionarios y los modelos adecuados.

En relación con el objetivo de este trabajo -la creación de un extractor automático de verbos diferenciales del español chileno- se puede señalar que, en esta primera etapa, se logró desarrollar un diccionario electrónico que contiene chilenismos verbales puros, pronominales y no pronominales, a cuyos lemas les fueran asignados modelos morfológicos que permiten detectar las posibles flexiones en un corpus.

En el caso de los chilenismos verbales de sentido, se establecieron modelos que, además de los rasgos morfológicos, contienen información de nivel sintáctico, esto es: (i) características de la flexión verbal y (ii) propiedades sintácticas. En (ii) se consignaron los tipos de palabras con las que debía relacionarse el verbo para convertirse en chilenismo. Como ya se mencionó, un verbo de estas características sería 'comer' que se combina con un clítico, la preposición 'a' y un SN, o bien 'abanderizar' ('simpatizar con una causa') que también va combinado con un clítico, una preposición (en este caso, 'con') y un SN: 'se abanderiza con una causa perdida'.

Cabe señalar que se hace necesario realizar una exploración exhaustiva del carácter sintáctico de los chilenismos verbales de sentido a fin de obtener información de su comportamiento sintáctico y evaluar la posibilidad de implantarlos de modo informatizado bajo reglas de lenguaje natural.

Otro de los desafíos consiste en detectar locuciones y neologismos verbales que puedan incorporarse al conjunto de verbos chilenos. Para ello, se podría apelar, en algunos casos, a las terminaciones distinguidas de Smorph. No obstante, algunas terminaciones pueden generar ambigüedades. Para ello, otra opción sería apelar al contexto sintáctico que rodea a la expresión neológica.

Por último, se requiere afinar las reglas de detección a fin de evitar etiquetados erróneos como los ya señalados en la sección anterior. Para ello, una opción sería extenderse más allá de los pronombres y preposiciones que requiere el verbo.

Se espera, una vez establecidas las reglas de detección automática, por un lado, contar con una herramienta que ayude a las tareas lexicográficas y, por otro, corroborar las hipótesis lingüísticas acerca de la estructura morfológica y sintáctica del chilenismo verbal.

El trabajo a futuro se organizará en torno a los siguientes ejes: (1) ampliar el corpus y combinar las reglas establecidas con estrategias estadísticas; (2) mejorar la precisión en la detección de chilenismos verbales de sentido mediante una afinación de reglas; (3) elaborar reglas de detección de neologismos verbales; (4) Analizar y detectar automáticamente locuciones verbales chilenas.

Bibliografía

Abacci, F. 1999. Développement du Module Post-Smorph. Clermont-Fd.: Memoria del DEA de Linguistique et Informatique. Universidad Blaise-Pascal/GRIL.

Academia Chilena de la Lengua. 2010. *Diccionario de uso del español de Chile*. Santiago, MN Editorial Ltda.

Aït-Mokthar, S. 1998. SMORPH: Guide d'utilisation. Rapport technique. Clermont-Fd.: Universidad Blaise Pascal/GRIL.

Aït-Mokthar, S. y Lázaro, M. 1995. Segmentación y análisis morfológico en español utilizando el sistema Smorph. *Procesamiento del lenguaje natural*, 17, 29-41.

Chávez, S. 2010. Ideas lingüísticas enprólogos de diccionarios diferenciales del español de Chile. Etapa 1875–1928. *Boletín de filología*, XLV(2) 49-69.

Matus, A. 2010. Un diccionario para la lexicografía clásica chilena. En Morales Pettorino, F. 2010. *Nuevo Diccionario Ejemplificado de Chilenismos. Edición refundida y actualizada*. Suplemento. Valparaíso, Edit. Puntángeles (pp. VII-XIII).

Morales Pettorino, F. 1984. *Diccionario ejemplificado de chilenismos*. Valparaíso, Academia Superior de Ciencias Pedagógicas de Valparaíso.

Morales Pettorino, F., Quiroz, O. y Mayorga, D. 1969. Los verbos en -ear en el español de Chile. Santiago, Editorial del pacífico.

Sáez, L. 2011. El léxico del dialecto chileno: Diccionario de uso del español de Chile DUECh. *Estudios filológicos*, 49, 137-15.

Procesamiento del Lenguaje Natural, Revista nº 54, marzo de 2015, pp 77-84 recibido 24-11-14 revisado 27-01-15 aceptado 11-02-15

Polarity analysis of reviews based on the omission of asymmetric sentences

Análisis de la polaridad de comentarios basado en la omisión de oraciones asimétricas

John Roberto, Maria Salamó, M. Antònia Martí
University of Barcelona
Gran Via 585, 08007 Barcelona, Spain
{roberto.john,maria.salamo,amarti}ub.edu

Resumen: En este artículo presentamos una aproximación novedosa para el tratamiento de la polaridad en comentarios sobre productos. Nuestro método se centra en identificar y eliminar las oraciones que tienen una polaridad opuesta a la del comentario (oraciones asimétricas) como paso previo a la identificación de los comentarios positivos y negativos. Nuestra hipótesis de partida es que las oraciones asimétricas son morfo-sintácticamente más complejas que las oraciones simétricas (oraciones con la misma polaridad que la del comentario) por lo que es posible mejorar la detección de la polaridad eliminado este tipo de oraciones del texto. Para validar esta hipótesis, hemos medido la complejidad sintáctica de ambos tipos de oraciones en diferentes dominios y hemos contrastado tres configuraciones de datos diferentes basadas en el uso y la omisión de las oraciones asimétricas.
Palabras clave: Análisis de la polaridad, minería de opiniones, complejidad sintáctica

Abstract: In this paper, we present a novel approach to polarity analysis of product reviews which detects and removes sentences with the opposite polarity to that of the entire document (asymmetric sentences) as a previous step to identify positive and negative reviews. We postulate that asymmetric sentences are morpho-syntactically more complex than symmetric ones (sentences with the same polarity to that of the entire document) and that it is possible to improve the detection of the polarity orientation of reviews by removing asymmetric sentences from the text. To validate this hypothesis, we measured the syntactic complexity of both types of sentences in a multi-domain corpus of product reviews and contrasted three relevant data configurations based on inclusion and omission of asymmetric sentences from the reviews.
Keywords: Polarity analysis, opinion mining, syntactic complexity

1 Introduction

In recent years, there has been a growing interest in mining opinions from user-generated content on the Web. This interest is motivated in part by an increase in freely available online reviews of products and services.

According to Ricci and Wietsma (2006), a product review can be defined as a subjective piece of text describing the user's experiences, product knowledge and opinions, together with a numerical rating. A common characteristic of a posted review is the presence of an overall opinion polarity, which describes the positive or negative opinion of the author with respect to the evaluated item.

A review, like all other opinionated documents, often consists of some evaluative text units and non-evaluative text units that jointly contribute to the overall polarity of the document. These units have either the same or the opposite polarity as that of the entire review. In this regard, in traditional approaches to polarity analysis, the overall polarity of a text is the average polarity of all its units, mostly words (e.g. adjectives), phrases, and sentences. In contrast to those studies that consider the overall polarity as the result of the average polarity of sentences, in this work we retrieve sentences expressing similar and opposite polarity orientation in relation to the entire review and analyze the differences between them. This paper starts from the premise that both types of sentences use different language constructs becau-

ISSN 1135-5948 © 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

se sentences with the same semantic orientation of the review have less complex structures than sentences with the opposite polarity orientation. On the basis of this assumption, we made the following hypothesis:

Hypothesis 1

> *We hypothesize that sentences with the same polarity to that of the entire review are syntactically different from those with the opposite polarity.*

In this paper we call "symmetric sentences" those sentences that have the same polarity as that of the entire review, and "asymmetric sentences" those that do not. Based on this, a second hypothesis can be stated:

Hypothesis 2

> *We hypothesize that it is possible to improve the detection of the polarity orientation of reviews by removing from the text the asymmetric sentences.*

With the aim of verifying *hypotheses 1* and *2*, we conducted two experiments with a multi-domain corpus of product reviews in English. These experiments are designed to demonstrate: a) that it is possible to predict accurately when a particular sentence is symmetric or asymmetric, and b) that it is possible to improve the automatic detection of the polarity orientation of reviews by removing asymmetric sentences. The results from both experiments are promising. In particular we show that removing asymmetric sentences improves the performance of the *baseline* to determine the overall polarity of positive and negative reviews.

The rest of this paper is organized as follows: Section 2 looks at the related work on polarity analysis of customer reviews. Next, Section 3 describes syntactic complexity measures. Section 4 explains the experimental analysis (data, tools and results). Finally, we present conclusions in Section 5.

2 Related Work

The traditional state-of-the-art approaches classify polarity of natural language text by analyzing vector representations using, e.g., machine learning (ML) techniques (Pang, Lee, and Vaithyanathan, 2002). ML solutions involve building classifiers from a collection of annotated texts, where each text includes some linguistic-related processing for preparing features such as lemmatization or stemming. Alternative approaches are semantic / lexicon-based (Turney, 2002; Taboada et al., 2011), which renders them robust across domains and texts and enables linguistic analysis at a deeper level. Semantic-based methods involve the use of dictionaries where different kinds of words are tagged with their semantic orientation (SO).

The great majority of works in polarity analysis have mainly focused on analysis of sentences expressing a direct or comparative opinion[1] (Dastjerdi, Ibrahim, and Ghorashi, 2012; Ganapathibhotla and Liu, 2008; Jindal and Liu, 2006). There are few studies analysing how other type of sentences affect the polarity of the entire reviews (cp. Roberto, Salamó, and Martí (2014); Wu and He (2011); Ramanand, Bhavsar, and Pedanekar (2010); Goldberg et al. (2009); and Kim and Hovy (2006)). More specifically, Kim and Hovy (2006) presented a system that automatically extracts the *pros* and *cons* sentences from online reviews. They focused on extracting *pros* and *cons* which include not only sentences that contain opinion-bearing expressions about products and features but also sentences with reasons why an author of a review writes the review.

Goldberg et al. (2009) conducted a novel study on building general "wish detectors"[2] for natural language text, and demonstrated their effectiveness on domains as diverse as consumer product reviews and online political discussions. In the same vein, Ramanand, Bhavsar, and Pedanekar (2010) described rules that can help detect "wishes" from texts such as reviews or customer surveys. Wu and He (2011) analyzed the problem of automatically identifying wishes in product reviews. They built an approach towards such detections, by the use of keyword set constructed by modal words and sequential patterns. Finally, Roberto, Salamó, and Martí (2014) analyzed the role played by narrative sentences in determining the polarity of reviews. Specifically, they applied an algorithm to de-

[1] According to Liu (2010), "direct opinions" give a positive or negative opinion about a object without mentioning any other similar objects and "comparative opinions" declare a preference relation of two or more objects based on some of their shared features.

[2] Wishes are sentences in which authors make suggestions about a product or service or show intentions to purchase a product or service.

tect sentences containing events semantically connected (narrative chains).

3 Syntactic Complexity

As we stated in Section 1, we hypothesize that symmetric sentences are syntactically different from asymmetric ones. In this section, we define syntactic complexity and we present a number of different measures of syntactic complexity.

Syntactic complexity refers to "the range of forms that surface in language production and the degree of sophistication of such forms" (Ortega, 2003). Even though there is no single agreed-upon measurement of syntactic complexity, it is mostly a matter of sentence embedding[3] (compare sentences a. and b. in example 1) and non-canonical word order (compare sentences a. and b. in example 2).

(1) a. I eat and you cook.

b. I eat *if you cook.*

(2) a. The student that met the teacher (subject relative clause).

b. The student *that the teacher met* (direct object relative clause).

Some measures of syntactic complexity are common in first and second language acquisition and development (e.g. Index of Productive Syntax or Developmental Sentence Scoring (Moyle and Long, 2013)). However, the act of giving an opinion is a cognitive activity that does not concern with the language acquisition or development. For this reason, we selected three measures that are not directly linked to language acquisition processes but quantify the demand of cognitive processing of different types of syntactic constructions: Yngve's depth algorithm (Yngve, 1960), Frazier's local nonterminal count (Frazier, 1985), and Pakhomov's length of grammatical dependencies (Pakhomov et al., 2011).

Yngve (1960) assumes that the production of a sentence imposes demands on a limited-capacity working memory. The depth of any word in a sentence represents the number of planned grammatical constituents that have not yet been realized during the production of the sentence. Yngve depth is determined by numbering the branches below each node from right to left in a syntactic tree, starting with zero. The depth of each word was the sum of all the branches connecting the word to the root or top-most node of the sentence. Figure 1 illustrates the calculation of the Yngve depth measure in the sentence "it was still starting but a bit sluggish". In this figure, *Total_Ydepth* is the sum of the depth of each word in the sentence and *mean_Ydepth* is the total divided by the number of words.

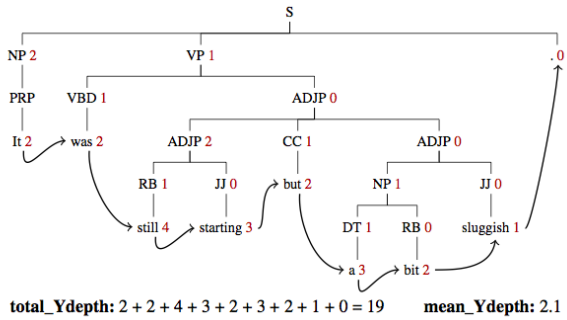


Figura 1: Parse tree fragments with scores for Yngve depth analysis.

Frazier's complexity metric (1985) is based on the idea that syntactic complexity involves the number of non-terminal nodes that the parser must construct when it processes a sentence. The Frazier's approach proceeds in a bottom-up fashion. It traces a path from a word up the tree until reaching either the root of the tree or the lowest node which is not the leftmost child of its parent. Each non-terminal node in the path contributes a score of 1, with 1.5 points for branches from a sentence node (S). Figure 2 illustrates the calculation of the Frazier local non-terminal count measure for the same example sentence "it was still starting but a bit sluggish". *Total_Fdepth* is the sum of the scores for each word in the sentence and *mean_Fdepth* is the total divided by the number of words.

The Pakhomov's scoring method (2011) is inspired in Gibson (1998). It computes the length of grammatical dependencies between lexical items in a sentence based on the Stanford syntactic parser. In the Pakhomov's approach, each dependency relation receives a distance score calculated as the absolute difference between the serial positions of the words that participate in the relation. For example, the distance for the nominal subject relation (*nsubj*) is 4-1 = 3. *Total_SynDepLen* is the sum of all dependencies in the sentence

[3]Embedding refers to the combining of simple sentences into a more complex sentence.

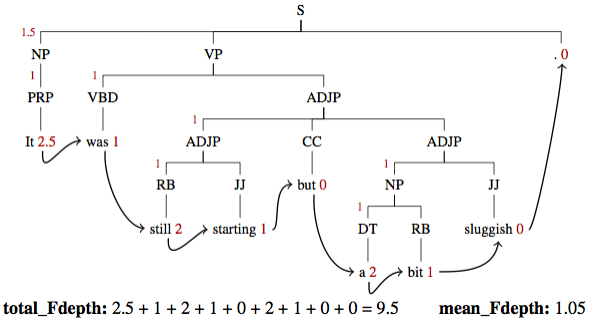
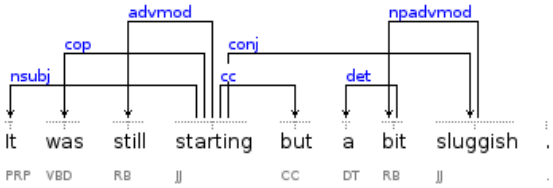


Figura 2: Parse tree fragments with scores for Frazier's node count.

and *mean_SynDepLen* is the total divided by the number of dependencies.



nsubj(starting-4, It-1) =3
cop(starting-4, was-2) =2
advmod(starting-4, still-3) =1
cc(starting-4, but-5) =1
det(bit-7, a-6) =1
npadvmod(sluggish-8, bit-7) =1
conj(starting-4, sluggish-8) =4

total_SynDepLen: 3 + 2 + 1 + 1 + 1 + 1 + 4 = 13
mean_SynDepLen: 1.8

Figura 3: A graph view of typed dependencies of a sentence as computed by Stanford parser and the calculation of the dependency length for Pakhomov scoring method.

In the experiments we apply the Yngve, Frazier, and Pakhomov's indices of complexity to characterize symmetric and asymmetric sentences.

4 *Experiments and results*

This section evaluates the hypotheses stated in Section 1. First, it describes the data and tools used in the experiments. Second, it presents the results obtained in the two experiments addressed to evaluate the hypotheses.

4.1 Data and tools

The data used in our experimental analysis is the multi-domain corpus of product reviews by Cruz Mata (2012). Originally, this corpus is a collection of 2,547 reviews extracted from www.ciao.com. This corpus has been chosen in order to analyze symmetric and asymmetric sentences because each review is annotated with the overall polarity and the polarity of its features in every single sentence. A pre-processing of the corpus has been performed automatically in order to remove sentences that do not express any polarity or induce inconsistencies (noise) in the data set. In the latter case, first, we eliminate all text passages that contain multiple sentences with a unique polarity assigned to it. Second, we also subtract those sentence with mixed polarity (e.g. "It is a nice hotel, small but very nice and clean."). Finally, we removed one-word sentences because they are not relevant for our analysis (e.g. "ok", "avoid", "duuuhhh"). Table 1 describes the corpus used in our experiments after removing those sentences.

Sentence	**Cars**	**Headphones**	**Hotels**	**All**
# *Symm*	403	194	334	994
# *Asymm*	403	194	334	994
Total	**806**	**388**	**668**	**1988**

Tabla 1: Number of symmetric (*Symm*) and asymmetric (*Aymm*) sentences in each domain.

As we mentioned in Section 1, sentences with the same polarity as that of the entire review have been referred to as symmetric (*symm*) and those with different polarity as asymmetric (*asymm*).

We used the Computerized Linguistic Analysis System (CLAS) (Pakhomov et al., 2011) for the computation of syntactic complexity measures. CLAS system implements Yngve (1960), Frasier (1985), Gibson (1998), and other computational approaches to establish the syntactic complexity of English sentences. This software uses the Stanford syntactic parser, which provides basic information on the hierarchical constituent structure of the sentence as well as syntactic dependencies between lexical items.

We used the Semantic Orientation CALculator System (SO-CAL) (Taboada et al., 2011) for calculating the polarity orientation of reviews. SO-CAL is a general purpose system that was designed for determining semantic orientation on the level of complete texts. SO-CAL uses manually built dictionaries of words (adjectives, nouns, verbs, and adverbs) annotated with their polarity and strength, and incorporates negation and intensification (e.g., *very*, *slightly*).

4.2 Symmetric and asymmetric sentences classification

This experiment attempts to answer the following question: Is it possible to predict accurately when a particular sentence will be symmetric or asymmetric?

To answer this question, we analyzed the syntactic complexity of opinionated sentences from reviews using the Computerized Linguistic Analysis System (Pakhomov et al., 2011). The fifteen scores obtained with this tool (listed in Table 2) were used as attributes to train and test different classifiers in Weka (Witten et al., 1999).

N.	Attributes
1	Mean of Frazer depth scores on individual tokens (*mean_Fdepth*).
2	Sum of Frazer depth scores on individual tokens (*total_Fdepth*).
3	Mean of Yngve depth scores on individual tokens (*mean_Ydepth*).
4	Sum of Yngve depth scores on individual tokens (*total_Ydepth*).
5	Mean of syntactic dependency lengths in the dependency parse (*mean_SynDepLen*).
6	Sum of syntactic dependency lengths in the dependency parse (*total_SynDepLen*).
7	Number of "S" nodes in the parse tree.
8	Raw count of nouns.
9	Raw count of adjectives.
10	Raw count of adverbs.
11	Raw count of verbs.
12	Raw count of determiners.
13	Raw count of conjunctions.
14	Raw count of prepositions.
15	Raw count of proper nouns.

Tabla 2: List of the fifteen scores/attributes generated by the Computerized Linguistic Analysis System (CLAS).

With the aim of determining the consistency of the scores obtained automatically by CLAS, we randomly selected 30 sentences from the corpus (10 for each domain) that were labeled and scored by a trained linguist. To compare automatic and human scores, we have used a Kappa statistic approach. The average Kappa score was 0.76, showing an acceptable degree of agreement for the task.

Additionally, we performed a linear transformation on the original data to scale the value of all features in the range [0..1] using the R package "ppls"(Krämer and Sugiyama, 2011). For classification, a 10-fold cross validation methodology was performed from which we report *accuracies*.

Table 3 shows the accuracies obtained for each one of the classifiers analyzed. First column contains the list of classification algorithms that have been tested. Subsequent columns list the distribution of accuracies per domains (cars, headphones, hotels) and all domains as a whole. Finally, the last row contains the average accuracies obtained for each domain.

Algorithm	Cars	Headphones	Hotels	All
BayesNet	70.3	70.1	67.2	73.1
LWL	65.0	71.9	66.5	76.3
DTNB	**72.3**	70.6	69.5	75.8
Decis.Table	71.5	72.7	71.0	76.0
JRip	70.3	70.6	68.7	76.3
Ridor	71.7	70.4	69.9	76.4
ADTree	70.6	**73.7**	71.1	76.0
BFTree	71.7	71.1	70.4	76.4
LADTree	71.3	72.9	71.7	75.5
REPTree	70.7	71.6	69.5	75.5
SimpleCart	70.7	71.6	**72.2**	76.3
Average	*70.6*	*71.6*	*69.8*	*75.8*

Tabla 3: Percentage of symmetric and asymmetric sentences correctly classified in each domain.

The findings of this study reveal that a good accuracy can be obtained using syntactic complexity for determining symmetric and asymmetric sentences. Note that on average all the results are around 70 %. In particular, a 70.6 % in the cars domain, a 71.6 % in the headphones domain, and 69.8 % in the hotels domain. In the cars domain, the best classifier achieves an accuracy of 72.3 % for distinguishing symmetric from asymmetric sentences. The best accuracy estimated using the same syntactic complexity measures is 73.7 % and 72.2 % for headphone and hotel domains, respectively. The best results are achieved bringing all domains: all classification accuracies are above 73 % and the general average is 75.8 %.

Additionally, we apply four well known selection methods to pick up the five most informative attributes that are used to classify symmetric and asymmetric sentences, as shown in Table 4. In general, we have found that the attributes based on the use of syntactic complexity measures (*total_SynDepLen*, *total_Ydepth*, *total_Fdepth*, *mean_Fdepth*, *mean_Ydepth*, and *mean_SynDepLen*) are among the five most discriminative attributes.

CARS			
Chi-squared	**Gain Ratio**	**Info. Gain**	**ReliefF**
total_SynDepLen	total_Ydepth	total_SynDepLen	total_SynDepLen
total_Ydepth	total_SynDepLen	total_Ydepth	total_Ydepth
total_Fdepth	det_count	total_Fdepth	total_Fdepth
mean_Ydepth	total_Fdepth	mean_Ydepth	mean_Ydepth
mean_SynDepLen	mean_Ydepth	mean_SynDepLen	mean_SynDepLen
HEADPHONES			
Chi-squared	**Gain Ratio**	**Info. Gain**	**ReliefF**
total_Ydepth	total_Ydepth	total_Ydepth	total_Fdepth
total_SynDepLen	total_SynDepLen	total_SynDepLen	total_Ydepth
total_Fdepth	mean_Ydepth	total_Fdepth	total_SynDepLen
mean_Ydepth	total_Fdepth	mean_Ydepth	adj_count
num_clauses	num_clauses	num_clauses	conj_count
HOTELS			
Chi-squared	**Gain Ratio**	**Info. Gain**	**ReliefF**
total_Ydepth	total_SynDepLen	total_Ydepth	total_Fdepth
total_SynDepLen	total_Ydepth	total_SynDepLen	num_clauses
noun_count	mean_Ydepth	mean_Ydepth	verb_count
mean_Ydepth	noun_count	noun_count	mean_Ydepth
total_Fdepth	total_Fdepth	total_Fdepth	total_Ydepth

Tabla 4: The most relevant features retained by the attribute selection methods for each domain.

In summary, the most discriminative features are the ones based on the syntactic complexity measures and the accuracies obtained using these features support the hypothesis that it is possible to predict accurately when a particular sentence express the same (symmetric) or the opposite (asymmetric) polarity to that of the entire review.

4.3 Polarity classification

This experiment attempts to answer the following question: Is it possible to improve the accuracy of polarity classifiers by removing asymmetric sentences from reviews?

To answer this question, we calculated the overall polarity of reviews using the Semantic Orientation CALculator System (Taboada et al., 2011). We contrasted three relevant data configurations based on the extraction of different types of sentences from the reviews. These configurations are:

1. **Gold standard:** the polarity analysis was performed using only symmetric sentences based on a hypothetical prediction accuracy of 100 % for the detection of this type of sentences. The input to the SO-CAL system is formed by all the sentences labeled with the same polarity to that of the entire review.
2. **Baseline:** the polarity analysis was performed in standard fashion, that is, by using the entire review. The input to the SO-CAL system is formed by all the sentences from the review.
3. **Approach:** the polarity analysis was performed removing some of the asymmetric sentences from reviews based on the factual categorization accuracies obtained in experiment one (see Table 3). The input to the SO-CAL system is formed by all the sentences from the review except the 70 % of asymmetric sentences for cars, the 71 % of asymmetric sentences for headphones, and the 69 % of asymmetric sentences for hotels.

The results of this experiment are summarized in Table 5. The performance of the *baseline* is consistent with other published studies (Taboada, 2011). The so-called *gold standard* configuration improves from 88 % to 93.2 % the performance of the *baseline* for positive reviews and from 76.8 % to 84.5 % the performance of the negative reviews. Recall that the *gold standard* configuration is based on hypothetical accuracies for symmetric and asymmetric sentences categorization.

Approach is the second best configuration but, in contrast to the *gold standard*, it is based on the factual data gathered from the first experiment. Under this configuration, the positive polarity obtains an average accuracy of

Configurations	**Gold**		**Baseline**		**Approach**	
Review Polarity	+	–	+	–	+	–
Cars	93.4	85	88.1	77.3	**89.4**	**83.2**
Headphones	89.2	80.6	81	74.3	**83.3**	**78.7**
Hotels	97	88	95	78.8	**96.7**	**81.9**
Averages	93.2	84.5	88	76.8	**89.8**	**81.2**

Tabla 5: Polarity analysis of product reviews (based on SO-CAL system) under four different configurations. The reported values are classification accuracies, that is, the percentage of correct choices.

89.8 %. This is a worthy improvement over the 88 % that results when all sentences are used (*baseline*). Nevertheless, the most significant increment is observed in the case of negative reviews: *approach* configuration improves from 76.8 % to 81.2 % the average accuracy of the *baseline* for negative reviews. This huge improvement is shared by every domain.

In summary, these results show that the polarity analysis of reviews improves by removing their asymmetric sentences, as shown in the *Averages* at the bottom of Table 5.

5 Conclusions

In this paper we analyze the function of symmetric and asymmetric sentences (opinionated sentences expressing similar and opposite polarity orientation in relation to the entire document) with the aim of improving the polarity detection of reviews. For this purpose we have performed two tasks.

The first task consists of the evaluation of the usefulness of different syntactic complexity measures to characterize both symmetric and asymmetric sentences. To this end, we have trained a cascade of classifiers using the Weka Environment. Our experiments show that syntactic complexity is an effective way to characterize symmetric and asymmetric sentences and it is possible to detect accurately when a particular sentence is symmetric or asymmetric.

The second task consists of the classification of the reviews as being positive or negative. To this end, we contrasted three relevant data configurations based on the removal of different types of sentences from the reviews. The experimental results indicate that removing asymmetric sentences increases the performance on the determination of the overall polarity of reviews. There is a noticeable improvement in the case of the negative reviews.

6 Acknowledgments

This work was supported by projects SGR-2014-623, TIN2012-38603-C02-02 and TIN2012- 38584-C06-01, as well as a FI grant (2010FI_B 00521).

References

Cruz Mata, Fermín L. 2012. *Extracción de opiniones sobre características: un enfoque práctico adaptable al dominio.* Colección de monografías de la Sociedad Española para el Procesamiento del Lenguaje Natural. Sociedad Española para el Procesamiento del Lenguaje Natural.

Dastjerdi, Niloufar Salehi, Roliana Ibrahim, and Seyed Hamid Ghorashi. 2012. Product feature extraction using natural language processing techniques. *Journal of Computing*, 4(7):39–43.

Frazier, Lyn, 1985. *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, chapter Syntactic complexity, pages 129–189. Cambridge University Press, Cambridge, UK.

Ganapathibhotla, Murthy and Bing Liu. 2008. Mining opinions in comparative sentences. In *Proc. of the 22Nd International Conference on Computational Linguistics*, volume 1 of *COLING '08*, pages 241–248, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Goldberg, Andrew B., Nathanael Fillmore, David Andrzejewski, Zhiting Xu, Bryan Gibson, and Xiaojin Zhu. 2009. May all your wishes come true: a study of wishes and how to recognize them. In *Proc. of Human Language Technologies: The 2009*

Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL '09, pages 263–271, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jindal, Nitin and Bing Liu. 2006. Mining comparative sentences and relations. In *Proc. of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, pages 1331–1336. AAAI Press.

Kim, Soo-Min and Eduard Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proc. of the Workshop on Sentiment and Subjectivity in Text*, SST '06, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Krämer, Nicole and Masashi Sugiyama. 2011. The degrees of freedom of partial least squares regression. *Journal of the American Statistical Association*, 106(494):697–705.

Liu, Bing, 2010. *Handbook of Natural Language Processing*, chapter Sentiment Analysis and Subjectivity, pages 627–666. CRC Press, Connecticut, USA.

Moyle, Maura and Steven Long, 2013. *Encyclopedia of Autism Spectrum Disorders*, chapter Index of Productive Syntax (IPSyn), pages 1566–1568. Springer.

Ortega, Lourdes. 2003. Syntactic complexity measures and their relationship to l2 proficiency: A research synthesis of college-level l2 writing. *Applied Linguistics*, 4(24):492–518.

Pakhomov, Serguei, Dustin Chacon, Mark Wicklund, and Jeanette Gundel. 2011. Computerized assessment of syntactic complexity in alzheimer's disease: a case study of iris murdoch's writing. *Behavior Research Methods*, 43(1):136–144.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP '02: Proc. of the ACL-02 conference on Empirical methods in natural language processing*, pages 79–86, Morristown, NJ, USA. Association for Computational Linguistics.

Ramanand, J., Krishna Bhavsar, and Niranjan Pedanekar. 2010. Wishful thinking: finding suggestions and 'buy' wishes from product reviews. In *Proc. of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, pages 54–61, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ricci and Wietsma. 2006. Product reviews in travel decision making. In *Information and Communication Technologies in Tourism 2006*, pages 296–307.

Roberto, John, Maria Salamó, and M. Antònia Martí. 2014. The function of narrative chains in the polarity classification of reviews. *Procesamiento del Lenguaje Natural*, 52:69–76.

Taboada, Maite. 2011. Stages in an online review genre. *Text and Talk. An Interdisciplinary Journal of Language, Discourse & Communication Studies*, 31(2):247–269.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307, June.

Turney, Peter. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 417–424, Philadelphia, Pennsylvania.

Witten, Ian, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Cunningham. 1999. *Weka: Practical Machine Learning Tools and Techniques with Java Implementations. (Working paper 99/11).* Hamilton, New Zealand: University of Waikato, Department of Computer Science.

Wu, Xing and Zhongshi He. 2011. Identifying wish sentence in product reviews. *Journal of Computational Information Systems*, 7(5):1607–1613.

Yngve, Victor. 1960. A model and an hypothesis for language structure. *Proc. of the American Philosophical Society*, 104(5):444–466.

Procesamiento del Lenguaje Natural, Revista nº 54, marzo de 2015, pp 85-92 recibido 25-11-14 revisado 27-01-15 aceptado 13-02-15

Exploiting Geolocation, User and Temporal Information for Monitoring Natural Hazards on Twitter

Uso de Información de Geolocalización, Usuario y Temporal para la Monitorización de Desastres Naturales en Twitter

Víctor Fresno
UNED
E.T.S.I. Informática
vfresno@lsi.uned.es

Arkaitz Zubiaga
Uni. of Warwick
Coventry, UK
arkaitz@zubiaga.org

Heng Ji
Rensselaer Pol. Institute
Troy, NY, USA
jih@rpi.edu

Raquel Martínez
UNED
E.T.S.I. Informática
raquel@lsi.uned.es

Resumen: Cuando se producen eventos relacionados con situaciones de emergencia, es importante acceder a tanta información como sea posible relacionada con dicho evento. En este contexto algunas redes sociales como Twitter suponen un importante recurso de información en tiempo real. La técnicas clásicas de filtrado de información suelen centrarse en el análisis de coocurrencia de términos con el conjunto de palabras clave inicialmente consideradas. Sin embargo, estas aproximaciones pueden perder información, ya que no son capaces de recuperar información relevante que venga expresada con palabras que no coocurran con las palabras clave inicialmente usadas, y que expresan nuestra necesidad de información. Considerar información de geolocalización, usuario o temporal dentro de un enfoque de pseudo-relevance feedback, nos permite encontrar terminología relacionada con el evento, pero no coocurrente con las palabras clave inicialmente consideradas. Por otro lado, considerando el aspecto temporal se puede modificar una función de expansión de consultas como la divergencia de Kullback-Leibler con el fin de mejorar el filtrado de información en estas situaciones de emergencia. Nuestras propuestas se han evaluado en dos colecciones de eventos del mundo real obteniéndose resultados alentadores.
Palabras clave: Recuperación de Información, Realimentación por relevancia, Análisis de Redes Sociales en Tiempo Real, Twitter, Seguimiento de Desastres Naturales

Abstract: During emergency situation events it is important to acquire as much information about the event as possible, and social media sites like Twitter offer important real-time user contributed data. Typical Information Filtering techniques are keyword-based approaches or focused on co-occurrence with keywords. However, these approaches can miss relevant local information if messages do not contain an initially considered event-related keyword. Considering geolocation, user and temporal information within a pseudo-relevance feedback approach we can find event-related terminology but not co-occurring with initially considered keywords. Thus, taking into account the temporal aspect we can modify a query expansion function like Kullback-Leibler divergence in order to improve the Information Filtering process. Our proposed approaches have been evaluated in two Twitter datasets associated with real-world events, obtaining encouraging results.
Keywords: Information Retrieval, Pseudo-Relevance Feedback, Real-Time Social Media Analysis, Twitter, Natural Hazards Monitoring

1 Introduction

Recent years have seen the explosive growth of the social volume of information. Social media sites like Twitter aggregate a large volume of real-time user contributed data for a wide variety of events (Zubiaga et al., 2011). These events range from popular and widely known pre-scheduled events, to unexpected natural hazards, e.g., earthquakes, hurricanes, etc. In the case of emergency situations, it is important to retrieve as much information as possible about the event to make sure that no relevant information is missed. This can be helpful for humanitarian aid workers to assist citizens effectively, for the relief ac-

ISSN 1135-5948
© 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

tivity management during such events, and for the people to stay abreast of the latest details. Therefore, the fact that social media plays an important role in monitoring the information shared by users in these situations, and the importance of avoiding to miss out relevant information, emphasize the need for developing effective Information Filtering techniques to carry out this monitoring process in the best possible way.

The most common and widely used approach to deal with this kind of Information Filtering tasks is the use of keyword-filtering techniques, which tend to obtain high precision but low recall values, i.e., tweets containing specific keywords that have been manually crafted by a user will most probably be relevant, but it is very likely that other relevant tweets will not contain these keywords. To facilitate tracking event-related contents, users on Twitter usually come to an agreement during emergency situations to use a common hashtag, which can help others follow relevant content related to these events; still, some local information will often be missed. Additional approaches that can help improve this computationally include Latent Semantic Indexing (Deerwester et al., 1990) or Topic Modeling (Steyvers and Griffiths, 2007), used as content-based filtering techniques that can help improve recall values; however, in these cases, the discovery of new relevant keywords is restricted to keywords that co-occur in the tweets that contain the keywords in the user-defined query. Other tweets which contain neither the initial keywords nor the co-occurring ones will never be retrieved using these techniques.

To the end of enhancing this process of monitoring natural hazards on Twitter, we delve into the use of three additional features of tweets, namely "user", "geolocation" and "temporal" information, which we rely on to discover new keywords which are related to the natural hazard. The contribution of our paper revolves around this idea, for which we set forth the following hypotheses:

- If a user posted a tweet message in Twitter about a natural hazard (e.g., a hurricane) in an affected area, we can expect that their immediately previous or later messages will be related with the event, irrespective of these tweets containing or not any of the initially considered hurricane-related keywords.
- If we find a tweet about a natural hazard in a specific geolocation and time, we can expect that tweets within a nearby geolocation and posted at the same time will be also related to the event, the nearest messages around this geolocation and at that time will be also related with it, irrespective of these tweets containing or not any of the initially considered hurricane-related keywords.

We introduce a new preliminary approach for harnessing social information in order to acquire as much information about a natural hazard as possible, and beyond initially considered event-related keywords. We define the Information Filtering problem as a pseudo-relevance feedback task, and propose a query expansion method using the geolocation, user and temporal information inherent to tweets. We incorporate new messages, where the initially event-related keywords are not necessarily used, to the initial set of most relevant documents. Thus, we introduce a modified Kullback-Leibler divergence as a query expansion function that considers the temporal aspect of tweets.

These approaches have been evaluated using two Twitter datasets associated with real-world events: "Hurricane Isaac" in late-August 2012 near Baton Rouge (Louisiana, US); and "Hurricane Sandy" when it affected New York City on October 29th, 2012. The results presented in this paper prove the effectiveness of the proposed approach, motivating further study of the exploitation of this kind of social information.

2 Related Work

Since its creation Twitter has become an important source of information for coverage of crisis events (Imran et al., 2014). For example, the events unfolding during the Sichuan earthquake were first reported by Twitter users. Similarly, the first report that a plane landed in the Hudson river in New York in 2009 was posted on Twitter by an eyewitness. In Mills et al. (2009), an early study on emergency events, the authors found that Twitter had a great impact in distributing crisis-related information. Twitter was crucial during events such as the Californian fires, New England Ice Storm, Gulf of Mexico Hurricane, Cyclone Nagris in Myanmar

and Mississippi Hurricane (Sinnappan, Farrell, and Stewart, 2010). In De Longueville, Smith, and Luraschi (2009) an analysis of tweets related to a fire near Marseille was carried out, and it was shown that Twitter updates about the event were generally well synchronized to temporal and spatial dynamics of the event itself. People access Twitter during crisis events to complement information they obtain from traditional sources ((Sorensen and Sorensen, 2007); (Shklovski, Palen, and Sutton, 2008); (Sutton, Palen, and Shklovki, 2008)) and it is increasingly being considered as a primary source to learn what is happening on the ground (Palen et al., 2010).

Hashtag use is the main mechanism for accessing information in order to design solutions in emergencies (Starbird et al., 2010), and Twitter activity and the nature of its use during emergencies have been subject of both mass media attention and academic research ((Sorensen and Sorensen, 2007); (Starbird et al., 2010); (Hughes and Palen, 2009); (Vieweg et al., 2010)). In order to sense and analyze disaster information from social media, microblogs as a source of social data have recently attracted attention; combining messages is helpful for understanding the impact of an event. From the point of view of event management, and considering geolocation information, Twitter as a social sensor has attracted much attention ((Sakaki, Okazaki, and Matsuo, 2010), (Vieweg et al., 2010), (Sinnappan, Farrell, and Stewart, 2010)). All these approaches consider geolocation information. In Sasaki et al. (2012) the authors consider each Twitter user as a sensor, and tackle an event detection task based on sensory observations. They use Kalman filtering and particle filtering to estimate the locations of an event and they developed an earthquake reporting system that shows the tweets relating to the earthquake on a map. Their system is based on the difference between the number of tweets posted while the event is occurring, and while not. Several map-based systems have been developed on the web to share local knowledge. In these systems, users can enter local safety/hazard incident information on a map (Shinohara et al., 2011). Recently, in Schnebele and Cervone (2013) a methodology for the generation of flood hazard maps is presented fusing remote sensing and volunteered geographical data.

To the best of our knowledge, there is no research exploiting geolocation, user and temporal information to the end of detecting new related terminology, which do not necessarily co-occur with event-related keywords. This is the main contribution of this work.

3 Query Expansion for Information Filtering using Social Information

In this paper, we propose an Information Filtering approach based on a boolean IR model (Hiemstra, 2009). The reason for not ranking tweets is that we assume all tweets containing event-related keywords will be surely relevant. We apply a query expansion approach for selecting new event-related terms by means of social information.

Pseudo-Relevance Feedback (PRF) via query expansion (Xu and Croft, 1996) has been proven to be effective in many Information Retrieval (IR) tasks. Carpineto et al. (2001) presented a method for term scoring for PRF, where the authors tried to maximize the divergence between the probability distributions of the terms estimated in the pseudo relevance set (p_{PR}) and the distribution estimated over the whole collection (p_C). They used the Kullback-Leibler divergence (KLD) calculated as in Equation 3.1 given that it captures the relative entropy between both distributions (Kullback and Leibler, 1951). To build the expanded query they selected the terms (w) that contribute most to the divergence of both distributions (i.e., higher KLD score).

$$KLD(w) = p_{PR}(w).log\frac{p_{PR}(w)}{p_C(w)} \quad (1)$$

3.1 Increasing the Pseudo Relevant Set

We re-examine the PRF assumption considering the social information stated in the initial hypotheses. Then, we consider the following tweets in order to extend the Pseudo Relevance set (PR) and extended Pseudo Relevance set ($PR*$), and then we apply the PRF process to PR^*:

- The immediately previous ($user_{pre}$) or later ($user_{post}$) tweets from a user that posted a message containing an initial keyword. Both approaches were also combined and tested.

- The messages posted from the nearest geolocations where a tweet containing a initial keyword was found, and considering a radius of 0.1 degrees in latitude and longitude ($geo_{0,1}$). These tweets should be constrained in a reasonably short timeframe (depending on collection). We tried different radius values, but we empirically found that the best results were achieved by using 0.1.
- Tweets containing a hashtag caught within a message where a initial keyword was found.

Hence, our proposal consists in extending the initial PR set to other PR^* set adding tweets from an user, geolocation and time information.

$$KLD^*(w) = \sum_{w \in V} p_{PR^*}(w).log\frac{p_{PR^*}(w)}{p_C(w)} \quad (2)$$

3.2 Temporal_KLD: Modifying KLD for Adding Twitter Temporal Aspect

In this paper we introduce Temporal-KLD (TKLD), a modification of Kullback-Leibler divergence for considering the Twitter temporal aspect within the PRF process. The main idea is to combine the query expansion process applied on PR and PR^* sets with tweets posted in a short timeframe, corresponding to the time when the natural hazard is happening, in order to find wider event-related terminology. It is important to remark that evaluation datasets were collected in a specific geolocation and time (where and when the natural hazard was hitting), as it will be seen in Section 4. We expect to find relevant information within the timeframe when event is happening. We look for into phrases not sharing event-related previously found terminology.

The inherent idea is to combine two PRF processes, one considering the PR or PR^* sets, and the other one considering a tweet set corresponding to the short period of time when hazard is hitting ($p_{Time}(w)$). Then, we propose:

$$TKLD(w) = p_{PR}(w).log\frac{p_{PR}(w)}{p_C(w)}.log\frac{p_{Time}(w)}{p_C(w)} \quad (3)$$

4 Datasets

We collected two Twitter datasets associated with real-world events: "hurricane Isaac" and "hurricane Sandy". These were two hurricanes that hit different parts of the United States in 2012.

For the "**hurricane Isaac**" dataset, we sampled 1,000 geolocated tweets posted from the Lousiana area in August 29^{th}. The selected timeframe was from 8 a.m to 9 a.m, and the specific geolocation was $\pm 2^{\circ}$ latitude and longitude degrees from the hurricane eye according to information of the US National Hurricane Center[1]. We also downloaded tweets from August 19^{th} to August 28^{th} (in the same timeframe and area) to consider background for the estimation of divergences, trying to capture the everyday vocabulary.

For the "**hurricane Sandy**" dataset, we sampled 1,000 geolocated tweets sent from NYC area in October 29^{th}. The selected timeframe was from 8:30 p.m to 8:35 p.m., and the specific geolocation was +-2^{o} from the hurricane eye according to information of the US National Hurricane Center. We also downloaded tweets for background from June 15^{th} to July 15^{th}.

In both cases, the preprocessing was the same: Porter stemming and removing stopword and tokens beginning with numbers or not-alphanumeric characters. With regard to gold-standard creation, tweets from both of the datasets were annotated as being either "related" or "non-related" to the event in question, based on the following guidelines. A tweet is "related" if:

- It explicitly refers to the hurricane.
- It refers to consequences of the hurricane (e.g., power outage).
- It is aware of the hurricane, and would not have been posted otherwise (e.g., concerned about safety of friends)

If it does not provide evidence to be considered as related, it should be categorized as non-related.

Each tweet was annotated via Amazon Mechanical Turk[2] by five US-based users. After the annotation process, 321 tweets were

[1] http://www.nhc.noaa.gov

[2] https://www.mturk.com/mturk/

annotated as related and 679 as non-related in the "hurricane Isaac" dataset, and 318 and 682 as related and non-related tweets in the "hurricane Sandy"dataset.

A Recall-Precision graph is used as a combined evaluation measure.The area under the curve is used as a simple metric to define how an algorithm performs over the whole space. Such a graph, given an arbitrary recall point, tells us the corresponding precision value. At this point it is important to remark that we are interested in increasing recall values, maintaining reasonable precision values. For this reason, we must pay attention to the right upper regions of the Figures.

5 Results

Figures 1 and 2 show the performance that can be achieved following the different query expansion approaches in a Recall-Precision space, and for the first 50 expansion terms. The initial queries are "hurricane OR isaac" and "hurricane OR sandy" and afterwards, a boolean IR model is applied to obtain all tweets containing the query terms and for creating the PR set, that will be used in the later PRF process.

Both Figures 1 and 2 show that, in the case of the baseline, when the PR set only contains tweets with original queries, precision values decrease (increasing recall values) in higher levels than when any social information approach is taken into account (approaches using PR^*), and thus the baseline area under the curve is smaller than when social information is considered. This shows that the use of social information leads to better performance than that achieved by using classic query expansion approaches.

With regard to the query expansion function, the behavior of KLD and TKLD is similar for the first expanded terms in both datasets. However, these positions represent low recall values, far from our objective of increasing the initial recall values (0.49 in "hurricane Isaac" and 0.44 in "hurricane Sandy"). In general, we can observe that TKLD expansion function obtains the best results in both datasets. Considering social information in PRF process we can achieve almost 0.75 in terms of both in precision and Recall, while not using temporal information the performance values are around 0.7 both in terms of precision and recall. While it is true that selection of the first expansion terms by TKLD function in "hurricane Sandy" dataset is worse than using KLD, these positions represent low recall values (around 0.55-0.63). The subsequent expansion term selection is able to maintain precision values in a slight decrease while recall values are increasing (up to 0.8-0.85), as it can be seen in Figure 2.

Analyzing the different social information in both datasets, we do not observe a similar performance. While for the "hurricane Isaac" dataset the geolocation information obtains encouraging results, the user information only works with high values of expanded terms (around $40 - 50$). In addition, in this dataset the hashtag approach does not work, since its results are similar to TKLD using PR. Similar conclusions cannot be drawn for the hurricane Sandy dataset, which occurred in New York City. In this case, the geolocation information decreases the recall values achieved with TKLD using PR; and the user and hashtag information do not contribute since their recall values are similar to the TKLD ones using PR. Their contrasting population densities and different social composition of these populations could be one reason. We think that initial hypotheses, related to geolocation and user, fit better in a small city like Baton Rouge and not in a high density urban environment like NYC.

In summary, using temporal information, natural hazard-related terminology can be found beyond considering co-occurrences with event-related keywords. Nevertheless, for considering user, geolocation and hashtag information, a deeper study must be carried out. There is another important aspect as well. When we expand with user, geolocation or hashtag information, and do not obtain new related terminology, in most of the cases (with the exception of geolocation information in NYC) it does not decrease the Recall-Precision values because the relative entropy between the added tweets and the background is low. The reason is that the terms contained in those added tweets are terms with similar relative frequency in the background, and then the terms are not selected in the first positions by query expansion function.

6 Conclusions

In this paper we have introduced a new approach for information filtering in the context of Twitter coverage of emergency events. We

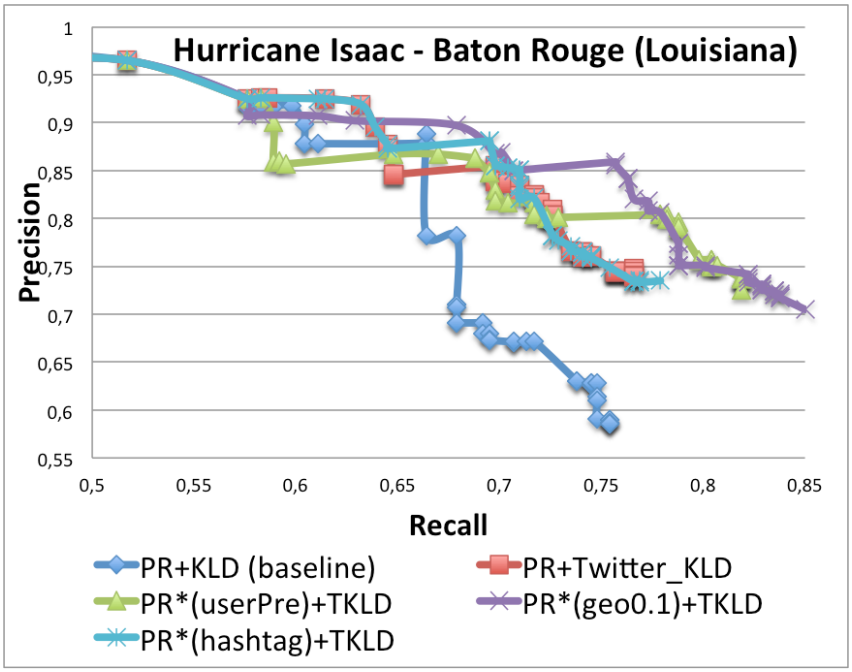


Figura 1: Recall-Precision curve for "Hurricane Isaac" dataset.

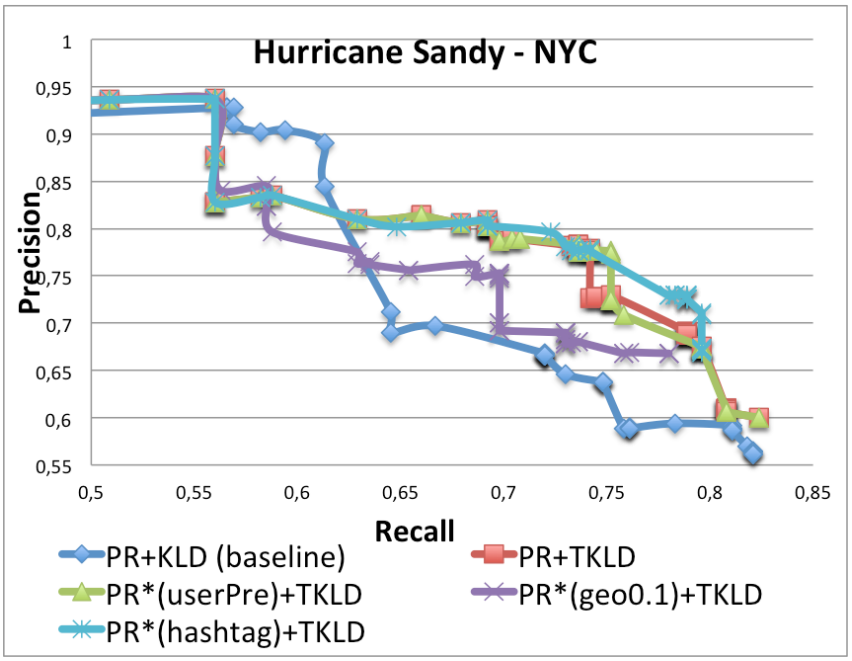


Figura 2: Recall-Precision curve for "Hurricane Sandy" dataset.

have proposed a set of novel approaches that rely on non-textual social features to capture new keywords that are related to an event. The approaches we introduce and experiment in this paper rely on geolocation, user and hashtag information, as well as temporal information in a Pseudo-Relevance Feedback via query expansion approach. Through experiments on two hand-labeled datasets associated with two natural hazards, our preliminary research shows that especially the use of temporal information can have a significant impact in the performance, improving recall values. Moreover, our results suggest that the study that the use of social information for query expansion so as to discover

new keywords related to an event can help boost the performance of the tweet retrieval.

Our plans for future work include a further exploration of the social features inherent in tweets to improve the tweet retrieval. Also, and especially motivated by the fact that previous studies found that the use of Twitter during emergencies is different than its use in other contexts, we would like to explore the applicability of our approaches to other types of events. This study, accompanied by an iterative refinement of the filtering techniques, will allow us to come up with a more generalizable approach.

Acknowledgments

This work has been part-funded by the Spanish Ministry of Science and Innovation (MED-RECORD Project, TIN2013-46616-C2-2-R) and by UNED Project (2012V/PUNED/0004). This research was also partially supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Carpineto, C., R. de Mori, G. Romano, and B. Bigi. 2001. An information-theoretic approach to automatic query expansion. *ACM Trans. Inf. Syst.*, 19(1):1–27.

De Longueville, B., R.S. Smith, and G. Luraschi. 2009. OMG, from here, i can see the flames!: A use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 International Workshop on Location Based Social Networks*, LBSN '09, pages 73–80, New York, NY, USA. ACM.

Deerwester, S., S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Hiemstra, D. 2009. *Information Retrieval Models*. Information Retrieval: Searching in the 21st Century, Wiley.

Hughes, A.L. and L. Palen. 2009. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(3):248–260.

Imran, M., C. Castillo, F. Diaz, and S. Vieweg. 2014. Processing social media messages in mass emergency: A survey. *CoRR*, abs/1407.7071.

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

Mills, A., R. Chen, J. Lee, and H. R. Rao. 2009. Web 2.0 emergency applications: How useful can twitter be for emergency response. *Journal of Information Privacy and Security*, 5:3–26.

Palen, L., K. Anderson, G. Mark, J. Martin, D. Sicker, and D. Grunwald. 2010. A vision for technology-mediated public participation and assistance in mass emergencies and disasters. In The University of Edinburgh, editor, *International Academic Research Conference*, 14–16 April.

Sakaki, T., M. Okazaki, and Y. Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM.

Sasaki, K., S. Nagano, K. Ueno, and K. Cho. 2012. Feasibility study on detection of transportation information exploiting twitter as a sensor. In *Sixth International AAAI Conference on Weblogs and Social Media. Workshop on When the City Meets the Citizen, AAAI Technical Report WS-12-04*.

Schnebele, E. and G. Cervone. 2013. Improving remote sensing flood assessment using volunteered geographical data. *Natural Hazards and Earth System Science*, 13(3):669–677.

Shinohara, M., A. Hattori, S. Ioroi, H. Tanaka, H. Hayami, H. Fujioka, and Y. Harada. 2011. Design and trial of a cellphone-based hazard information sharing system for residents living close to an incident. In *Next Generation Mobile Appli-*

cations, Services and Technologies (NGMAST), 2011 5th International Conference on, pages 13–18, Sept.

Shklovski, I., L. Palen, and J. Sutton. 2008. Finding community through information and communication technology in disaster events. In *Proceedings of the ACM 2008 Conference on Computer Supported Cooperative Work (CSCW 2008*, pages 127–136.

Sinnappan, S., C. Farrell, and E. Stewart. 2010. Priceless tweets! a study on twitter messages posted during crisis: Black saturday. In *Proceedings of Information Systems: Defining and Establishing a High Impact Discipline*, number 39, Brisbane, Australia, 01-03 December 2010. 21st International Conference on Information Systems (ACIS 2010).

Sorensen, J.H. and B. Sorensen. 2007. Community processes: Warning and evacuation. In *Handbook of Disaster Research*, Handbooks of Sociology and Social Research. Springer New York, pages 183–199.

Starbird, K., L. Palen, A.L. Hughes, and S. Vieweg. 2010. Chatter on the red: What hazards threat reveals about the social life of microblogged information. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, pages 241–250, New York, NY, USA. ACM.

Steyvers, M. and T. Griffiths, 2007. *Latent Semantic Analysis: A Road to Meaning*, chapter Probabilistic topic models. Erlbaum, L.

Sutton, J., L. Palen, and I. Shklovki. 2008. Backchannels on the front lines: Emergent use of social media in the 2007 southern california fires. In *Proceedings of the 2008 Information Systems for Crisis Response and Management Conference (ISCRAM 2008)*, pages 624–631, Washington, D.C.

Vieweg, S., A.L. Hughes, K. Starbird, and L. Palen. 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1079–1088. ACM.

Xu, J. and W.B. Croft. 1996. Query expansion using local and global document analysis. In *Proceedings of SIGIR 96*, pages 4–11, Zurich. ACM.

Zubiaga, A., D. Spina, V. Fresno, and R. Martínez. 2011. Classifying trending topics: a typology of conversation triggers on twitter. In C. Macdonald, I. Ounis, and I. Ruthven, editors, *CIKM*, pages 2461–2464. ACM.

Procesamiento del Lenguaje Natural, Revista nº 54, marzo de 2015, pp 93-100 recibido 28-11-14 revisado 27-01-15 aceptado 09-02-15

Recomendación de puntos de interés turístico a partir de la web*

Recommendation of Tourist Points of Interest using the Web as source

Eladio M. Blanco-López, Arturo Montejo-Ráez
Fernando J. Martínez-Santiago, Miguel Á. García-Cumbreras
Universidad de Jaén
23071 - Jaén (Spain)
{emblanco, amontejo, dofer, magc}@ujaen.es

Resumen: Este artículo presenta un sistema de recomendación híbrido, basado en contenido y comunidad de usuarios, para recomendar a los usuarios los lugares próximos más afines a sus gustos. El contenido se extrae de forma automática de la web oficial del punto de interés. Destacamos los buenos resultados obtenidos cuando la información recuperada para cada lugar de su sitio web es descriptiva. Nuestros experimentos se han realizado sobre los datos ofrecidos por la organización del Contextual Suggestion Track en TREC 2014, una tarea exigente donde la información de los usuarios es dispersa y cuyas recomendaciones se deben obtener a partir de coordenadas geográficas y poca información adicional.
Palabras clave: Sistemas de recomendación, extracción de información, turismo

Abstract: This paper introduces a hybrid recommender system, based on both content and communiy of users, to suggest places according to user's interests. The content has been automatically extracted from official web page of each place. We remark the promising results obtained when the official web site provides descriptive content. Our experiments have been performed on the Contextual Suggestion Track dataset from TREC 2014, a competitive task where information about users is very sparse and recommendations must come from only GPS coordinates and few additional information.
Keywords: Recommender systems, information extraction, tourism

1 Introducción

En este trabajo se presenta un sistema de recomendación de puntos de interés turístico próximos a la ubicación geográfica de un usuario. Se propone un sistema híbrido, basado en contenido y comunidad de usuarios, para recomendar a los usuarios los lugares próximos más afines a sus gustos. El contenido se extrae de forma automática de la web oficial del punto de interés. El sistema desarrollado obtiene buenos resultados cuando la información recuperada para cada lugar de su sitio web es descriptiva. El sistema ha sido evaluado sobre la colección de datos de la tarea Çontextual Suggestion Track"del TREC 2014 (Text Retrieval Conference).

Los sistemas de recomendación tienen como objetivo el sugerir, de entre un conjunto de ítems candidatos, aquellos que pueden resultar de mayor interés para el usuario (Ricci, Rokach, y Shapira, 2011). Esto implica disponer de información sobre el usuario y sobre los ítems, aunque se trate únicamente de las evaluaciones (puntuaciones) de los primeros sobre los segundos. Son muchas las aplicaciones de los sistemas de recomendación y se utilizan tres tipos de estrategias principales: basados en contenido (Pazzani y Billsus, 2007), con filtrado colaborativo (Schafer et al., 2007) o mixtos (Adomavicius y Tuzhilin, 2005). Los sistemas basados en contenido analizan la información disponible de cada ítem para extraer una serie de características que puedan cotejarse con las preferencias del usuario. Esto implica, por ende, la necesidad de dicho perfil de usuario, lo cual no siempre es posible. En el caso de los sistemas basa-

* Este trabajo se ha desarrollado gracias a la financiación parcial del proyecto ATTOS (TIN2012-38536-C03-0) del Gobierno de España y del proyecto CEATIC-2013-01 de la Universidad de Jaén.

ISSN 1135-5948 © 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

dos en filtrado colaborativo se trabaja sobre una matriz usuario-ítem cuyos valores suelen ser las puntuaciones (*ratings*) de los usuarios sobre los productos. De esta forma, podemos construir un vector de características de un usuario sobre la base de los ítems puntuados o, por el contrario, vectores de ítems con base en las puntuaciones de los usuarios. El filtrado colaborativo aprovecha estos vectores para buscar usuarios o ítems afines.

Gracias a la llegada de los dispositivos móviles con sensores GPS, ha surgido la posibilidad de añadir dicho elemento como parte de los condicionantes de un usuario. Así, para sugerir puntos de interés turístico (*POI: Point of Interest*) la información de posición representa un elemento fundamental en el ámbito turístico (Horozov, Narasimhan, y Vasudevan, 2006). A partir de este hecho, y debido al creciente aumento de información facilitada por distintos sensores, así como otras cuestiones de relevancia como el momento del día o la meteorología, surgen los denominados *sistemas de recomendación contextuales* (Adomavicius y Tuzhilin, 2011).

Según el informe del The Second Strategic Workshop on Information Retrieval en Lorne (Allan et al., 2012): "*Los sistemas de recomendación futuros deben anticiparse a las necesidades del usuario y responder con la información apropiada al contexto sin que el usuario tenga que realizar una consulta explícita [...]. En un contexto móvil, el sistema se corresponderá a una app que recomendará lugares y actividades interesantes según la ubicación del usuario, preferencias personales, historia pasada y factores del entorno como el clima, tiempo [...]. Al contrario que muchos sistemas de recomendación tradicionales, estos sistemas serán de dominio abierto, capaces de realizar sugerencias y sintetizar información de múltiples fuentes [...]*".

Por ejemplo, se podría imaginar a un investigador de Tecnologías del Lenguaje Humano (TLH) con una tarde libre en el congreso de la SEPLN en Gerona; conociendo algunos lugares visitados en otras ediciones de la SEPLN, un sistema de sugerencias contextual le podría recomendar cenar en el restaurante Massana, visitar los baños árabes o tomar una copa en la terraza del Hotel Gran Ultonia. Nuestro sistema pretende ofrecer una solución a este nuevo paradigma considerando el procesamiento de los textos descriptivos de un lugar parte fundamental del proceso.

El artículo se organiza del siguiente modo: En primer lugar, se presenta una síntesis del estado del arte en esta tarea tomando como referencia principal los experimentos realizados en la Contextual Suggestion Track del 2013. En segundo lugar, se pasa a describir los datos utilizados para la realizar la experimentación. A continuación se describe el sistema propuesto, detallando los módulos implementados en cada una de sus fases. Después se hace una discusión de los resultados obtenidos. Por último, se presentan las conclusiones y las líneas para el trabajo futuro.

2 Estado del arte

En este punto se describe el estado del arte en la tarea, basándonos en los experimentos realizados por los participantes en el TREC 2013 Contextual Suggestion Track y otras publicaciones relevantes en el ámbito de los sistemas de recomendación contextuales orientados a turismo.

En el Contextual Suggestion Track de 2013, los sistemas de Jiang y He (2013), Avula, O'Connor, y Arguello (2013) y Yang y Fang (2013) utilizaron Yelp[1], una red social para la puntuación de lugares, desde restaurantes u hoteles hasta clínicas de fisioterapia, para obtener sugerencias candidatas y comprobar las más similares con los perfiles de usuario. Otros como Bellogin et al. (2013) han utilizado el conjunto de datos cerrado ClueWeb12, recuperando una subcolección de los documentos más relevantes para cada contexto y ordenándola según el perfil. En Luo y Yang (2013) además extraen nombres de lugares de WikiTravel[2] y hacen consultas para crear la colección.

Una estrategia común entre los participantes ha sido la de alimentarse de redes sociales donde se comparten, comentan y puntúan lugares de interés. Así por ejemplo, en McCreadie et al. (2013), Roy, Bandyopadhyay, y Mitra (2013) y Drosatos et al. (2013) se usan redes sociales basadas en posición como Google Places[3], FourSquare[4] y Facebook Places[5] para recopilar lugares. Generalmente la descripción de los mismos se llevaba a cabo mediante información recuperada de sus webs de motores de búsqueda como Google y

[1] http://www.yelp.com
[2] http://wikitravel.org
[3] http://www.google-places.com
[4] http://foursquare.com
[5] http://facebook.com/places

Bing.

En general, estos sistemas se fundamentan en técnicas de recuperación de información donde la función de distancia suele ser probabilística, una combinación lineal de pesos o mediante algoritmos de aprendizaje automático. La diferencia no radica tanto en la tecnología utilizada, sino en la forma en que se genera el contenido sobre el que se calcula el peso final en el ranking de recomendaciones (Dean-Hall et al., 2013). Así por ejemplo, Yang y Fang (2013) (equipo que obtuvo los mejores resultados) proponen un sistema basado en contenido donde el perfil de usuario se construye a partir de los comentarios de otros usuarios a los lugares preferidos de dicho usuario. El equipo de Lugano (Rikitianskii, Harvey, y Crestani, 2013) crea perfiles positivos y negativos a partir de expansiones sobre las descripciones de los sitios encontradas con Yandex y Google Custom Search API. Drosatos y sus colegas (Drosatos et al., 2013) generan el contenido mediante técnicas de *crowdsourcing* a partir de los fragmentos descriptivos que devuelven los motores de búsqueda Google y Bing.

El comercio electrónico ha cambiado la industria del turismo y como se afirma en Werthner y Ricci (2004) es un buen área para la investigación aplicando sistemas de recomendación. En Fesenmaier, Wöber, y Werthner (2006) se hace un análisis de los sistemas de recomendación aplicados al turismo y en Staab et al. (2002) destacan los sistemas de recomendación para viajes haciendo hincapié en la personalización y ubicación como prerrequisitos imprescindibles para el éxito de aplicaciones turísticas. Nuestro grupo tiene experiencia en la construcción de sistemas de recomendación para guiado, con el sistema GeOasis (Martínez-Santiago et al., 2012), si bien este sistema estaba enmarcado dentro de una ontología controlada, en un dominio muy concreto y fundamentado en reglas definidas para la toma de decisiones.

3 Datos para la experimentación

Como entrada para la tarea del *Contextual Suggestion Track*, a los participantes se les proporciona un conjunto de perfiles de usuario, un conjunto de ejemplos de sugerencias y un conjunto de contextos. Cada perfil se corresponde con un usuario e indica sus preferencias con unas sugerencias de ejemplo. Por ejemplo, una sugerencia puede ser tomar una cerveza en el Dogfish Head Alehouse y en el perfil del usuario puede haber una preferencia negativa con respecto a dicha sugerencia. Cada sugerencia de entrenamiento incluye un título, una descripción y una URL asociada. Cada contexto se corresponde con una localización geográfica (una ciudad), como por ejemplo Gaithersburg o Maryland (todas en Estados Unidos).

Para cada par perfil/contexto, los participantes tienen que elaborar una lista ordenada de 50 sugerencias. Cada sugerencia debe ser apropiada para el perfil (basada en las preferencias del usuario) y el contexto (según la localización). Los perfiles están formados por preferencias de usuarios reales, de entre estudiantes y graduados universitarios, los cuales evalúan las sugerencias propuestas. Para los experimentos se ha asumido que los usuarios tienen edad legal de beber, disponen de hasta 5 horas para llegar hasta el lugar sugerido y tienen acceso al transporte apropiado. Esta restricción de tiempo, determinada por los organizadores de la tarea es lo suficientemente holgada como para que la gran mayoría de los sistemas participantes descarten la componente temporal como característica decisiva en la propuesta de sugerencias.

3.1 Ejemplos de sugerencias y perfiles

Los perfiles constan de dos puntuaciones para una serie de sugerencias, una puntuación para el título y la descripción y otra puntuación para el sitio web de la sugerencia en cuestión. Por tanto, el perfil proporciona información sobre qué sugerencias gustan o no a un determinado usuario. Las puntuaciones están basadas en una escala de 5 puntos según lo interesado que esté el usuario en realizar la actividad si estuviese visitando la ciudad en la que se halla el lugar propuesto:

- 4: Altamente interesado
- 3: Interesado
- 2: Neutral
- 1: Desinteresado
- 0: Altamente no interesado
- -1: La web no cargó o no se dio puntuación

Toda la información relativa a usuarios, contextos y sugerencias se proporciona de la siguiente forma:

- `examples2014.csv`: Contiene 100 sugerencias de ejemplo que han sido puntuadas por los usuarios. El formato es `id`, `tittle`, `description`, `URL`.
- `profiles2014-100.csv`: Contiene las puntuaciones dadas por los usuarios a las 100 sugerencias de ejemplo. Se compone de 11.600 puntuaciones realizadas por 115 usuarios. El formato es `id`, `attraction_id`, `description`, `website`.
- `profiles2014-70.csv`: Contiene las puntuaciones dadas por los usuarios a un subconjunto de 70 sugerencias de ejemplo. Se compone de 12.810 puntuaciones realizadas por 182 usuarios. El formato es `id`, `attraction_id`, `description`, `website`.
- `contexts2014.csv`: Son los contextos para realizar las sugerencias se han elaborado con las ciudades principales de 50 áreas metropolitanas seleccionadas aleatoriamente, excluyendo las dos ciudades utilizadas en las sugerencias de ejemplo (Chicago, IL y Santa Fe, NM).

4 *Descripción del sistema propuesto*

Se ha desarrollado un sistema de recomendación basado en contenido donde se ha creado un modelo de espacio vectorial para cada sugerencia (en adelante POI - *Point Of Interest*) y, a partir de las puntuaciones dadas en cada perfil a los POIs facilitados, se ha elaborado un modelo del usuario como agregado ponderado de los vectores de cada POI.

A continuación se presentan las tres fases llevadas a cabo con los módulos implementados en cada una de ellas.

4.1 Fase 1: Generación de los vectores de perfil

En primer lugar, se han obtenido las categorías de los POIs. Para ello, se han comprobado manualmente las categorías de Google Places que se corresponden con los términos de los ejemplos de lugares suministrados. El conjunto de categorías que finalmente considera son 14: *aquarium*, *bakery*, *bar*, *cafe*, *cemetery*, *church*, *food*, *lodging*, *museum*, *park*, *restaurant*, *school*, *spa* y *store*.

En esta primera fase se generan los vectores de perfil de usuario a partir de las puntuaciones y la información extraída de la web de cada POI. En la Figura 1 se muestra el esquema seguido.

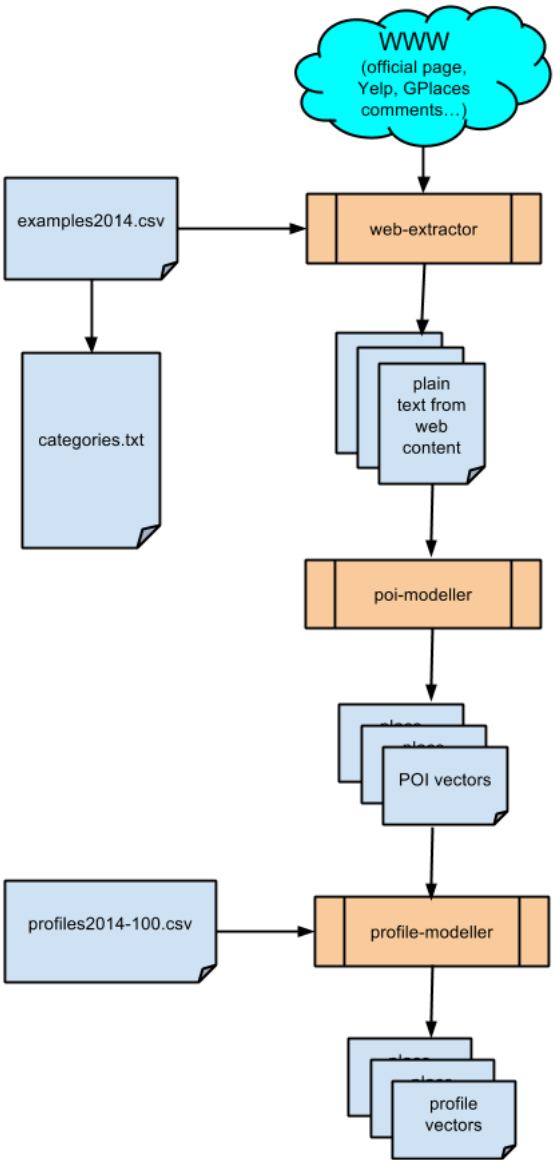


Figura 1: Fase 1

- **web-extractor**. Partiendo de las URLs de los POIs de ejemplo, se extrae información de la web mediante dos procesos: un *crawling* para obtener, a partir de la URL, páginas hasta un segundo nivel de jerarquía, y extracción de términos/características (limpieza del HTML), que genera un documento de texto plano. Se genera así para cada POI un archivo en texto plano y además, se enriquece con información extraída de otras fuentes como Google Places y Yelp.
- **poi-modeller**. A continuación se toma

la colección de documentos generada de POIs para crear los vectores de POIs asignando un peso a cada término mediante TF-IDF. Esto no es sino representar los POIs en el Modelo de Espacio Vectorial (*bag of words*)

- **profile-modeller**. Cada usuario ha realizado una puntuación para la descripción y título del POI y otra para su página web. Para los cálculos realizados se ha tenido en cuenta la media de ambas puntuaciones como puntuación general de un usuario a un POI. Por tanto, a partir de las puntuaciones a POIs de cada usuario y los vectores de POIs anteriores, se genera el vector de cada usuario, construido como la media ponderada por las puntuaciones de los vectores de los POIs evaluados por el usuario.

4.2 Fase 2: Generación de vectores de POIs candidatos

En esta segunda fase, se genera una base de conocimiento de lugares de interés (POIs) usando la API de Google Places. A diferencia de la fase anterior, donde ya teníamos identificados los POIs de entrenamiento evaluados por los usuarios, en esta fase buscamos nuevos POIs para cada uno de los contextos (ciudades) considerados en la competición. Esto implica crear la base de conocimiento a partir de la cual podremos generar recomendaciones. En la Figura 2 se muestra el esquema seguido.

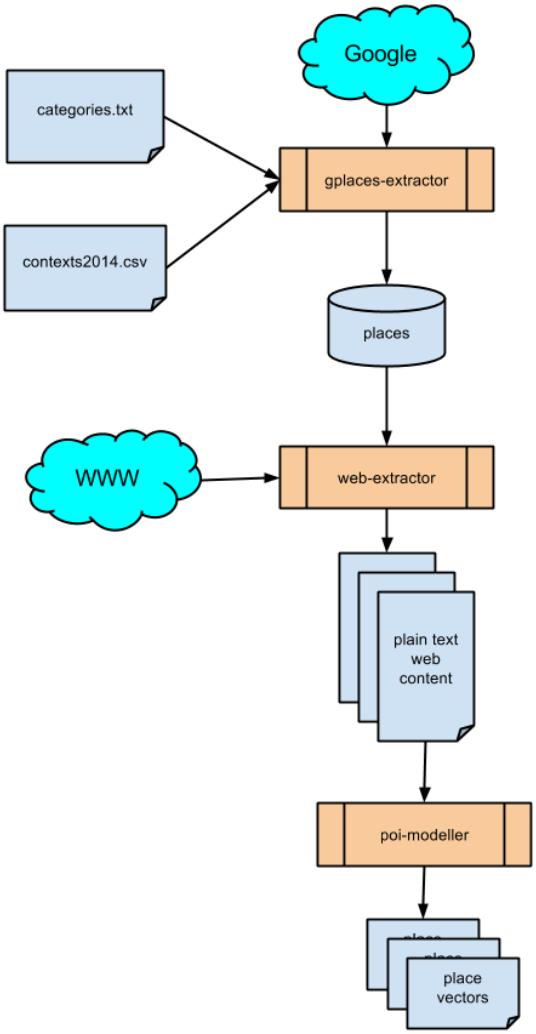


Figura 2: Fase 2

- **gplaces-extractor**. Se parte de las categorías obtenidas en la fase anterior y de un listado de contextos o ciudades donde buscar los POIs. Ambos sirven de entrada a este módulo para extraer los POIs mediante la API de Google Places. Dado un lugar (ciudad) y una lista de tipos de POIs (categorías) se puebla una base de datos NoSQL (MongoDB) con metadatos (incluyendo la URL de la web del POI).
- **web-extractor**. Partiendo de las URLs de los POIs obtenidos, se extrae información de la web mediante *crawling* (sacando, a partir de la URL, páginas hasta un segundo nivel de jerarquía) y limpiado (extracción de términos/características), generando un documento de texto plano por POI, al igual que se vio en la fase anterior. De nuevo, cada archivo de texto plano generado por POI se enriquece con información extraída de otras fuentes como Google Places.
- **poi-modeller**. A continuación se toma la colección de documentos generada de POIs para crear los vectores de POIs asignando un peso a cada término mediante TF-IDF.

4.3 Fase 3: Generación de las sugerencias

En la tercera fase ya tenemos el sistema preparado para generar sugerencias pues disponemos de una representación de los usuarios (fase 1) y una base de conocimiento de POIs (fase 2). Ahora se generan las sugerencias a partir de los vectores de perfiles y los vectores de lugares para cada una de las ciudades

o contextos. Este proceso muestra en la Figura 3.

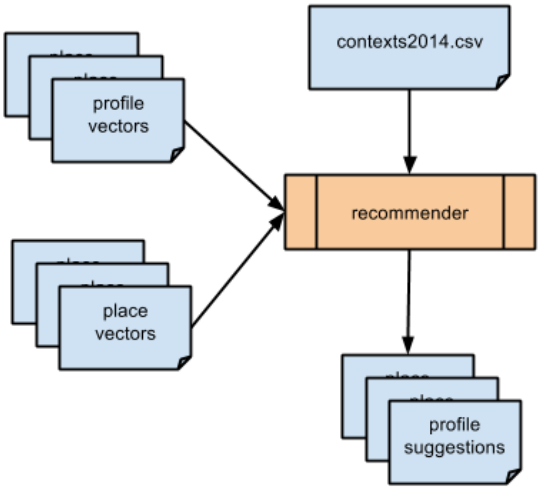


Figura 3: Fase 3

En primer lugar, los vectores de lugares obtenidos en la fase anterior se distribuyen en colecciones según la ciudad o el contexto al que pertenecen. Para cada usuario se hace un ranking de lugares en cada una de las 50 colecciones, utilizando la similitud del coseno entre el vector del usuario y el vector de lugar. Por último, se obtienen las recomendaciones dejando los 50 lugares más similares a cada usuario en cada una de las colecciones.

5 Resultados obtenidos

La organización del Contextual Suggestion Track utiliza las siguientes métricas para evaluar las sugerencias:

- P@5: precisión en los primeros 5 resultados devueltos (Craswell, 2009a).
- MRR: (Mean Reciprocal Rank): medida calculada sobre la base de la posición del primer documento seleccionado por el usuario dentro del ranking propuesto (Craswell, 2009b).
- TBG (Time-Biased Gain): medida calculada basada en el tiempo que le lleva a un usuario validar un resultado. Esto puede tener repercusión según la calidad de la sugerencia (calidad del título y la descripción) así como la calidad del sitio web oficial (Smucker y Clarke, 2012).

En la Tabla 1 se presentan los resultados obtenidos con el sistema propuesto sobre 25 ejecuciones (25 evaluadores) sobre el total de resultados devueltos (50) para cada evaluador y cada contexto.

Medida	**Mejor**	**Media**	**Peor**
P@5	0,7986	0,3491	0,0053
MRR	0,9738	0,5350	0,0069
TBG	3,8452	1,3685	0,0164

Tabla 1: Puntuaciones detalladas

Como puede verse en estos resultados, el sistema tiene un valor de dispersión en su comportamiento bastante alto. El mejor resultado (el mejor valorado por el usuario) tiene una puntuación media alta (P@5 de 0,79) y, además, suele ubicarse en posiciones superiores del ranking (MRR de 0,93). Las mejores sugerencias suelen llevar más tiempo de inspección al evaluador (TBG de 3,84 segundos). Por contra, como hemos comentado acerca de la dispersión, la valoración media de los resultados no es tan alta. Esto nos sugiere que es más deseable una estrategia "menos es más", y proponer un número reducido de sugerencias. Es importante notar que cuando se disponía de webs con información descriptiva, sí se obtenían buenos resultados, como se puede observar en la columna "Mejor"de la tabla 1.

Con respecto a los valores obtenidos en la competición Contextual Suggestion Track del TREC 2014[6], nuestros resultados quedan en la parte final de los rankings (valores concretos no facilitados en el momento de escribir este artículo). Si comparamos estos resultados con los de la edición anterior (2013), vemos que los mejores valores para P@5, MRR y TBG fueron 0,5094, 0,6320 y 2,4474 respectivamente, por lo que nos sentimos satisfechos con el sistema propuesto. Nos anima ver que, a pesar de ser nuestra primera participación, hemos sido capaces de construir un sistema completo usando tecnologías muy similares a las del resto de grupos. Estamos convencidos de que podríamos mejorarlo trabajando el componente de generación de contenido para los POIs, ya que el sistema usado daba un alto porcentaje de contenido vacío.

[6]http://trec.nist.gov/proceedings/proceedings.html (las actas no estaban disponibles en el momento de redactar este artículo)

6 Conclusiones y trabajo futuro

El sistema propuesto contempla una arquitectura completa para la sugerencia de lugares de interés turístico próximos al usuario. La solución adoptada es altamente modular y permite la adición de nuevas fuentes de lugares. El sistema ha sido evaluado sobre los datos facilitados por la organización del Contextual Suggestion Track celebrado en el seno del TREC 2014. Los resultados obtenidos son esperanzadores, pero sin duda nos sugieren la necesidad de mejorar determinados módulos críticos.

El principal problema se ha encontrado en el módulo **web-extractor**, ya que muchas de las webs consultadas no tenían textos descriptivos. Se limitaban a ofrecer contenido multimedia o texto estructurado. Esto supone todo un reto pues no basta con el procesamiento del contenido textual, sino que es necesario ir más allá en la extracción de características concretas que nos permitan generar un descripción lo suficientemente informativa para el usuario. Una de las posibles soluciones es el aplicar técnicas de *slot filling* (Ji y Grishman, 2011) para completar una ficha modelo del lugar.

También como parte del trabajo futuro se pretende adaptar el sistema para utilizarlo como herramienta de Orientación Profesional para alumnos de Educación Primaria en España. Se pretende crear un sistema de recomendación basado en contenido de vídeos profesionales mediante fichas de información de los mismos, adaptando los recursos construidos al español.

Bibliografía

Adomavicius, G. y A. Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749.

Adomavicius, G. y A. Tuzhilin. 2011. Context-aware recommender systems. En *Recommender systems handbook*. Springer, páginas 217–253.

Allan, J., B. Croft, A. Moffat, y M. Sanderson. 2012. Frontiers, challenges, and opportunities for information retrieval: Report from swirl 2012 the second strategic workshop on information retrieval in lorne. En *ACM SIGIR Forum*, volumen 46, páginas 2–32. ACM.

Avula, S., J. O'Connor, y J. Arguello. 2013. A Nearest Neighbor Approach to Contextual Suggestion. En NIST (NIST, 2013).

Bellogin, A., G. Gebremeskel, J. He, A. Said, T. Samar, A. de Vries, J. Lin, y J. Vuurens. 2013. CWI and TU Delft Notebook TREC 2013: Contextual Suggestion, Federated Web Search, KBA, and Web Tracks. En NIST (NIST, 2013).

Craswell, N. 2009a. Mean Reciprocal Rank. *Encyclopedia of Database Systems*, página 1776.

Craswell, N. 2009b. Precission at n. *Encyclopedia of Database Systems*, páginas 2127–2128.

Dean-Hall, A., C. Clarke, N. Simone, J. Kamps, P. Thomas, y E. Voorhees. 2013. Overview of the TREC 2013 Contextual Suggestion Track. En NIST (NIST, 2013).

Drosatos, G., G. Stamatelatos, A. Arampatzis, y P. Efraimidis. 2013. DUTH at TREC 2013 Contextual Suggestion Track. En NIST (NIST, 2013).

Fesenmaier, D., K. Wöber, y H. Werthner. 2006. *Destination recommendation systems: Behavioural foundations and applications*. CABI.

Horozov, T., N. Narasimhan, y V. Vasudevan. 2006. Using location for personalized poi recommendations in mobile environments. En *Applications and the Internet, 2006. SAINT 2006. International Symposium on*, páginas 6–pp. IEEE.

Ji, H. y R. Grishman. 2011. Knowledge base population: Successful approaches and challenges. En *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, páginas 1148–1158. Association for Computational Linguistics.

Jiang, M. y D. He. 2013. PITT at TREC 2013 Contextual Suggestion Track. En NIST (NIST, 2013).

Luo, J. y H. Yang. 2013. Boosting Venue Page Rankings for Contextual Retrieval–Georgetown at TREC 2013 Contextual Suggestion Track. En NIST (NIST, 2013).

Martínez-Santiago, F., F. Ariza-López, A. Montejo-Ráez, y A. Ureña-López. 2012. Geoasis: A knowledge-based geo-referenced tourist assistant. *Expert Systems with Applications*, 39(14):11737 – 11745.

McCreadie, R., M. Albakour, S. Mackie, N. Limosopathan, C. Macdonald, I. Ounis, y B. Dincer. 2013. University of Glasgow at TREC 2013: Experiments with Terrier in Contextual Suggestion, Temporal Summarisation and Web Tracks. En NIST (NIST, 2013).

NIST, editor. 2013. *TREC 2013*.

Pazzani, M. y D. Billsus. 2007. Content-based recommendation systems. En *The adaptive web*. Springer, páginas 325–341.

Ricci, F., L. Rokach, y B. Shapira. 2011. *Introduction to recommender systems handbook*. Springer.

Rikitianskii, A., M. Harvey, y F. Crestani. 2013. University of Lugano at the TREC 2013 Contextual Suggestion Track. En NIST (NIST, 2013).

Roy, D., A. Bandyopadhyay, y M. Mitra. 2013. A Simple Context Dependent Suggestion System. En NIST (NIST, 2013).

Schafer, J., D. Frankowski, J. Herlocker, y S. Sen. 2007. Collaborative filtering recommender systems. En *The adaptive web*. Springer, páginas 291–324.

Smucker, M. y C. Clarke. 2012. Modeling user variance in time-biased gain. En ACM, editor, *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*, páginas 1–10, New York, NY, USA.

Staab, S., H. Werthner, F. Ricci, A. Zipf, U. Gretzel, D. Fesenmaier, C. Paris, y C. Knoblock. 2002. Intelligent systems for tourism. *IEEE Intelligent Systems*, 17(6):53–64.

Werthner, H. y F. Ricci. 2004. E-commerce and tourism. *Communications of the ACM*, 47(12):101–105.

Yang, P. y H. Fang. 2013. An Opinion-aware Approach to Contextual Suggestion. En NIST (NIST, 2013).

Tesis

Procesamiento del Lenguaje Natural, Revista nº 54, marzo de 2015, pp 103-106 recibido 23-11-14 revisado 27-01-15 aceptado 10-02-15

SSG: Simplified Spanish Grammar. An HPSG Grammar of Spanish with a reduced computational cost

SSG: Simplified Spanish Grammar. Una gramática del español de tipo HPSG de coste computacional reducido

Benjamín Ramírez González
Qindel Group
Príncipe de Vergara, 204, 28002 Madrid
bramirez@qindel.com/benjaminramirezg@gmail.com

Abstract: PhD Thesis written by Benjamín Ramírez González at the Universidad Complutense de Madrid, under the supervision of Dr. Fernando Sánchez León (Real Academia Española, Technology Department). It was defended on February 25th, 2014 at the Instituto Universitario Ortega y Gasset, and it was awarded Summa Cum Laude. The members of the committee were José Lázaro Rodrigo (Universidad Complutense de Madrid), Guadalupe Aguado de Cea (Universidad Politécnica de Madrid), Montserrat Marimón Felipe (Universidad de Barcelona), Olga Fernández Soriano (Universidad Autónoma de Madrid) and Cristina Sánchez López (Universidad Complutense de Madrid).
Keywords: HPSG, computational grammar, Spanish grammar, computational complexity, reduction of computational cost, lexical rules reduction, diathesis alternations, clitics, word order.

Resumen: Tesis escrita por Benjamín Ramírez González en la Universidad Complutense de Madrid, bajo la dirección del doctor Fernando Sánchez León (Departamento de Tecnología de la Real Academia Española). La tesis fue defendida el 25 de febrero de 2014 en el Instituto Universitario Ortega y Gasset y obtuvo una calificación de sobresaliente cum laude. El tribunal lo formaron los doctores José Lázaro Rodrigo (Universidad Complutense de Madrid), Guadalupe Aguado de Cea (Universidad Politécnica de Madrid), Montserrat Marimón Felipe (Universidad de Barcelona), Olga Fernández Soriano (Universidad Autónoma de Madrid) y Cristina Sánchez López (Universidad Complutense de Madrid).
Palabras clave: HPSG, gramática computacioanal, gramática del español, complejidad computacional, reducción de coste computacional, reducción de reglas léxicas, alternancias de diátesis, clíticos, orden de palabras.

1 Objectives and motivation

This PhD Thesis presented SSG (Simplified Spanish Grammar), an HPSG (Head-driven Phrase Structure Grammar) Spanish Grammar.

Every computational grammar of a natural language must face the challenging problem of ambiguity. In order to analyze a sentence in a natural language, an HPSG grammar must generate all possible behavioral patterns of every word in the sentence in the first stages of the process, and then try all possible combinations. In fact, the result in non-trivial cases is a combinational explosion of hypothetical behavioral patterns.

This thesis aims to develop the core of an HPSG grammar of Spanish with a really small amount of lexical rules, which has been named Simplified Spanish Grammar (SSG). It is claimed that SSG analysis are elegant and theoretically motivated, and such analysis significantly reduces the computational cost of grammar and improves analysis times.

2 Structure of the thesis

Three main groups of central phenomena in Spanish have been implemented in SSG.

The first phenomenon is diathesis alternations. From a computational point of

ISSN 1135-5948 © 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

view, this is one of the most challenging phenomena in natural languages as verbs can usually behave in very different ways: they may have both active and passive versions, they may accept certain optional complements, and so on. HPSG lexical rules are meant to deal with these alternations.

Traditional computational grammars usually deal with this diversity by means of specialized lexical rules or lexical units to: transitive verbs with nominal object, transitive verbs with nominal object and dative, transitive verbs with clausal object, transitive verbs with clausal object and dative, and so on. This traditional approach fails to capture due generalizations. Every grammatical reality (transitivity, passive, and a certain kind of dative complement) should be implemented just once. Moreover, argumental positions can be filled with different types of phrases, which mean that both clausal and nominal objects should be considered different fillers available to the same argumental position in the same pattern. This thesis develops a system in which every intuitive verbal pattern is implemented with a unique lexical rule.

The second central grammatical phenomenon implemented in SSG is the Spanish clitics system. Clitization in HPSG has always been formalized by means of lexical rules. By following this approach, many lexical rules and clitization patterns can be added to grammar, which can become a great source of complexity. In Spanish, both accusative and dative arguments can suffer clitization. Moreover, depending on the context, a clitic can appear instead of its canonical object or beside it. Therefore, this thesis develops an analysis of clitics that avoids using any rule or lexical unit intended to deal with clitics.

The last grammatical phenomenon implemented in an innovative way in SSG is word order. The possibilities of word order are a great source of complexity in every Spanish computational grammar. First of all, canonical preverbal subjects can be inverted in several contexts. That inversion has been implemented in traditional HPSG grammars by means of a lexical rule, which leads to a bigger combinational explosion of patterns. At the same time, post-verbal complements can switch their canonical positions, maybe only in a specific context, with certain intonation patterns and with different informational purposes. SSG proposes an analysis of subjects as postverval complements. This proposal is plausible in a theoretical way and contributes to reduce the combinational explosion of grammar. At the same time, in SSG, post-verbal linearization of complements is implemented, according to the classical Linearization Theory in HPSG, as non-continuous constituents.

Finally, it has been added a compared analysis of the same test suit both with SSG and NSSG (Non Simplified Spanish Grammar). NSSG is a traditional grammar whose analysis of diathesis alternation, clitics and word order use the traditional lexical rules. In order to analyze this test suite, as a part of this thesis, SGP (Simplified Grammars Parser) has been developed. SGP is a bunch of libraries written in Perl. SGP provides all the needed tools to analyze written text with HPSG grammars. Moreover, it provides all the needed tools to analyze with SSG, such as a library that joins clitics and verb, as well as a parser compatible with discontinuous constituents.

3 Contributions and future work

It is claimed that SSG analysis are elegant and theoretically motivated, and such analysis significantly reduces the computational cost of grammar and its analysis times. Specifically, these are the main contributions of SSG.

3.1 Theoretical contributions: non-destructive lexical rules

In this thesis it has been coined the term non-destructive rule. Usually, in HPSG, all verbs are supposed to have a canonical characterization, and lexical rules are intended to change that canonical pattern into another. These rules destruct a feature structure and create another one. Crucially, input and output are not supposed to be necessarily compatible. The result is that an HPSG rule is able to change its input in almost every way: it can add or remove an argument, change its category, its case, its position and so on. Unlike previous grammars, lexical rules used by SSG are non-destructive rules. Non-destructive rules never change their input structure, they only specify them. In a non-destructive rule, input and output must share their feature structure and both structures must be identical. Those rules take an underspecified verb and specify it by adding information compatible with their original characterization. The non-destructive rule system is easier to implement and maintain than

a traditional system. This approach has theoretical significance. Every science aims to explain as much data as possible with a theoretical system in the simplest way possible. HPSG lexical rules can operate almost every conceivable change in input and this power reduces HPSG's explanatory capacity. A non-destructive lexical rules system can entirely solve this problem. All non-destructive rules can be reduced, in fact, to a single universal operation: specification, application of an independently-legitimated behavioral pattern.

3.2 A drastic reduction of lexical rules by means of a linguistically motivated analysis

SSG deduces syntactic behavior of verbs from their semantic characterization. Verbs in SSG are really under-specified in a syntactic sense, but they feature a rich semantic characterization. It has been assumed that syntactic alternatives share a common semantic background. A classic semantic characterization has been used: verbs can be accomplishments, achievements, activities or states. According to this main classification, the semantic feature structure of verbs informs about the possible presence of an external argument, an inner argument, and the ability of the verb to receive a certain kind of dative complements or certain controlled predicates. Verbs are also crucially characterized by relevant syntactic features: their ability to assign accusative case or government idiosyncrasies. All these features are well-known verbal characterization criteria, so it is safe to say that they are natural and linguistically motivated. The interesting point is that, just by means of a system of several simple, classic notions, it is possible to develop a general grammar of diathesis alternations of Spanish verbs in a non-destructive fashion. On the other hand, lexical rules restring the nature of their arguments in an interesting way. SSG has a general description of the general notion of argument and it also has a description of case: nominative, accusative, dative and obliq cases. The confluence of all these notions, as well as several semantic idiosyncrasies of certain verbs, successfully regulates the nature of the fillers of every argumental position.

Moreover, in SSG clitics are verbal affixes. Thanks to this morphological approach, SSG avoids using a grammatical rule to merge clitics and verb. In SSG, clitics information is added to the verb by means of an inflectional rule. Note that inflectional rules do not trigger combinational explosion, because they are applied separately and only if pre-syntactic analysis (tokenization) has found actual clitics in the verb. In SSG, clitics are not considered fillers available to an argumental position. Rather, they are only the morphological mark that certain words have left in the verb when they have filled their accusative or dative position. These words are personal pronouns, elliptic pronouns and traces left in topicalization processes. This thesis claims that these words exist in grammar independently of clitics. The outcome is a system of clitics that does not add complexity to the grammar.

Finally, SSG features innovative analysis of Spanish word order. In Spanish, subjects are typically pre-verbal arguments. But a grammar with canonical preverbal subjects features a systematic ambiguity between local and topicalized subjects. In order to reach a simplified and computationally efficient analysis of subject linearization, SSG regards subjects as originally post-verbal arguments where pre-verbal subjects are the result of a topicalization. It is claimed that this approach is plausible in theoretical terms, it solves ambiguity (all preverbal subjects are topics) and reduces the computational cost of grammar. Post-verbal complements in Spanish can be sorted in many ways (scrambling). SSG analysis of scrambling leads to a great simplification of grammar. This solution is a technical application for Spanish of a well known theoretical proposal in HPSG. The key idea is to use discontinuous constituents: all arguments are always listed in the same order in the verb. However, the parser is able to merge two constituents no matter if they are adjacent. In that case, all these arguments, which are always listed in the same order, can be found in different relative positions. This approach has not been applied to traditional computational grammars because traditional parsers cannot deal with this kind of discontinuous constituents. In this thesis, it has been implemented a parser able to do that. For this reason, SSG does not need any rule to deal with scrambling as all complements are always listed in the verb according to a unique increasing order of obliquity.

3.3 A drastic reduction of analysis time

SSG proposals significantly reduces the computational cost of grammars and its analysis times, as it was proved by this work in an empirical way.

For a future work, it would be interesting to improve the current version of SGP (Simplified Grammars Parser). It should include a wrapper to a C library in order to perform feature structures unification tasks in an efficient way, and it should include support to statistical information. On the other hand, SSG coverage should be expanded.

4 References

Bender, E., D. Flickinger and S. Oepen. 2002. The Grammar Matrix: An Open-Source Starter-Kit for the Rapid Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammars. CSLI, Stanford University.

Chomsky, N. 1956. Three Models for the Description of Language. *IRE Transactions PGIT* 2, pp. 113–124.

Donohue, C. and I. Sag. 2006. Domains in Warlpiri. Stanford University.

Fernández Soriano, O. 1993. Sobre el orden de lapabras en español. *Cuadernos de Filología Hispánica* II, pp. 113–152.

Flickinger, D. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering* 6, pp. 15–28.

Gazdar, G., E. Klein, G. Pullum and I. Sag. 1985. *Generalized Phrase Structure Grammar*. Harvard University Press.

Levin, B. and M. Rappaport. 1995. *Unaccusativity at the Syntax-Lexical Semantics Interface*. Massachusetts Institute of Technology (MIT Press).

Marimon, M. 2013. The Spanish DELPH-IN Grammar. *Languages Resources and Evaluation* 47.

Monachesi, P. 1998. Decomposing Italian Clitics. In *Romance in HPSG*, pp. 305–357. CSLI Publications.

Müller, S. 2004. Continuous or Discontinuous Constituents? A Comparison between Syntactic Analyses for Constituents Order and Their Processing System. *Research on Language and Computation* 2, pp. 209-257.

Pineda, L. and I. Meza 2005. The Spanish Pronominal Clitic System. *Procesamiento del lenguaje natural* 34, pp. 67–104.

Pollard, C., R. Kasper and R. Levine 1993. Studies in Constituent Ordering: Toward a Theory of Linearization in HPSG. Grant Proposal to the National Science Foundation, Ohio State University.

Ramírez González, B. 2014. Hacia un modelo computacional unificado del lenguaje natural. *Linguamática* 5:2.

Sánchez León, F. 2006. Gramáticas y Lenguajes Formales. Departamento de Lingüística Computacional de la Real Academia Espñola.

Wall, L. and R. Schwartz. 1991. *Programming Perl*. O'Reilly Media.

Procesamiento del Lenguaje Natural, Revista nº 54, marzo de 2015, pp 107-110 recibido 28-11-14 revisado 27-01-15 aceptado 10-02-15

Negation and Speculation Detection in Clinical and Review Texts[1]

Detección de la Negación y la Especulación en Textos Médicos y de Opinión

Noa P. Cruz Díaz
Dpto. de Tecnologías de la Información. Universidad de Huelva
Ctra. Huelva - Palos de la Frontera s/n. 21819 La Rábida (Huelva)
noa.cruz@dti.uhu.es

Abstract: PhD Thesis written by Noa P. Cruz Díaz at the University of Huelva under the supervision of Dr. Manuel J. Maña López. The author was examined on 10th July 2014 by a committee formed by the doctors Manuel de Buenaga (European University of Madrid), Mariana Lara Neves (University of Berlin) and Jacinto Mata (University of Huelva). The PhD Thesis was awarded Summa cum laude (International Doctorate).
Keywords: Negation and speculation detection, machine learning, biomedicine, sentiment analysis.

Resumen: Tesis doctoral realizada por Noa P. Cruz Díaz en la Universidad de Huelva bajo la dirección del Dr. Manuel J. Maña López. El acto de defensa tuvo lugar el jueves 10 de julio de 2014 ante el tribunal formado por los doctores Manuel de Buenaga (Universidad Europea de Madrid), Mariana Lara Neves (Universidad de Berlín) y Jacinto Mata (Universidad de Huelva). Obtuvo mención internacional y la calificación de Sobresaliente Cum Laude por unanimidad.
Palabras clave: Detección de la negación y la especulación, aprendizaje automático, biomedicina, análisis de sentimientos.

1 Introduction

Negation and speculation are complex expressive linguistic phenomenas which have been extensively studied both in linguistic and philosophy (Saurí, 2008). They modify the meaning of the phrases in their scope. Negation denies or rejects statements transforming a positive sentence into a negative one, e.g., Mildly hyperinflated lungs without focal opacity. Speculation, also known as hedging, it is used to express that some fact is not known with certainty, e.g., Atelectasis in the right mid zone is, however, possible. These two phenomenas are interrelated (De Haan, 1997) and have similar characteristics in the text: they both have scope, so affect part of the text which is denoted by the presence of negation or speculation cue words.

Nowadays, negation and speculation detection is an emergent task in Natural Language Processing (henceforth, NLP). In recent years, several challenges and shared tasks have included the extraction of these language forms such as the BioNLP'09 Shared Task 3 (Kim et al., 2009), the CoNLL-2010 Shared Task (Farkas et al., 2010) or the SEM 2012 Shared Task (Morante and Blanco, 2012).

Detecting uncertain and negative assertions is relevant in a wide range of applications such as information extraction (henceforth, IE), interaction detection, opinion mining, sentiment analysis, paraphrasing and recognising textual entailment (Farkas et al., 2010; Konstantinova et al., 2012; Morante and Daelemans, 2009a; Morante and Daelemans, 2009b). For all of these tasks it is crucial to know when a part of the text should have the opposite meaning (in the case of negation) or should be treated as subjective and non-factual (in the case of speculation). This part of the text is what is known as scope.

At first glance, negation and speculation might seem easy to deal with. The problem could be broken down into finding negative and hedge cues and determining their scope. However, it is much more problematic.

[1] This thesis can be downloaded from http://www.sepln.org/wp-content/uploads/2014/09/NEGATION-AND-SPECULATION-Q9.pdf

ISSN 1135-5948 © 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

Negation and speculation play a remarkable role towards understanding text and pose considerable challenges. They interact with many other phenomenas and they are used for so many different purposes that a deep analysis is needed (Blanco and Moldovan, 2011b).

This thesis is focused on the two domains in which negation and hedging have drawn more attention: the biomedical domain and the review domain. In the first one, negation and speculation detection can help in tasks like Protein-Protein interaction or Drug-Drug interaction. This particular area has been the focus of much current research, mainly due to the availability of the BioScope corpus (Szarvas et al., 2008); a collection of clinical documents, full papers and abstracts annotated for negation, speculation and their scope. In the review domain; opinion mining, sentiment analysis and polarity identification are examples of improvable tasks through the identification of negation and speculation. In all these tasks, distinguishing between objective and subjective facts is crucial and therefore negative and speculative information must be taken into account. Despite its importance and the interest of some authors to explore other areas apart from biomedical (Morante and Daelemans, 2012), the impact of negation and speculation detection in the review domain has not been sufficiently considered compared to the biomedical domain.

2 Contributions

The aim of this thesis is to contribute to the ongoing research on negation and speculation in the Language Technology community. In the medical domain, a system based on machine-learning techniques that identifies negation and speculation cues and their scope in clinical texts is proposed (Cruz et al., 2012).

Additionally, and due to the tokenization problems encountered during the pre-processing of the BioScope corpus and the lack of guidance in this respect, this thesis closely describes this issue and provide both a comprehensive overview analysis and evaluation of tokenization tools. This means, the first comparative evaluation study of tokenizers in the biomedical domain which could help developers to choose the best tokenizer to use.

In the sentiment analysis and opinion mining domains, and contrary to what happens in the biomedical field, there are no publicly available standard corpora of reasonable size annotated with negation and hedging. Therefore, the first step was the participation in the annotation process of the SFU Review corpus with negative and speculative keywords and their linguistic scope. It represents the first corpus annotated with this kind of information in the review domain. Next, using the corpora previously described as well as following the approach used in the biomedical domain, a system to automatically detect negation and hedge cues and their scope is presented.

3 Structure of the thesis

An outline of the thesis is described below.

Chapter 2 begins with an introduction to the definition of negation and speculation from different perspectives, including a classification of the different types of each of them. After briefly motivating the importance of processing these language forms, this chapter presents the related work that inspired and motivated our work, both in the biomedical domain and in sentiment analysis.

Chapter 3 is dedicated to the tokenization problem in the biomedical domain with the aim of helping developers in the decision of choosing the best tokenizer to use. Therefore, this chapter provides an analysis of the problematic cases that the nature of the biomedical field introduces as well as a comprehensive comparative study of the available tools. Finally, it includes the evaluation of the 2 tokenizers that show better features and more accuracy and consistency in the previous study.

Chapter 4 is an in-depth description of the negation and speculation detection system for the clinical domain, explaining every step of the development process. It also presents the corpora used to build the system that accompanies it. Finally, this chapter describes how the system is evaluated and gives details about the experimentation, showing the results obtained and the discussion and error analysis around them.

Chapter 5 presents the developed system for the negation and hedging detection in review texts. It includes the description of the corpora used to train and test the system and the methodology followed. The corpora have been

previously annotated for this task so their annotation process is also specified. It describes the evaluation process; the experiments performed as well as it details the system performance. A discussion and error analysis are also presented in this chapter.

Chapter 6 sums up the outcomes of the work done in this thesis and discuss the possibilities for future work.

4 Conclusions

This thesis tackles negation and speculation treatment in computational linguistics in the two fields which have received more attention: biomedical and review.

In the biomedical domain, a machine-learning system that identifies the negation/speculation cues and their scope in clinical texts has been developed, using the clinical sub-collection of the BioScope corpus as a learning source and for evaluation purposes. For this reason, the proposed approach may not be generalisable to other domains because the expectations in terms of effectiveness could be different if it was used in a corpus with other features, such as scientific texts. The proposed approach achieves an F_1 of 97.3% and 94.9% in negation and speculation cue detection, respectively. In the scope recognition, the system reports F_1 values of 90.9% in negation and 71.9% in speculation. These results show the superiority of the machine-learning-based approach regarding the use of regular expressions. In fact, in the detection of negation expressions, the developed system outstrips the F_1 of NegEx (Chapman et al., 2001) by 30%. In speculation, the proposed method beats the F_1 of the best system by more than 10%. In addition, compared to other approaches based on machine-learning techniques, the developed global system correctly determines approximately 20% more than the scopes identified by Morante and Daelemans (2009b) in negation. In speculation, this difference is greater and the proposed approach correctly recognises nearly twice the number of scopes identified by Morante and Daelemans (2009a). This means improving the results to date for the sub-collection of clinical documents. However, much still remains to be done since scope detector performance is far from having reached the level of well established tasks such as parsing, especially in speculation detection.

Also in the biomedical field, this thesis includes a comprehensive overview study of tokenization tools. Choosing the right tokenizer in this domain is a non-trivial task so this contribution aims to provide a valuable guideline for NLP developers in the biomedical field to select the appropriate tokenizer as the first phase of a text mining task. Specifically, all the biomedical domain difficulties, together with what could be considered to be the correct tokenization in each of these difficult cases are detailed. The process followed to create the list of tools for tokenizing texts to analyse is also explained, including a description of the technical, functional and usability criteria employed to asses each of these tokenizers. After analyzing 21 tools according to the criteria, 13 of them are tested on a set of 28 sentences from the BioScope corpus. Finally, the two tokenizers that show better features and more accuracy and consistency in the examples tested in the previous phase are evaluated in a subset of sentences of this corpus. This contribution means, as far as we are aware, the first comparative evaluation carried out on tokenizers in the biomedical field.

In the review domain, although negation and speculation recognition can help to improve the effectiveness of sentiment analysis and opinion mining tasks, there is just a few works on detecting negative information. Besides, there is, as far as we are aware, no work in identifying speculation. Therefore, this thesis aims to fill this gap through the development of a system which automatically identifies both negation and speculation keywords and their scope. It means the first attempt to detect speculation in the review domain. The novelty of this contribution also lies in the fact that, to the best of our knowledge, this is the first system trained and tested on the SFU Review corpus (Konstantinova et al., 2012). This corpus is extensively used in opinion mining and consists of 400 documents annotated with negative and speculative information. Overall, the results are competitive and the system is portable. In fact, the results reported in the cue detection task (92.37% and 89.64% in terms of F_1 for negation and speculation, respectively) are encouragingly high. In the case of the speculation, the results are comparable to those obtained by a human annotator doing the same task. In the scope detection task, the results are

promising and the system correctly identifies 80.26% full scopes in negation and 71.43% in speculation. The proposed approach outstrips the baseline by as much as about 20% in the negation cue detection and improves it up by roughly 13% in scope detection.

Acknowledgements

This thesis has been funded by the University of Huelva (PP10-02 PhD Scholarship), the Spanish Ministry of Education and Science (TIN2009-14057-C03-03 Project) and the Andalusian Ministry of Economy, Innovation and Science (TIC 07629 Project).

References

Blanco, E. and Moldovan, D. I. 2011b. Some issues on detecting negation from text. FLAIRS Conference.

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F. and Buchanan, B. G. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34, 301-310.

Cruz Díaz, N. P., Maña López, M. J., Vázquez, J. M. and Álvarez, V. P. 2012. A machine-learning approach to negation and speculation detection in clinical texts. *Journal of the American Society for Information Science and Technology*, 63(7), 1398-1410.

De Haan, F. 1997. *The interaction of modality and negation: A typological study Taylor and Francis.*

Farkas, R., Vincze, V., Móra, G., Csirik, J. and Szarvas, G. 2010. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. En *Proceedings of the Fourteenth Conference on Computational Natural Language Learning---Shared Task*, páginas 1-12.

Kim, J., Ohta, T., Pyysalo, S., Kano, Y. and Tsujii, J. 2009. Overview of BioNLP'09 shared task on event extraction. En *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, páginas 1-9.

Konstantinova, N., de Sousa, S., Cruz, N., Maña, M. J., Taboada, M. and Mitkov, R. 2012. A review corpus annotated for negation, speculation and their scope. LREC, 3190-3195

Morante, R. and Blanco, E. 2012. * SEM 2012 shared task: Resolving the scope and focus of negation. En *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, páginas 265-274.

Morante, R. and Daelemans, W. 2009a. Learning the scope of hedge cues in biomedical texts. En *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, páginas 28-36.

Morante, R. and Daelemans, W. 2009b. A metalearning approach to processing the scope of negation. En *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, páginas 21-29.

Morante, R. and Daelemans, W. 2012. ConanDoyle-neg: Annotation of negation in conan doyle stories. En *Proceedings of the Eighth International Conference on Language Resources and Evaluation*.

Saurí, R. 2008. *A factuality profiler for eventualities in text*.

Szarvas, G., Vincze, V., Farkas, R. and Csirik, J. 2008. The BioScope corpus: Annotation for negation, uncertainty and their scope in biomedical texts.

Información General

SEPLN 2015

XXXI CONGRESO DE LA SOCIEDAD ESPAÑOLA PARA EL PROCESAMIENTO DEL LENGUAJE NATURAL

Universidad de Alicante – Alicante (España)
16-18 de septiembre 2015
http://www.sepln.org/ y http://gplsi.dlsi.ua.es/sepln15/

1 Presentación

La XXXI edición del Congreso Anual de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) se celebrará los días 16, 17 y 18 de septiembre de 2015 en la Universidad de Alicante.

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.
- Análisis automático del contenido textual.

ISSN 1135-5948
© 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

- 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 2015

Presidente del Comité Organizador

- Patricio Martínez-Barco (Universidad de Alicante)

Colaboradores

- Rafael Muñoz (Universidad de Alicante)
- Andrés Montoyo (Universidad de Alicante)
- Estela Saquete (Universidad de Alicante)
- Paloma Moreda (Universidad de Alicante)
- David Tomás Universidad de Alicante)
- Maite Romá (Universidad de Alicante)
- José Manuel Gómez (Universidad de Alicante)
- Armando Suárez (Universidad de Alicante)
- Javier Fernández (Universidad de Alicante)
- Yoan Gutiérrez (Universidad de Alicante)
- Fernando Peregrino (Universidad de Alicante)
- Elena Lloret (Universidad de Alicante)
- Isabel Moreno (Universidad de Alicante)
- Ester Boldrini (Universidad de Alicante)

6 Consejo Asesor

Miembros:

- Manuel de Buenaga Rodríguez (Universidad Europea de Madrid, España)
- 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)
- 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)
- Raquel Martínez Unanue (Universidad Nacional de Educación a Distancia, España)
- Leonel Ruiz Miyares (Centro de Lingüística Aplicada de Santiago de Cuba, Cuba)
- Ruslan Mitkov (Universidad de Wolverhampton, Reino Unido)
- Investigador Manuel Montes y Gómez (Instituto Nacional de Astrofísica, Óptica y Electrónica, México)
- Lidia Ana Moreno Boronat (Universidad Politécnica de Valencia, España)
- Lluis Padró (Universidad Politécnica de Cataluña, España)
- Manuel Palomar Sanz (Universidad de Alicante, España)
- Ferrán Pla (Universidad Politécnica de Valencia, España)
- Germán Rigau (Universidad del País Vasco, España)
- Horacio Rodríguez Hontoria (Universidad Politécnica de Cataluña, España)
- Kepa Sarasola Gabiola (Universidad del País Vasco, España)
- Emilio Sanchís (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)
- Mª Felisa Verdejo Maillo (Universidad Nacional de Educación a Distancia, España)
- Rafael Valencia García (Universidad de Murcia, España)
- Manuel Vilares Ferro (Universidad de la Coruña, 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 2015.
- Notificación de aceptación: 1 de mayo de 2015.
- Fecha límite para entrega de la versión definitiva: 15 de mayo de 2015.
- Fecha límite para propuesta de talleres y tutoriales: 23 de marzo de 2015.

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 Maillo
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)

© 2015 Sociedad Española para el Procesamiento del Lenguaje Natural

M. Antonia Martí Antonín	Universidad de Barcelona (España)
M. Teresa Martín Valdivia	Universidad de Jaén (España)
Patricio Martínez-Barco	Universidad de Alicante (España)
Raquel Martínez	Universidad Nacional de Educación a Distancia (España)
Leonel Ruiz Miyares	Centro de Lingüística Aplicada de Santiago de Cuba (Cuba)
Ruslan Mitkov	Universidad de Wolverhampton (Reino Unido)
Manuel Montes y Gómez	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)
Lidia Moreno	Universidad Politécnica de Valencia (España)
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)
Kepa Sarasola	Universidad del País Vasco (España)
Emilio Sanchís	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
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)
Felisa Verdejo Maillo	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)

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/

© 2015 Sociedad Española para el Procesamiento del Lenguaje Natural