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

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

Comité Editorial

Consejo de redacción

ISSN: 1135-5948
ISSN electrónico: 1989-7553
Depósito Legal: B:3941-91
Editado en: Universidad de Alicante
Año de edición: 2015
Editores: Patricio Martínez Barco, Universidad de Alicante, patricio@dlsi.ua.es
Borja Navarro Colorado, Universidad de Alicante, borja@dlsi.ua.es
Sonia Vázquez Pérez, Universidad de Alicante, svazquez@dlsi.ua.es
M. Teresa Romá Ferri, Universidad de Alicante, mtr.ferri@ua.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

Leonel Ruiz Miyares	Centro de Linguística 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 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

Ester Boldrini	Universidad de Alicante
Arantza Casillas	Universidad del País Vasco
Noa Cruz Díaz	Universidad de Huelva
Manuel C. Díaz-Galiano	Universidad de Jaén
Javier Fernández Martínez	Universidad de Alicante
Víctor Fresno	UNED
Diego Gachet	Universidad Europea de Madrid
Miguel Ángel García Cumbreras	Universidad de Jaén
José Manuel Gómez	Universidad de Alicante
Yoan Gutiérrez	Universidad de Alicante
Salud M. Jiménez Zafra	Universidad de Jaén
Gorka Labaka	Universidad del País Vasco
Elena Lloret	Universidad de Alicante
Manuel Maña	Universidad de Huelva
Eugenio Martínez Cámara	Universidad de Jaén
Fernando Martínez Santiago	Universidad de Jaén
Soto Montalvo	Universidad Rey Juan Carlos
Arturo Montejo Ráez	Universidad de Jaén
Andrés Montoyo	Universidad de Alicante
Paloma Moreda	Universidad de Alicante
Rafael Muñoz	Universidad de Alicante
Alicia Pérez Ramírez	Universidad del País Vasco
Enrique Puertas	Universidad Europea de Madrid
Estela Saquete	Universidad de Alicante
Miguel Anxo Solla Portela	Universidad de Vigo

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

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 55 de la revista de la Sociedad Española para el Procesamiento del Lenguaje Natural contiene tres apartados: comunicaciones científicas, resúmenes de proyectos de investigación y descripciones de herramientas (demostraciones). 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 45 trabajos para este número de los cuales 31 eran artículos científicos y 14 correspondían a resúmenes de proyectos de investigación y descripciones de herramientas. De entre los 31 artículos recibidos 16 han sido finalmente seleccionados para su publicación, lo cual fija una tasa de aceptación del 51,6%. Autores de otros 7 países han participado en los trabajos publicados en la revista. Estos países son: Portugal, Holanda, UK, Irlanda, Arabia Saudi Croacia y Grecia.

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 4 sobre 7, y exista al menos dos recomendaciones de aceptación.

Septiembre de 2015
Los editores

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 55th issue of the *Procesamiento del Lenguaje Natural* journal contains scientific papers, investigation projects and tools descriptions summaries. All of these were accepted by the traditional peer reviewed process. We would like to thank the Advisory Committee members and additional reviewers for their work.

Forty-five papers were submitted for this issue of which thirty-one were scientific papers and fourteen were either projects or tool description summaries. From these thirty-one papers, we selected sixteen (51.6%) for publication. Authors from other seven countries have submitted papers to the journal. These countries are: Portugal, The Netherlands, UK, Ireland, Saudi Arabia, Croatia and Greece.

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, as long as this has been equal to or higher than 4 out of 7, and having at least two "accept" recommendations.

September 2015
Editorial board

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

Proyectos

Demostraciones

Información General

Artículos

Semántica, pragmática y discurso. Resolución de la ambigüedad léxica.

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 15-22 recibido 26-03-2015 revisado 27-04-2015 aceptado 14-05-2015

Topic Modeling and Word Sense Disambiguation on the Ancora corpus

Modelado de Categorías y Desambiguación del Sentido de las Palabras en el corpus Ancora

Rubén Izquierdo **Marten Postma** **Piek Vossen**
VU University of Amsterdam, The Netherlands
{ruben.izquierdobevia, m.c.postma, piek.vossen}@vu.nl

Resumen: En este artículo se presenta una aproximación a la Desambiguación del Sentido de las Palabras basada en Modelado de Categorías (LDA). Nuestra aproximación consiste en dos pasos diferenciados, donde primero un clasificador binario se ejecuta para decidir si la heurística del sentido más frecuente se debe aplicar, y posteriormente otro clasificador se encarga del resto de sentidos donde esta heurística no corresponde. Se ha realizado una evaluación exhaustiva en el corpus en español Ancora, para analizar el funcionamiento de nuestro sistema de dos pasos y el impacto del contexto y de diferentes parámetros en dicho sistema. Nuestro mejor experimento alcanza un acierto de 74.53, lo cual es 6 puntos superior al *baseline* más alto. Todo el software desarrollado para estos experimentos se ha puesto disponible libremente para permitir la reprodubilidad de los experimentos y la reutilización del software
Palabras clave: Modelado de categorías, LDA, Sentido más frecuente, WSD, corpus Ancora

Abstract: In this paper we present an approach to Word Sense Disambiguation based on Topic Modeling (LDA). Our approach consists of two different steps, where first a binary classifier is applied to decide whether the most frequent sense applies or not, and then another classifier deals with the non most frequent sense cases. An exhaustive evaluation is performed on the Spanish corpus Ancora, to analyze the performance of our two–step system and the impact of the context and the different parameters in the system. Our best experiment reaches an accuracy of 74.53, which is 6 points over the highest baseline. All the software developed for these experiments has been made freely available, to enable reproducibility and allow the re–usage of the software.
Keywords: Topic Modeling, LDA, Most Frequent Sense, WSD, Ancora corpus

1 Introduction

Word Sense Disambiguation (WSD) is a well–known task within the Natural Language Processing (NLP) field which consists of assigning the proper meaning to a word in a certain context. A very large number of works and approaches have addressed this task from different perspectives in the last decades. Despite all this effort, the task is considered to be still unsolved, and the performance achieved is not comparable to other tasks such as PoS–tagging (with an accuracy around 98%). This is especially problematic if we consider that sense information is used in almost all the high levels NLP tasks (event extraction, NER...). An extensive description of the WSD task and their approaches can be found in Agirre and Edmonds (2007).

Lately, more and more WSD unsupervised approaches have been exploiting, with a reasonable performance under some circumstances, the large resources that are becoming available. Nevertheless, the most widely applied techniques to WSD have been those based on supervised Machine Learning. These approaches tackle WSD as a classification problem, where the goal is to pick the best sense from a predefined list of possible values for a word in a given context, being WordNet (Fellbaum, 1998) the main sense repository selected.

Traditionally, one Machine Learning algorithm is selected (SVM, MaxEnt...), and local and topical features are used to represent the training examples and induce the models.

ISSN 1135-5948

Nevertheless, the size of the context considered to model the problem is usually quite narrow (quite often not more than one sentence), and this may not be sufficient in some cases. Little attention has been paid to consider the role of broader contexts, such as the whole document, or even background information that could be found in external resources and it is not implicit in the document.

The most frequent sense (MFS) heuristic has been extensively used as a baseline for comparison and evaluation. This heuristic has turned to be very difficult to beat by any WSD system. Indeed, we think that in many cases the systems are too skewed towards assigning the MFS, and they do not address properly the problem, specially in the cases where the MFS does not apply. In this direction, we performed an error analysis on the previous SensEval/SemEval evaluations (Izquierdo, Postma, and Vossen, 2015). We found that the participant systems perform very well when the MFS is the correct sense (68% in average in SensEval2, 78% in SensEval3 or 80% in SemEval-2013), but the performance dramatically goes down when the correct label is not the MFS (20% for SensEval2, 18% for SensEval3 or 22% for SemEval2013). Besides to this, we found that, considering the SensEval–2 test dataset, when the correct sense is not the MFS (799 cases), in the 84% of the cases the systems still pick the MFS, which shows clearly the bias towards assigning the MFS that was mentioned before.

In this paper we propose to use topic modeling to perform WSD in the Ancora corpus (Taulé, Martí, and Recasens, 2008), which is a multilevel annotated corpus for Catalan and Spanish, and from which we will make use of the sense annotations for Spanish. Topic modeling is a statistical approach within the Machine Learning field, that tries to discover automatically what are the main topics for a given document or text. We will exploit this technique to create a supervised WSD system that automatically learns the topics related with different senses of a target word and uses these topics to select the proper sense for a new unknown word. The impact of the context and the number of topics on the performance of the WSD system will be also explored. Besides to this, the phenomenon of the most frequent sense will be analyzed and considered as an individual step in the entire WSD problem. To our knowledge there is no other work presenting such an analysis based on topic modeling for Spanish. With the experimentation presented in this paper, an improvement around 6 points in accuracy is obtained over the most frequent sense baseline. All the software developed and the data used for these experiment has been made freely available at `http://kyoto.let.vu.nl/lda_wsd_sepln2015` enabling the reproducibility of these experiments as well as the reuse of the code and data created by the NLP community.

Section 2 will introduce some works applying topic modeling to perform WSD. Then section 3 will present our system architecture. The evaluation framework will be introduced in section 4. Finally the results will be presented in section 5 and some conclusions and future work will be drawn in section 6.

2 Related work

Latent Dirichlet Analysis (LDA) and Topic Modeling in general have been largely applied in NLP tasks, mainly in document classification, topic classification and information retrieval. In these areas, the strong relation between the definition and objective of the task and the application and relevance of topics is quite obvious. In addition, Topic modeling has been also applied in some works to perform Word Sense Disambiguation, which is the main focus of our paper. For instance in Cai, Lee, and Teh (2007), LDA is applied to extract topic features from a large unlabeled corpus. These features are fed into a Naïve Bayes classifier, together with traditional features (part-of-speech, bag-of-words, local collocations...). They perform the evaluation on the SensEval-3 corpora, showing a significant improvement with the use of the topic features. Also in Boyd-Graber and Blei (2007) the authors extend the predominant sense algorithm presented in McCarthy et al. (2004) to create an unsupervised approach for WSD. The topics obtained via LDA are used to calculate similarity measures and pre-dictions for each word in the document, also considering frequencies and features from the surrounding words.

In Li, Roth, and Sporleder (2010) the task of WSD is approached by selecting the best sense based on the conditional probability of sense paraphrases given a context. Two mod-

els are proposed for WSD. One requires prior knowledge of the conditional probability of senses; the second one uses the cosine similarity of two topic-document vectors (sense and context). They prove to get good results (comparable to state-of-the-art) when evaluating at different granularity levels on several SemEval and SenseEval datasets.

The structure of WordNet is exploited in another unsupervised approach presented by Boyd-Graber, Blei, and Zhu (2007). WordNet senses are incorporated as additional latent variables. Each topic is associated not just with simple words, but with a random walk through the WordNet hierarchy. Top-ics and synsets are generated together. An improvement is obtained in some cases, but in some other cases the structure of WordNet affects the accuracy of the system.

Topics and topic modeling have been extensively applied to word sense induction. For instance in (Brody and Lapata, 2009) sense induction is performed as a Bayesian problem by modeling the contexts of the ambiguous word as samples from a multinomial distribution over senses which are in turn characterized as distributions over words. Other works facing word sense induction from a topic modeling point of view are (Wang et al., 2015) or (Knopp, Völker, and Ponzetto, 2013).

3 Our WSD approach

Our WSD system[1] is a supervised machine learning framework based on topic modeling, in particular Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan, 2003). LDA is one of the algorithms for topic modeling that has shown a higher performance and some advantages compared to others such as Latent Semantic Indexing (LSI) or Random Indexing (RI). In our case we have used the LDA implementation available in the Gensim python library[2]. The main idea is to induce a topic model for every sense of every polysemous word based on token features within a certain context. Giving a new word (on the tagging or evaluation phase), we will pick the sense that maximizes the similarity of the feature document created for the new word with each of the models induced for every sense of the same lemma (in the training phase). The features used for representing one target example are the bag-of-words (token based) within a certain number of sentences around the target word. These classifiers assign the proper sense for a target word, and we will refer to them as *sense–LDA* classifiers.

In order to analyze what is the effect of the most frequent sense phenomenon (MFS) in our WSD system, we isolate the problem by considering two different steps in the classification task for a specific case:

1. Decide if the MFS applies in this case
2. If it applies, the MFS is selected
3. Otherwise, the sense returned by the *sense–LDA* is selected[3]

This means that basically a binary classifier is applied first (the MFS classifier) to decide if the most frequent sense applies in this case or not, and the second classifier (sense–LDA) is only queried in the cases where the MFS classifier does not apply. In fact the two tasks could be quite different in nature. On the one hand, deciding if the MFS applies can depend on clues found in larger contexts, related to the topics of the document or even derived from external knowledge sources. On the other hand, learning the topics for less frequent senses could rely on different types of information, linked to more specific and small contexts. Tackling both tasks in one step would not allow to specialize and exploit the proper information for each task. The classifiers derived for deciding on the MFS are based also in LDA and will be named *mfs–LDA* classifiers. The features in this case are the same bag-of-words on larger contexts.

4 Evaluation framework

For the evaluation of our WSD system we performed a folded–cross validation (3-FCV) on the Ancora corpus[4]. We first converted the Ancora corpus to NAF[5] format, as it is the format used by all the tools and linguistic processors developed in our group. Then the folds for training and evaluation were created for every lemma in the Ancora corpus,

[1] As stated previously, all the software and data used for these experiments can be found at `http://kyoto.let.vu.nl/lda_wsd_sepln2015`

[2] `http://radimrehurek.com/gensim/`

[3] The MFS can not be selected anymore in this step

[4] The folds created for our evaluation are available at `http://kyoto.let.vu.nl/lda_wsd_sepln2015/data/`

[5] `http://www.newsreader-project.eu/files/2013/01/techreport.pdf`

making sure to keep the sense distribution in every fold for a fair evaluation.

Our evaluation has been focused only to the polysemous lemmas, and, from these, just in those with at least three manually annotated instances on all the corpus (otherwise the 3-FCV is not possible). There are a total of 7119 unique lemmas annotated in the Ancora corpus. Out of these, 4907 (almost 69%) are monosemous (or annotated just with one sense). From the remaining 31% polysemous[6], 589 lemmas fulfill the requirement of having at least three annotated instances per sense. This set of 589 lemmas compose our evaluation set.

For obtaining the evaluation figures, we use the traditional precision, recall and F-score, micro-averaging across the three folds to get the figures per lemma, and micro–averaging over all the lemmas to get the overall performance of the system. As the total coverage is 100% (the system always provides an answer for every test instance), precision, recall and F–score have the same value and we will refer to this value as *accuracy*. All the lemma output files for the the different experiments presented in next section can be found at `http://kyoto.let.vu.nl/lda_wsd_sepln2015/data/`.

5 Results

In this section the figures obtained by our WSD system for different configurations are shown. As we explained previously, we focus on the polysemous lemmas annotated with at least three instances for each sense (589 lemmas in total). All the results shown in this section refer to that set of 589 lemmas. In order to establish a reference for comparison, three baselines on the Ancora corpus have been derived following different heuristics:

- *Random*: selecting a random sense in each case
- *MFS–overall*: the well–known most frequent sense baseline considering the whole corpus to obtain the sense distribution
- *MFS–folded*: the most frequent sense heuristic using the evaluation folds to calculate the MFS

The *MFS–folded* baseline establishes a better comparison for our WSD system, as the information available for both is exactly the same. In table 1 we can see the figures for these baselines.

Exp	Accuracy
Random	40.10
MFS–overall	67.68
MFS–folded	68.63

Table 1: Baselines on the Ancora corpus

Both MFS baselines are quite high, as expected *a priori* and similarly to the same heuristics calculated for other languages and other sense annotated corpora.

Our first experiment evaluates the behavior of our WSD system when the disambiguation process is done just in one step by the *sense–LDA* classifier (so no MFS classifier is involved). As mentioned previously, one topic model is induced for every sense of each lemma (regardless the MFS or non MFS cases), and the classifier picks the sense that maximizes the similarity of a test instance against the possible sense models. In all our experiments involving LDA classifiers, there are two main parameters that can play a crucial role and that will be analyzed:

- Sentence size: number of sentences considered around a target word to extract the bag-of-word features. The possible values for this parameter are 0, 3 or 50 (where 0 means only the same sentence where the target word is contained). With these values we aim to examine what is the impact of three different context sizes (small, medium and large) on the topic induction task.
- Number of topics: the number of topics set to build the LDA models. In this case this parameter can take the values 3, 10 or 100, which represent three different levels of abstraction.

Combining these three values for the sentence window with the three values for the number of topics we obtain nine possible experiments. The results of these parameter combinations in our first experiment (just *sense–LDA* classifiers) can be seen in Table 2

[6] There 1318 with 2 senses, 449 with 3, 227 with 4, 110 with 5, 41 with 6, 38 with 7, 11 with 8, 10 with 9, 5 with 10 senses, 2 lemmas with 11 senses and one lemma with 12 senses

(*MFSfolded* baseline is included for easy comparison).

Sentences	Topics	Accuracy
MFSfolded	-	68.63
0	3	67.54
	10	65.56
	100	58.34
3	3	66.30
	10	64.62
	100	60.07
50	3	66.04
	10	63.42
	100	59.06

Table 2: Results the sense– classifiers (no MFS–classifier)

As can be seen, the *sense–LDA* classifier is not able to reach the *MFSfolded* baseline in any case. This could mean that indeed considering the task in just one step (with no MFS specialization) makes it very difficult for the LDA models to induce the correct topics. The best results are obtained by considering only the same sentence of the target word to get the features and three as the number of topics for LDA (67.54%). It seems that in this task, the most informative clues are to be found in near contexts. Besides to this, apparently there is certain relation between the two parameters. For instance, the result for $\{sentences = 0; topics = 100\}$ is 58.34 while the result for the same number of topics with $\{sentences = 10\}$ is 59.06, which could imply than for modeling broader contexts (with a larger number of tokens and features), a higher number of topics is required in order to get good results.

The second experiment consists in evaluating our two–steps approach, chaining together the *mfs–LDA* and the *sense–LDA* classifiers. Before doing this, we will evaluate which would be the performance of the whole WSD system if we could use a perfect *mfs–LDA* classifier. In order to simulate this, all the test instances where the correct label is the MFS are considered to be classified correctly, and the rest of instances are classified automatically by the *sense–LDA* classifier (this classifier does not assign the MFS in any case). In other words, this evaluation will examine the performance of the *sense–LDA* classifier on just the non–MFS instances. The results are shown in Table 3.

Sentences	Topics	Accuracy
MFSfolded	-	68.63
0	3	92.48
	10	92.12
	100	90.5
3	3	92.45
	10	92.11
	100	91.60
50	3	92.41
	10	92.12
	100	91.43

Table 3: Results of the WSD system with 2 steps: perfect *mfs–LDA* and automatic *sense–LDA*

The figures in this case are extremely high. This indicates that the *sense–LDA* is able to classify the non MFS cases with a high accuracy, which reinforces our idea of separating both tasks. The conclusions drawn about the combinations of number of sentences and topics are the same as in the previous experiment with only the *sense–LDA* classifier.

The next two experiments will show the evaluation of the two–steps WSD framework with both *mfs–LDA* and *sense–LDA* classifiers induced automatically. Two tables will be shown, the first one for a context of 5 sentences to build the *mfs–LDA* classifier, and the second one using 50 sentences instead.

The first table is Table 4. Each row presents the result for a certain combination of the sentence window and number of topics parameters for the *sense–LDA* classifiers. The last two columns represent the accuracy for different settings of the *mfs–LDA* classifier. In this table the number of sentences for the *mfs–LDA* classifier is set to 5, and there are two experiments for different values of the number of topics: 100 (column *"MFS s5 t100"*) and 1000 (column "*MFS s5 t1000*").

As derived from the table, in the all the cases using the *mfs–LDA* with options $\{sentences = 5; topics = 100\}$, the results are higher than the baseline. In concrete, the best experiment correspond to the *sense–LDA* with option $\{sentences = 0; topics = 3\}$ (74.53), with an improvement around 6 points over the Apparently, using a context of 5 sentences, 100 topics are more informative than 1000 to represent the main features that characterize the most frequent sense. Analyzing the different *sense–LDA* experiments

Sentences	Topics	MFS s5 t100	MFS s5 t1000
MFSfolded	-	68.63	68.63
0	3	74.53	66.73
	10	74.00	66.41
	100	72.61	64.91
3	3	74.30	66.61
	10	73.87	66.36
	100	73.39	65.76
50	3	74.26	66.48
	10	73.90	66.24
	100	73.53	65.75

Table 4: Results for different *sense–LDA* classifier with *mfs–LDA* (S=5 T=100) and *mfs–LDA* (S=5 T=1000)

Sentences	Topics	MFS s50 t100	MFS s50 t1000
MFSfolded	-	68.63	68.63
0	3	73.34	67.15
	10	72.92	66.76
	100	71.43	65.13
3	3	73.21	67.02
	10	72.88	66.60
	100	72.40	66.24
50	3	73.21	66.95
	10	72.83	66.58
	100	72.15	66.20

Table 5: Results for different *sense–LDA* classifier with *mfs–LDA* (S=50 T=100) and with *mfs–LDA* (S=50 T=1000)

in all the cases (the same sentence, 3 or 50 sentences) the results are quite similar. This would indicate that the most rich information to disambiguate the no MFS cases is to be found in local contexts (as the contexts of 3 and 50 sentences already include the smaller context of just the same sentence where the target words are found). Finally about what is the best number of topics to build the *sense–LDA* classifiers, the best performance is reached by using just 3 topics, indicating that larger number of topics may just introduce noise and no relevant information to the disambiguation process.

Following with the experimentation, Table 5 shows the same evaluation as in the previous table, but in this case the context used to build the *mfs–LDA* classifiers is 50 sentences. Similarly, there are two columns with the accuracy of the whole system when 100 or 1000 topics are selected to build the *mfs–LDA* classifiers.

The analysis of this table is similar to the previous one (with only 5 sentences used to generate the *mfs–LDA* classifiers). Comparing both tables, in this case the performance is a bit lower. This might point out that the clues for learning when the MFS applies or not are found in medium sizes contexts (at least for the simple bag–of–words feature model that is being used). Regarding the number of sentences or topics used to build the *sense–LDA* in this experiment, the behavior is the same as in the previous table with a context of 5 sentences.

Finally, we have evaluated individually the *mfs–LDA* classifiers on the task of deciding when the MFS applies or not. In next table, Table 6, the performance of the *mfs–LDA* classifiers for different settings of the parameters (*Sents* for the number of sentences considered as context and *Tops* as the number of Topics) is shown. Specifically, we show the accuracy on predicting the MFS cases, as these cases are those that can affect the overall performance of our system.

Sents. \\Tops.	**100**	**1000**
5	74.41	62.82
50	72.17	62.93

Table 6: Evaluation of the *mfs–LDA* classifiers on detecting the MFS cases

The results endorse the conclusions drawn from the previous experiments. The *mfs–LDA* classifier obtains a better performance by considering 100 topics to induce the models. Furthermore, and as expected, a *mfs–LDA* classifier with a better performance leads to a better overall accuracy when integrated in the two steps (*mfs–LDA* + *sense–LDA*) WSD system.

5.1 Lemma comparison

In this section we will compare the best of our experiments[7], with an accuracy of 74.53 with the baseline (68.63) at lemma level. Out of

[7] *mfs–LDA* with $\{sentences = 5; topics = 100\}$ and *sense–LDA* with $\{sentences = 0; topics = 3\}$

the 589 lemmas evaluated (those lemmas that are polysemous and at least with 3 senses annotated for each sense), a total of 399 (67.7%) lemmas were improved by our best run over the baseline, 126 were under the baseline (21.4%) and for 64 (10.9%) the accuracy of our system and the baseline was equal. In Table 7 we can see the top 5 lemmas with the highest improvement over the baseline. The columns *MFS* and *LDA* represent the accuracy for the MFS baseline and for our LDA system, the column *Var.* shows the variation of our system with respect to the baseline and the last column (*#S*) contains the number of senses of the lemma.

lemma	MFS	LDA	Var.	#S
castigo	50	100	+50	2
ética	50	100	+50	2
veto	50	100	+50	2
mediación	50	100	+50	2
rebeldía	50	100	+50	2

Table 7: Lemmas with highest improvement

We can see that in all the five cases, the number of senses of these lemmas is two. This makes sense with our two–step approach, so if the *mfs–LDA* detects correctly the MFS cases, the rest of the cases become monosemous for the *sense–LDA* classifier. In next table, Table 8, we show the variation in accuracy of our system compared to the MFS baseline for the top 10 lemmas with the highest number of annotations in Ancora (the number of annotations for the lemma is presented in the column *#A.*).

lemma	MFS	LDA	Var.	#A.
año	89.15	91.19	2.04	1275
país	72.29	83.55	11.26	695
presidente	70.31	73.94	3.63	690
partido	55.87	64.48	8.61	641
equipo	98.32	98.88	0.56	539
mes	54.29	80	25.71	315
hora	61.39	56.11	-5.28	305
caso	61.05	91.58	30.53	286
mundo	47.31	40.14	-7.17	279
semana	85.06	92.34	7.28	263

Table 8: Improvement on the most frequent lemmas

In this case we can see a general positive effect, mainly with improvement over the baseline. These lemmas with a large number of annotations are those that can be most affected by the MFS bias. The improvement in these cases could show the robustness of our two–step WSD system. Finally we include in Table 9 those lemmas where our LDA systems presents the largest decrease with respect to the MFS baseline.

lemma	MFS	LDA	Var.
colisión	66.67	33.33	-33.34
filosofía	66.67	33.33	-33.34
garantía	60	26.67	-33.33
prestigio	50	16.67	-33.33
congreso	56.25	27.08	-29.17

Table 9: Lemmas with highest reduction of accuracy

In the majority of these cases, the number of senses is around 2 or 3. This would indicate that either the *mfs–LDA* is not been modeling the context properly in these cases, or that the non most frequent senses for these lemmas are problematic and difficult to disambiguate (which is pointed out too by the discrete results of the MFS baseline in these cases).

6 Conclusions

In this paper we have presented an approach for WSD based on Topic Modeling (LDA), and it has been evaluated on the Ancora Spanish corpus. The whole WSD task is split into two tasks: when the most frequent sense heuristic applies and when it does not. These subtasks have different nature and they might need to be approached in different steps. Our WSD system implements a two–step approach, where first a classifier is applied to decide whether or not the most frequent sense heuristic should be applied. In the cases where this heuristic does not correspond, a traditional sense classifier is employed to return the proper sense.

An exhaustive evaluation of our system has been performed following fold–cross validation on the Ancora corpus, in order to analyze all the different parameters that can play a role in our system. We have found that the best run reaches an accuracy of 74.53 by using the two step system, which is 6 points better than the most frequent sense baseline (68.63). In general, it seems that the best clues for deciding on the most frequent sense

are to be found on contexts around 50 sentences (medium sized) and using 100 topics to induce the models. For the traditional sense classifier, the best models are induced using few topics (3) within small sentence windows (just the sentences where the training instances occur).

All the code and software developed for these experiments, as well as the evaluation data and experiment outputs, can be found freely available at `http://kyoto.let.vu.nl/lda_wsd_sepln2015`. This will enable the reproduction of our experiments as well as the reuse of our programs for further research within the NLP community. Also the data used in these experiments and the output files produced are available.

As future work, we plan to incorporate external knowledge through the detection of named entities and their links to DBpedia in the whole process to enrich the classifiers. Some experiments have been already conducted in this direction with promising results, but some analysis are further experiments are still required. Furthermore, we will carry on a similar evaluation for other languages, starting with English, to reproduce our experiments and analyze our approach in other resources.

References

Agirre, E. and P. Edmonds. 2007. *Word Sense Disambiguation: Algorithms and Applications*. Springer Publishing Company, Incorporated, 1st edition.

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

Boyd-Graber, J. and D. Blei. 2007. Putop: Turning predominant senses into a topic model for word sense disambigua-tion. In *Proceedings of the Fourth International Workshop SemEval-2007*, pages 277–281. Association for Computational Linguistics.

Boyd-Graber, J. L., D. M. Blei, and X. Zhu. 2007. A topic model for word sense disambiguation. In *EMNLP*, pages 1024–1033. ACL.

Brody, S. and M. Lapata. 2009. Bayesian word sense induction. In *Pro-ceedings of the 12th EACL Conference*, EACL '09, pages 103–111, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cai, J., W. Sun Lee, and Y. Whye Teh. 2007. Improving word sense disambiguation using topic features. In *Proceedings of the 2007 Joint EMNLP and CoNLL conferences)*, pages 1015–1023.

Fellbaum, C., editor. 1998. *WordNet: an electronic lexical database*. MIT Press.

Izquierdo, R., M. Postma, and P. Vossen. 2015. Error analysis of word sense disambiguation. In *Proceedings of the Computational Linguistics in The Netherlands (CLIN)*, volume 25.

Knopp, J., J. Völker, and S. P. Ponzetto. 2013. Topic modeling for word sense induction. In Iryna Gurevych, Chris Biemann, and Torsten Zesch, editors, *Language Processing and Knowledge in the Web*, volume 8105 of *Lecture Notes in Computer Science*, pages 97–103. Springer Berlin Heidelberg.

Li, L., B. Roth, and C. Sporleder. 2010. Topic models for word sense disambiguation and token-based idiom detection. In *Proceedings of the 48th ACL conference*, ACL '10, pages 1138–1147, Stroudsburg, PA, USA. Association for Computational Linguistics.

McCarthy, D., R. Koeling, J. Weeds, and J. Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42nd ACL conference*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Taulé, M., M. A. Martí, and M. Recasens. 2008. Ancora: Multi-level annotated corpora for catalan and spanish. In Nicoletta Calzolari et al., edi-tor, *Proceedings of the Sixth International LREC*. European Language Resources As-sociation (ELRA).

Wang, J., M. Bansal, K. Gimpel, B. Ziebart, and C. Yu. 2015. A sense-topic model for word sense induction with unsupervised data enrichment. *Transactions of the Association for Computational Linguistics*, 3:59–71.

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 23-30 recibido 27-03-2015 revisado 28-04-2015 aceptado 19-05-2015

Coreference Resolution for Morphologically Rich Languages.
Adaptation of the Stanford System to Basque*

Resolución de coreferencia para lenguajes morfológicamente ricas.
Adaptación del sistema de Stanford al euskera

Ander Soraluze, Olatz Arregi, Xabier Arregi and Arantza Díaz de Ilarraza
Ixa Group. University of the Basque Country
{ander.soraluze, olatz.arregi, xabier.arregi, a.diazdeilarraza}@ehu.eus

Resumen: Este artículo presenta el proceso de adaptación del sistema de resolución de coreferencia de Stanford para el euskera, un idioma aglutinante, de núcleo final y *pro-drop*. Este sistema ha sido integrado en una cadena de análisis lingüística de manera que recibe como entrada textos procesados y analizados para el euskera. Hemos demostrado que haciendo uso de las características lingüísticas del lenguaje se puede mejorar la resolución de la coreferencia. En el caso de los lenguajes aglutinantes el uso de características morfosintácticas mejora claramente el rendimiento del sistema obteniéondose un incremento en CoNLL F_1 de 5 puntos para el caso de menciones automáticas y de 7,87 puntos con menciones *gold*.
Palabras clave: coreferencia, euskera, lenguaje aglutinante

Abstract: This paper presents the adaptation of the Stanford coreference resolution system to Basque, an agglutinative head-final pro-drop language. The adapted system has been integrated into a global linguistic analysis pipeline so that the input of the system are original Basque raw texts linguistically processed, and annotated. We demonstrate that language-specific characteristics have a noteworthy effect on coreference resolution. In the case of agglutinative languages the use of morphosyntactic features improves substantially the system's performance, obtaining a gain in CoNLL F_1 results of 5 points when automatic mentions are used and of 7.87 points when gold mentions are provided.
Keywords: coreference, Basque, agglutinative language

1 Introduction

Coreference resolution consists of identifying textual expressions (mentions) that refer to real-world objects (entities) and determining which of these mentions refer to the same entity. It is well known that coreference resolution is helpful in NLP applications where a higher level of comprehension of the discourse leads to better performance. Information Extraction, Question Answering, Machine Translation, Sentiment Analysis, Machine Reading, Text Summarization, and Text Simplification, among others, can benefit from coreference resolution.

In this paper we present the adaptation of the Stanford Deterministic Coreference Resolution System (henceforth SDCRS) (Lee et al., 2013) to resolve coreferences in Basque written texts. The SDCRS applies a succession of ten independent deterministic coreference models (or "sieves") to resolve coreference in written texts. During the adaptation process, firstly, we have created a baseline system which receives as input texts processed by Basque analysis tools and uses specifically adapted static lists to identify language dependent features. Afterwards, improvements over the baseline system have been applied, adapting and replacing some of the original sieves. Our final goal is to create a robust end-to-end coreference resolution system for the Basque language.

Basque is a non-Indo-European language and differs considerably in grammar from the languages spoken in its surroundings. It is, indeed, an agglutinative head-final pro-drop isolated language. Furthermore, Basque is a free word order language; this means

* This work has been supported by Ander Soraluze's PhD grant from Euskara Errektoreordetza, the University of the Basque Country (UPV/EHU) and by the Ber2Tek project, Basque Government (IE12-333).

ISSN 1135-5948

that the order of phrases in the sentence can vary (Laka, 1996). The rich morphology of Basque requires that one takes into account the structure of words (morphological analysis) to improve coreference resolution results.

This paper is structured as follows. After reviewing related work in section 2, we describe the adaptation of the system for Basque in section 3. Section 4 presents the effects of considering some morphosyntactic characteristics of Basque and the improvements obtained with respect to the baseline system. The main experimental results are outlined in section 5 and discussed in section 6. Finally, we review the main conclusions and preview future work.

2 Related Work

The first coreference resolution systems were designed for English; nowadays, however, many conferences focus on multilingual coreference resolution. In CoNLL 2011 (Pradhan et al., 2011), participants had to model unrestricted coreference in the Ontonotes corpus (Pradhan et al., 2007) in English. Only one year later, the CoNLL 2012 shared task (Pradhan et al., 2012) involved predicting coreference in three languages, English, Chinese and Arabic. Participants adapted their systems to resolve coreference in these languages, taking into consideration the special characteristics of each language (Fernandes, dos Santos, and Milidiú, 2012). Björkelund and Farkas (2012) note that while Chinese and English are not morphologically rich languages, Arabic has a very complex morphology and this is why they had to use lemmas and unvocalized Buckwalter forms. Chen and Ng (2012), on the other hand, seek to improve the multi-pass sieve approach by incorporating lexical information using machine learning techniques. They employ different sieves depending on the language.

SemEval-2010 Task 1 (Recasens et al., 2010) was also dedicated to coreference resolution in multiple languages (Catalan, Dutch, English, German, Italian, and Spanish). This shared task addressed open questions like how much language-specific tuning is necessary to implement a general coreference resolution system portable to different languages or how helpful morphology, syntax and semantics are for solving coreference relations. Zhekova and Kübler (2010) suggest that an optimization of the feature set for individual languages should improve system performance. Broscheit et al. (2010) affirm that substantial improvements can be achieved by incorporating language specific information with the *Language Plugin* module of their BART system. The *Language Plugin* provides an effective separation between linguistic and machine learning aspects of the problem of coreference resolution. Recently, Kopeć and Ogrodniczuk (2012) have explained particularly well the process of adapting the BART system to Polish, a less-resourced language.

3 Adapting the Stanford Coreference Resolution System to Basque

The SDCRS applies a succession of ten independent deterministic coreference models (or "sieves") one at time from highest to lowest precision. It makes use of global information through an entity-centric model that encourages the sharing of features across all mentions that point to the same real-world entity. The architecture is highly modular, which means that additional coreference models can easily be integrated. The system gives state-of-the-art performance for English and has also been incorporated into hybrid state-of-the-art coreference systems for Chinese and Arabic.

The module that resolves coreference is used at the end of a pipeline process where raw English written texts are processed. In each step, common linguistic processors (tokenizer, POS tagger, named entity recognizer) are applied to the text, thus obtaining linguistically annotated data.

The coreference resolution module is dependent on the annotations that previous modules make. As the modules were created to process English, we had to adapt the output obtained by Basque linguistic processors in order to create appropriate annotations for the coreference module.

The Basque linguistic processors used to create annotations are the following: i) A morphological analyser that performs word segmentation and PoS tagging (Alegria et al., 1996), ii) A lemmatiser that resolves the ambiguity caused at the previous phase (Alegria et al., 2002), iii) A multi-word item identifier that determines which groups of two or more words are to be considered multi-word expressions (Alegria et al., 2004), iv) A named-

entity recogniser that identifies and classifies named entities (person, organization, location) in the text (Alegria et al., 2003), v) A numerical-entity recogniser that identifies and classifies numerical entities (date, time, percent, number...) in the text (Soraluze et al., 2011), vi) A dependency parser based on Maltparser (Nivre et al., 2007); its output is then used to create constituent trees (Bengoetxea and Gojenola, 2010), and vii) A mention detector that identifies mentions that are potential candidates to be part of coreference chains in Basque written texts (Soraluze et al., 2012).

Apart from the annotations, the coreference resolution module also makes use of some static lists that are organized to exploit relevant features like gender, animacy or number. Pronouns, too, are defined as static lists. These static lists have also been adapted to Basque.

The created baseline system uses the static lists adapted to Basque and the annotations created by Basque linguistic processors. The sieves of the coreference module have not been changed at all. The results obtained by this system are presented in section 5.

4 Improvements over the baseline system

In this section we explain how we modified the baseline system taking advantage of some of the Basque language features to improve the performance of the system.

4.1 The Exact Morphology String Match sieve

Firstly, we observed the need of creating a new sieve to deal with mentions that fulfilled the string match constraint except for some grammatical suffix. This need is closely related to the agglutinative nature of Basque. The new sieve, named Exact Morphology String Match, can be considered a replacement of the original Exact String Match sieve, which links two mentions if they contain the same extent text. The Exact String Match sieve works correctly when mentions are identical. Nevertheless, this constraint is too restrictive in agglutinative languages, since the role of prepositions is played by suffixes added to word forms, for example, *lehendakariarekin* "with the president" and *lehendakariarengana* "to the president". Exact Match String sieve would not link these two mentions because their extents do not match.

In order to treat these cases correctly, the Exact Morphology Match sieve assumes that two mentions are coreferent if i) the lemmas of each word in both mentions are identical, and ii) the number and definiteness are the same. If these conditions are not fulfilled, the mentions are not considered coreferent.

The examples in Table 1 illustrate the suitability of the Exact Morphology String Match. We can see that first mention *txori politak*, and the second one, *txori politekin*, are coreferent because their lemmas are identical and they satisfy the same number and same definiteness condition. Nevertheless, although the first and third mention are identical, they are not coreferent. The first mention *Txori politak* represents a plural mention in the absolutive case, and the same string in the third row corresponds with a mention in the singular ergative case (obviously this morphological information has been previously extracted by attending to the context). Finally, the first and fourth mentions have the same lemma and number but their definiteness differs (the first is definite while the second is indefinite), so they can not be considered coreferent.

4.2 The Relaxed String Match sieve

The SDCRS has a special sieve to treat relative clauses, called Relaxed String Match. This sieve considers two mentions coreferent if their strings are identical when the text of a relative clause following the head word is dropped, e.g., "Bush" and "Bush, who was president of the U.S.". In English relative clauses always follow the noun, but in Basque relative clauses can either follow or precede the noun. For example, the two possible equivalents in Basque for the sentence "the president, who accepted the new law" are the following: i) noun followed by the relative clause *Presidentea [zeinak lege berria onartu baitu]*$_{REL}$ and ii) noun preceded by the relative clause *[Lege berria onartu duen]*$_{REL}$ *presidentea*. Although, the two cases presented above are correct in Basque, the second one is more common.

To resolve this issue, we have adapted the Relaxed String match sieve to also take into account relative clauses that precede the

#	Mention	Translation	Lemmas	Number	Definiteness	Coreferent
1	txori politak	pretty bird	txori polit	plural	definite	-
2	txori politekin	with the pretty birds	txori polit	plural	definite	yes
3	txori politak	pretty bird	txori polit	singular	definite	no
4	txori politek	pretty birds	txori polit	plural	indefinite	no

Table 1: Examples of mentions that are coreferent and not based on their morphological features

noun. This way, the sieve is able to drop the text or relative clause preceding the head word, and mentions like *presidentea* and *lege berria onartu duen presidentea* can be linked.

4.3 The Strict Head Match sieve and its variants

The String Match sieve links two mentions if the following constraints are fulfilled: i) the candidate mention (mention to consider for resolution) head word matches any head word of mentions in the antecedent entity (*Entity Head Match*); ii) all non-stop words in the current entity to be solved are included in the set of non-stop words in the antecedent entity (*Word Inclusion*); iii) all the mention modifiers (whether nouns or adjectives) of the candidate are included in the modifiers of the antecedent (*Compatible Modifiers only*); in other words this constraint avoids clustering the mentions *autobia zuzena* "the correct motorway" and *autobia okerra* "the wrong motorway"; iv) two mentions are not in an i-within-i construct (*Not-i-within-i*), i.e, one cannot be child NP in the other's NP constituent (Chomsky, 1981).

We have adapted the first constraint and retained the others. In our proposal, the *Entity Head Match* constraint considers for comparison the head word lemmas, number and definiteness instead of the head word forms. In this way mentions like *Vitoria-Gasteizko Eusko Legebiltzarra* "Vitoria-Gasteiz Basque Parliament" and *Vitoria-Gasteizko Legebiltzarretik* "from Vitoria-Gasteiz Parliament" that would not be clustered following the original constraint are linked by means of our new adapted sieve.

In order to improve overall F_1 by improving recall, the following three variants of the Strict Head Match sieve are applied: 1) all the constraints are considered; 2) all the constraints are considered except *Compatible Modifiers Only*; 3) all the constraints are considered except *Word Inclusion*.

As the *Entity Head Match* constraint is applied in all the variants, our adaptation influences all of them.

4.4 The Proper Head Word Match sieve

The Proper Head Word Match sieve considers two mentions coreferent if the following are fulfilled: i) the two mentions are headed by proper nouns and the head is the last word; ii) the two mentions are not in an i-within-i construct; iii) the modifiers of the two mentions cannot have location mismatches; iv) the candidate mention cannot have a number that appears in the antecedent candidate.

The first constraint of this sieve is too restrictive, because Basque is a free word order language and the last word of a mention does not obligatorily have to be the head of a mention. Therefore, we have changed the constraint in the way that the mentions headed by proper nouns can have their heads in any position within a mention. For example, the mention *Frantzia ekialdea* "eastern France" would not be a possible candidate for resolution without any changes in the sieve as the head, the proper noun *Frantzia*, is not the last word.

We also translated the list of location modifiers and the list of written numbers defined inside the sieve from English to Basque.

5 Evaluation

The corpus used to develop and test the system is a part of EPEC (the Reference Corpus for the Processing of Basque) (Aduriz et al., 2006). EPEC is a 300,000 word sample collection of news published in *Euskaldunon Egunkaria*, a Basque language newspaper. The part of the corpus we have used has about 45,000 words and it has been manually tagged by two linguists. First of all, automatically tagged mentions obtained by our mention detector have been corrected; then, coreferent mentions have been linked in clusters.

Decisions about the annotation of singletons differ depending on the corpora. In the corpus used in SemEval-2010 Task 1 (Recasens et al., 2010) all the singletons were annotated, on the contrary, in the Ontonotes

corpus (Pradhan et al., 2007) singletons were not tagged. We decided to annotate all the singletons, although they had not coreference relations in the text. In our opinion singletons are significant for a deep text understanding.

To calculate the agreement between annotators, we used the *Strict Matching* protocol which considers two mentions correct if they are identically the same. Using this protocol we compared the annotations made by the two linguists and obtained an F-measure of 94.07% for agreement. All the annotation process has been carried out using the MMAX2 annotation tool (Müller and Strube, 2006).

We divided the dataset into two main parts: one for developing the system and the other for testing. More detailed information about the two parts can be found in Table 2.

	Words	Mentions	Clusters	Singletons
Devel	30434	8432	1313	4383
Test	15949	4360	621	2445

Table 2: EPEC corpus division information

We tested two systems using the corpus: i) the baseline system (henceforth BS), which is a copy of the original SDCRS taking as input only the output of the Basque linguistic processors and translated static lists, and ii) our improved system, Basque Coreference Resolver (henceforth BCR), which modifies and adds some sieves taking advantage of the morphosyntactic features of Basque.

The metrics used to evaluate the systems are MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), CEAF$_e$ (Luo, 2005), CEAF$_m$ (Luo, 2005), and BLANC (Recasens and Hovy, 2011). The scores have been calculated using the reference implementation of the CoNLL scorer (Pradhan et al., 2014).

Table 3 shows the results obtained by each system in the test set when automatic mentions are used.

We have also evaluated the two systems using the gold mentions, i.e, when providing all the correct mentions to the coreference resolution systems. The scores obtained are shown in Table 4.

6 Discussion

In the case of automatically detected mentions, BCR outperforms the BS according to F_1 on all the metrics. In CoNLL metric, BCR has a score of 53.67 , which is 5 points higher than BS, which scores 48.67.

		R	P	F_1
Mention Detection		73.01	74.86	73.92
MUC	BS	22.48	35.27	27.46
	BCR	36.63	44.34	40.11
B^3	BS	54.81	66.17	59.96
	BCR	58.34	64.08	61.08
$CEAF_m$	BS	56.13	57.6	56.86
	BCR	58.52	60.00	59.25
$CEAF_e$	BS	62.08	55.5	58.61
	BCR	60.99	58.71	59.83
BLANC	BS	33.47	44.96	36.75
	BCR	39.13	47.64	42.44
CONLL	BS	-	-	48.67
	BCR			53.67

Table 3: BS and BCR scores with automatic mentions

		R	P	F_1
Mention Detection		100	100	100
MUC	BS	31.6	43.32	36.55
	BCR	51.54	56.71	54.00
B^3	BS	76.32	86.92	81.28
	BCR	81.61	86.6	84.03
$CEAF_m$	BS	72.13	72.13	72.13
	BCR	76.3	76.3	76.3
$CEAF_e$	BS	80.44	72.11	76.05
	BCR	81.00	77.97	79.46
BLANC	BS	59.47	71.06	62.94
	BCR	67.54	75.56	70.76
CONLL	BS	-	-	64.62
	BCR			72.49

Table 4: BS and BCR scores with gold mentions

In Goenaga et al. (2012) an automatic coreference resolution system for Basque is presented, but unfortunately the results are not comparable with ours. The reasons of not being the works comparable are that the corpus used by the authors to evaluate pronominal anaphora resolution is not the same as the one we used, and the size of the corpora also differs considerably in size. Furthermore, some structures like relative clauses that we consider as mentions are not taken into account in the cited work.

As it is mentioned in Pradhan et al. (2014), where official updated scores of CoNLL 2011 and CoNLL 2012 participants are presented, the best system in 2011 ob-tained 51.5 official score and the worst 15.5 for English. One year later in CONLL 2012, systems obtained better results for English task: the best one scored 60.7 in CoNLL F_1. It is worth to note that the scores from CoNLL 2011 and 2012 are not directly com-

parable with ours, given that neither the language for resolution nor the corpus used are the same.

Comparing with the results for Arabic coreference resolution, a morphologically rich language as Basque, in CoNLL 2012 the best system obtained an score of 45.2. Clearly, the results are lower for Arabic than for English. Chen and Ng (2012), participants in the Arabic coreference resolution task, argue that their low results for Arabic are primar-ily because Arabic is highly inflectional and their knowledge of the language was poor.

German is also a morphologically rich language, but while Basque is an agglutinative language German is considered fusional language. Two system presented in SemEval-2010 Task 1 (Recasens et al., 2010) that resolved coreference for German are SUCRE (Kobdani and Schütze, 2010) and UBIU (Zhekova and Kübler, 2010). SUCRE obtained an score of 55.03 CoNLL F_1 and UBIU 33.93. While SUCRE's results are good enough, UBIU's are quite low considering that the system is described as a language-independent for coreference resolution.

Observing our results we can affirm that the knowledge of the language, such as the morphosyntactic information, benefits considerably the coreference resolution in highly inflectional languages.

Our preliminary results are quite good in all the metrics, taking into account that there is margin of improvement as the full adaptation of the BCR is not yet finished.

When gold mentions are used our system also outperforms the baseline system according to all the metrics. The official CoNLL metric is outperformed by 7.87 points.

It is interesting to compare the effect of the mention detection. When automatic mentions are provided the CoNLL F_1 of BCR is 53.67, while providing gold mentions raises this value to 72.49. There is a considerable improvement, 18.86 points, that shows how important is to obtain a good performance in the mention detection task and how errors in this step can decrease substantially the performance on coreference resolution. Similar ideas on the importance of mention detection are presented in Uryupina (2008) and in Uryupina (2010). The scores obtained when gold mentions are provided also shows that there is still a margin to improve coreference resolution.

7 Conclusions and future work

We have adapted the SDCRS to Basque and integrated it into a global architecture of linguistic processing. Firstly, we have defined a baseline system, and afterwards, improved it based on the principle that morpho-syntactic features are crucial in the design of the sieves for agglutinative languages like Basque. The initial changes consist of the addition of a new sieve, Exact Morphology Match, replacing the original Exact String Match and the modification of Relaxed String Match sieve, Strict Head Match sieve and its variants and Proper Head Word Match sieve. Our system outperforms the baseline system in all the metrics considered. The results obtained in CoNLL metric are quite good, 53.67 when automatic mentions are used and 72.49 with gold mentions, and point that we are in a good direction to obtain a robust coreference resolution for Basque.

In the future, our aim is to analyse the influence of agglutination and free word-order on other sieves, and to implement the necessary adaptations. We also want to adapt the Pronoun Resolution sieve to Basque, taking into account the characteristics of the Basque pronouns. Nowadays, the original sieve ordering of the SDCRS is used in our system, nevertheless, better ordering options could exist. We would like to investigate which ordering of the sieves would be the optimal for Basque. It would also be interesting to carry out a deep qualitative error analysis of the results in order to obtain information about how to improve the recall of the system while preserving the precision.

We expect that the incorporation of new Basque-oriented treatments into the system will improve the overall scores.

References

Aduriz, I., M. Aranzabe, J. M. Arriola, M. Atutxa, A. Díaz de Ilarraza, N. Ezeiza, K. Gojenola, M. Oronoz, A. Soroa, and R. Urizar. 2006. Methodology and Steps towards the Construction of EPEC, a Corpus of Written Basque Tagged at Morphological and Syntactic Levels for the Automatic Processing. Rodopi. Book series: Language and Computers., pages 1–15.

Alegria, I., O. Ansa, X. Artola, N. Ezeiza, K. Gojenola, and R. Urizar. 2004. Representation and Treatment of Multiword

Expressions in Basque. In *ACL workshop on Multiword Expressions*, pages 48–55.

Alegria, I., M. Aranzabe, N. Ezeiza, A. Ezeiza, and R. Urizar. 2002. Using Finite State Technology in Natural Language Processing of Basque. In *Implementation and Application of Automata*, volume 2494 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, pages 1–12.

Alegria, I., X. Artola, K. Sarasola, and M. Urkia. 1996. Automatic Morphological Analysis of Basque. *Literary & Linguistic Computing*, 11(4):193–203.

Alegria, I., N. Ezeiza, I. Fernandez, and R. Urizar. 2003. Named Entity Recognition and Classification for texts in Basque. In *II Jornadas de Tratamiento y Recuperación de Información*, (JOTRI 2003), pages 198–203, Madrid, Spain.

Bagga, A. and B. Baldwin. 1998. Algorithms for Scoring Coreference Chains. In *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566.

Bengoetxea, K. and K. Gojenola. 2010. Application of Different Techniques to Dependency Parsing of Basque. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 31–39, Stroudsburg, PA, USA.

Björkelund, A. and R. Farkas. 2012. Data-driven Multilingual Coreference Resolution Using Resolver Stacking. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 49–55, Jeju Island, Korea, July. Association for Computational Linguistics.

Broscheit, S., M. Poesio, S. P. Ponzetto, K. J. Rodriguez, L. Romano, O. Uryupina, Y. Versley, and R. Zanoli. 2010. BART: A multilingual Anaphora Resolution System. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, (SemEval 2010), pages 104–107, Stroudsburg, PA, USA.

Chen, C. and V. Ng. 2012. Combining the Best of Two Worlds: A Hybrid Approach to Multilingual Coreference Resolution. In *Joint Conference on EMNLP and CoNLL: Proceedings of the Shared Task*, pages 56–63.

Chomsky, N. 1981. *Lectures on Government and Binding*. Studies in generative grammar. Foris publications, Dordrecht, Cinnaminson (R.I.).

Fernandes, E. R., C. N. dos Santos, and R. L. Milidiú. 2012. Latent Structure Perceptron with Feature Induction for Unrestricted Coreference Resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, CoNLL '12, pages 41–48, Stroudsburg, PA, USA. Association for Computational Linguistics.

Goenaga, I., O. Arregi, K. Ceberio, A. Díaz de Ilarraza, and A. Jimeno. 2012. Automatic Coreference Annotation in Basque. In *11th International Workshop on Treebanks and Linguistic Theories*, Lisbon, Portugal.

Kobdani, H. and H. Schütze. 2010. SUCRE: A Modular System for Coreference Resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, (SemEval 2010), pages 92–95, Stroudsburg, PA, USA.

Kopeć, M. and M. Ogrodniczuk. 2012. Creating a Coreference Resolution System for Polish. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Laka, I. 1996. A Brief Grammar of Euskara, the Basque Language. http://www.ehu.es/grammar. University of the Basque Country.

Lee, H., A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky. 2013. Deterministic Coreference Resolution Based on Entity-centric, Precision-ranked Rules. *Compututational Linguistics*, 39(4):885–916.

Luo, X. 2005. On Coreference Resolution Performance Metrics. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Müller, C. and M. Strube. 2006. Multi-level Annotation of Linguistic Data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Peter Lang, Frankfurt a.M., Germany, pages 197–214.

Nivre, J., J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, and E. Marsi. 2007. MaltParser: A language-independent System for Data-driven Dependency Parsing. *Natural Language Engineering*, 13(2):95–135.

Pradhan, S., E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2007. OntoNotes: A Unified Relational Semantic Representation. In *Proceedings of the International Conference on Semantic Computing*, (ICSC 2007), pages 517–526, Washington, DC, USA. IEEE Computer Society.

Pradhan, S., X. Luo, M. Recasens, E. Hovy, V. Ng, and M. Strube. 2014. Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35. Association for Computational Linguistics.

Pradhan, S., A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012)*, Jeju, Korea.

Pradhan, S., L. Ramshaw, M. Marcus, M. Palmer, R. Weischedel, and N. Xue. 2011. CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, (CoNLL 2011), pages 1–27, Stroudsburg, PA, USA.

Recasens, M. and E. Hovy. 2011. BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Recasens, M., L. Màrquez, E. Sapena, M. A. Martí, M. Taulé, V. Hoste, M. Poesio, and Y. Versley. 2010. SemEval-2010 task 1: Coreference Resolution in Multiple Languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, (SemEval 2010), pages 1–8, Stroudsburg, PA, USA.

Soraluze, A., I. Alegria, O. Ansa, O. Arregi, and X. Arregi. 2011. Recognition and Classification of Numerical Entities in Basque. In *RANLP*, pages 764–769, Hissar, Bulgaria.

Soraluze, A., O. Arregi, X. Arregi, K. Ceberio, and A. Díaz de Ilarraza. 2012. Mention Detection: First Steps in the Development of a Basque Correference Resolution System. In *KONVENS 2012, The 11th Conference on Natural Language Processing*, Vienna, Austria.

Uryupina, O. 2008. Error Analysis for Learning-based Coreference Resolution. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Uryupina, O. 2010. Corry: A System for Coreference Resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 100–103, Uppsala, Sweden, July. Association for Computational Linguistics.

Vilain, M., J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. 1995. A Model-theoretic Coreference Scoring Scheme. In *Proceedings of the 6th Conference on Message Understanding*, MUC6 '95, pages 45–52, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhekova, D. and S. Kübler. 2010. UBIU: A Language-independent System for Coreference Resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, (SemEval 2010), pages 96–99, Stroudsburg, PA, USA.

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 31-38 recibido 27-03-2015 revisado 27-04-2015 aceptado 18-05-2015

Diseño y comparación de varias aproximaciones estadísticas a la comprensión del habla en dos tareas e idiomas distintos*

Design and comparison of several statistical approaches to Speech Understanding in two different tasks and languages

Fernando García, Marcos Calvo, Lluís-F. Hurtado, Emilio Sanchis, Encarna Segarra
Universitat Politècnica de València
Camí de Vera s/n, 46022 València, Spain
{fgarcia,mcalvo,lhurtado,esanchis,esegarra}@dsic.upv.es

Resumen: En este artículo se presenta un estudio de diversas aproximaciones al problema de la comprensión del habla en dominios semánticos restringidos. Se proponen dos sistemas basados en modelos generativos y se comparan con un sistema basado en un método discriminativo. La experimentación se ha realizado sobre dos tareas diferentes, DIHANA y MEDIA, que a su vez están en dos idiomas diferentes. El uso de las dos tareas tiene interés no sólo por las diferencias en la forma de expresar los conceptos en los dos idiomas, sino también por las diferencias en la forma de representar la semántica. Los resultados muestran la capacidad de los modelos estadísticos aprendidos automáticamente para representar la semántica, incluso cuando se trata con voz, que introduce errores generados en el proceso de reconocimiento.
Palabras clave: comprensión del habla, modelos estocásticos, modelos generativos, modelos discriminativos

Abstract: In this paper, a study of different approaches to the problem of speech understanding in restricted semantic domains is presented. Two systems based on generative models are proposed and they are compared with a system based on discriminative methods. The experiments were conducted on two different tasks, DIHANA and MEDIA, which are in two different languages. The use of the two tasks is of interest not only because of the differences in how concepts are expressed in both languages, but also because of the differences in the way of representing the semantics. The results show the ability of automatically learned statistical models to represent the semantics, even when dealing with voice input, which introduces errors that are generated in the recognition process.
Keywords: spoken language understanding, stochastic models, generative models, discriminative models

1 Introducción

La componente de comprensión del habla tiene una importancia capital en muchos de los sistemas de interacción con los ordenadores, bien oral o escrita. Si bien está todavía muy lejos la posibilidad de disponer de componentes que proporcionen una interpretación semántica de un texto en un universo semántico no restringido, su uso para tareas acotadas semánticamente proporciona resultados razonables.

Uno de los ámbitos de aplicación de estos componentes son los sistemas de diálogo hablado para dominios restringidos. En un gran número de sistemas de este tipo se abordan tareas cuyo objetivo final es la obtención de una plantilla con las informaciones necesarias para realizar una consulta a un sistema de información. Esto se realiza a lo largo de diversos turnos, por lo que para cada turno es necesario obtener la información semántica proporcionada por el usuario, es decir, los datos concretos que se han proporcionado y también información sobre la intención subyacente en el turno.

* This work is partially supported by the Spanish MEC under contract TIN2014-54288-C4-3-R and FPU Grant AP2010-4193

ISSN 1135-5948

Entre las aproximaciones desarrolladas para la obtención de los componentes de comprensión del habla podemos hablar de las basadas en reglas que determinan los patrones sintácticos de ciertos significados (Ward y Issar, 1994; Seneff, 1992) y de las basadas en modelos estadísticos (Hahn et al., 2010; Raymond y Riccardi, 2007) que pueden ser estimados a partir de conjuntos de muestras etiquetadas semánticamente. Los modelos estadísticos presentan diversas ventajas. Una de ellas es que representan adecuadamente la secuencialidad de las frases y los conceptos. Disponen de mecanismos para ampliar la cobertura de forma que pueden tratar frases con errores (habituales cuando su entrada es la salida de un reconocedor). En contrapartida necesitan que los corpus de entrenamiento estén segmentados y etiquetados.

Algunos de estos modelos estadísticos son los llamados generativos basados en Modelos de Markov o Gramáticas estocásticas (Servan et al., 2010; Ortega et al., 2010; Hurtado et al., 2004; Esteve et al., 2003; He y Young, 2003; Segarra et al., 2002). También se han utilizado modelos discriminativos, como son clasificadores bayesianos, SVM o CRF (Dinarelli, Moschitti, y Riccardi, 2009; Lefèvre, 2007; Lafferty, McCallum, y Pereira, 2001). En este tipo de modelos uno de los principales problemas que se tiene que abordar es el de la segmentación de la frase de entrada, ya que el objetivo no es solamente obtener una o más clases asociadas a una frase sino también el segmento de texto que se corresponde con cada significado semántico encontrado, considerando el contexto de toda la frase.

En este artículo presentamos tres aproximaciones estadísticas al problema de la comprensión del lenguaje, explorando su comportamiento sobre diferentes corpus de diálogo. Se presentan dos aproximaciones que modelizan la semántica con autómatas finitos, desarrolladas anteriormente para comprensión multilingüe (García et al., 2012; Calvo et al., 2013), y una aproximación que utiliza CRF. Se ha hecho una experimentación en la que se han utilizado dos corpus con tareas e idiomas diferentes. Estas aproximaciones funcionan en tiempo real y se han implementado en un prototipo de comprensión de habla multilingüe (Laguna et al., 2014).

Los corpus utilizados son el corpus DIHANA, en castellano, una tarea de acceso a información sobre trenes, y el corpus MEDIA, en francés, una tarea de información y reserva turística. Estos son dos de los principales corpus etiquetados semánticamente que existen en la actualidad.

2 El corpus DIHANA

La tarea del corpus DIHANA (Benedí et al., 2006) consiste en un sistema de información sobre horarios, precios y servicios de trenes de larga distancia españoles en castellano. El corpus consiste en 900 diálogos que se adquirieron por 225 hablantes empleando la técnica del Mago de Oz, que permite que estén presentes la mayor parte de las características del habla espontánea en la adquisición de los diálogos.

El corpus DIHANA se ha transcrito y etiquetado manualmente a nivel de concepto, para ello se definieron 30 etiquetas semánticas. Estas etiquetas semánticas se pueden agrupar en: independientes de la tarea como "cortesía", "nada" ..., otras cuyo segmento asociado contiene un valor que es relevante para la comprensión p.e. "ciudad_origen", "ciudad_destino", "hora", "tipo_tren" ..., y otras que son relevantes para la tarea e identifican el tipo de concepto del cuál se está hablando: "<hora>", "<hora_salida>", "<hora_llegada>", "<tipo_tren>" ..., (estas últimas vienen parentizadas por las marcas < y > para distinguirlas de las anteriores). A continuación se muestra un ejemplo de representación semántica de la frase: *"me podría decir horarios de tren intercity de Zamora a Valladolid"*.

Ejemplo:
"Me podría decir horarios de tren intercity de Zamora a Valladolid por favor"

Palabras	Concepto
me podría decir	consulta
horarios de tren	<hora>
intercity	tipo_tren
de Zamora	ciudad_origen
a Valladolid	ciudad_destino
por favor	cortesía

Algunas características del corpus DIHANA etiquetado semánticamente se muestra en la Tabla 1.

3 El corpus MEDIA

La tarea del corpus MEDIA (Bonneau-Maynard et al., 2005) consiste en un sistema de información turística y reservas de

Número de turnos de usuario:	6.229
Total de palabras:	47.222
Talla del vocabulario:	811
Media de palabras por turno de usuario:	7,6
Número total de segmentos semánticos:	18.588
Media de palabras por segmento semántico:	2,5
Media de segmentos por turno de usuario:	3,0
Media de muestras por unidad semántica:	599,6

Tabla 1: Características del corpus DIHANA etiquetado semánticamente.

Número de turnos de usuario:	16.279
Total de palabras:	114.969
Talla del vocabulario:	2.357
Media de palabras por turno de usuario:	7,1
Número total de segmentos semánticos:	53.942
Media de palabras por segmento semántico:	2,1
Media de segmentos por turno de usuario:	3,3
Media de muestras por unidad semántica:	709,8

Tabla 2: Características del corpus MEDIA etiquetado semánticamente.

hoteles en francés. Al igual que DIHANA se adquirió empleando la técnica del Mago de Oz. Se definieron ocho categorías de escenarios con diferentes niveles de complejidad.

El corpus adquirido está compuesto por 1.257 diálogos de 250 hablantes y contiene aproximadamente 70 horas de diálogos. Cada hablante grabó cinco escenarios diferentes de reservas de hotel. El corpus francés MEDIA se ha transcrito y etiquetado manualmente a nivel de concepto. Para realizar el etiquetado se definieron más de 80 conceptos semánticos de los cuales solo 72 están presentes en el corpus.

Un ejemplo de etiquetado semántico para la frase: *"je souhaiterais réserver pour la deuxième semaine de janvier à Paris à côté de l' arc de triomphe"* se muestra a continuación.

Ejemplo:
"Je souhaiterais réserver pour la deuxième semaine de janvier à Paris à côté de l' arc de triomphe"

Palabras	Concepto
je souhaiterais réserver	command-tache
pour la deuxième	rang-temps
semaine	temps-unite
de janvier	temps-mois
à Paris	localisation-ville
à côté de	localisation-distanceRelative
l' arc de triomphe	localisation-lieuRelatif

Algunas características del corpus MEDIA etiquetado semánticamente se muestra en la Tabla 2.

4 El sistema de comprensión

El problema de la comprensión del habla puede abordarse como la búsqueda de la secuencia de conceptos que se corresponden con el significado de una frase de entrada. Cada concepto representa el significado de una secuencia de palabras (un segmento) de la frase. Por ejemplo, el concepto "consulta" del corpus DIHANA puede ser asociada con los segmentos "Me podría decir", "Por favor dígame", "Cuál es", etc. De esta forma, el sistema de comprensión es capaz de asociar a cada frase de entrada una secuencia semántica (secuencia de conceptos), así como los segmentos de palabras consecutivas asociados a los conceptos, es decir, la segmentación subyacente.

En la Figura 1 se presenta un esquema del proceso de comprensión, tanto de la fase de entrenamiento como de la fase de test. En la fase de entrenamiento se dispone del corpus de entrenamiento que deberá presentar las frases etiquetadas y segmentadas en términos de conceptos, como se explica a continuación. En la fase de test, dada una frase de entrada $W = w_1, w_2, \dots w_{\mathrm{N}}$, obtenida a partir de la salida de un reconocedor del habla o bien proporcionada directamente como texto, el módulo de comprensión obtiene la secuencia de pares (*segmento*,*concepto*).

Con el fin de aprender los modelos de comprensión, se dispone de un conjunto de secuencias de conceptos asociadas a las frases de entrada, así como la asociación de segmentos de palabras correspondientes a cada concepto. En otras palabras, sea $\mathscr{W}$ el vocabulario de la tarea, y sea $\mathscr{C}$ el alfabeto de conceptos; el conjunto de entrenamiento es un conjunto de pares (s, c) donde:

$$s = s_1 s_2 \dots s_{\mathrm{n}}\ ,\ \ s_{\mathrm{i}} = w_{\mathrm{i}_1} w_{\mathrm{i}_2} \dots w_{\mathrm{i}_{|s_{\mathrm{i}}|}},$$
$$w_{\mathrm{i}_{\mathrm{j}}} \in \mathscr{W}\ ,\ i = 1, \dots, n\ ,\ j = 1, \dots, |s_{\mathrm{i}}|;$$
$$c = c_1 c_2 \dots c_{\mathrm{n}}\ ,\ c_{\mathrm{k}} \in \mathscr{C}\ ,\ k = 1, \dots, n.$$

cada frase de entrada en $W \in \mathscr{W}^*$ tiene un par (s,c) asociado, donde s es una secuencia de segmentos de palabras y c es una secuencia de unidades semánticas. Un ejemplo de par de entrenamiento para el

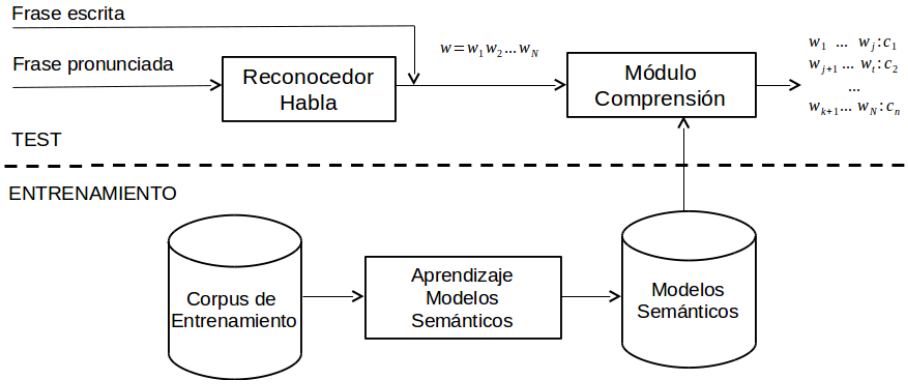

Figura 1: Esquema del proceso de comprensión

corpus DIHANA sería:

Frase de entrada W:

"me podría decir el que sale a las ocho qué tipo de tren es"

Par de entrada: $(s,c)=(s_1 s_2 s_3, c_1 c_2 c_3)$ donde:

s_1: me podría decir	c_1: *consulta*
s_2: el que sale a las ocho	c_2: *hora_salida*
s_3: qué tipo de tren es	c_3: $< tipo\_tren >$

La secuencia semántica s para el modelo semántico sería:

$$consulta\ hora\_salida < tipo\_tren >$$

Dado un conjunto de entrenamiento de este tipo, el problema de aprender el modelo de comprensión del lenguaje puede resolverse aplicando diferentes aproximaciones. En este trabajo presentamos tres aproximaciones estadísticas al problema de la comprensión del lenguaje: dos aproximaciones que modelizan la semántica con autómatas finitos y una aproximación que utiliza CRF.

Las dos aproximaciones basadas en autómatas finitos estiman dos tipos de modelos a partir de un conjunto de entrenamiento como el descrito anteriormente: un modelo semántico que representa el lenguaje de las concatenaciones de conceptos, y un modelo para cada concepto que representa el lenguaje de secuencias de palabras asociadas a ese concepto.

En el caso de la aproximación basada en CRF se establece para cada una de las palabras su concepto asociado empleando la notación IOB. Para ello se etiqueta el principio de un segmento con el prefijo "$B\_$" sobre el concepto, y para el resto de palabras del segmento se utiliza el segmento "$I\_$". Este etiquetado permite proporcionar información de los segmentos asociados a cada concepto al modelo CRF.

Palabra	Etiqueta
me	$B\_consulta$
podría	$I\_consulta$
decir	$I\_consulta$
el	B_hora_salida
que	I_hora_salida
sale	I_hora_salida
a	I_hora_salida
las	I_hora_salida
ocho	I_hora_salida
qué	$B\_ < tipo\_tren >$
tipo	$I\_ < tipo\_tren >$
de	$I\_ < tipo\_tren >$
tren	$I\_ < tipo\_tren >$
es	$I\_ < tipo\_tren >$

4.1 La aproximación 2-niveles

En esta aproximación, a partir de un conjunto de pares de entrenamiento (s,c) se estiman dos tipos de modelos (autómatas finitos): un modelo para el lenguaje de las concatenaciones de conceptos que llamaremos el lenguaje semántico $L_c \subseteq \mathscr{C}^*$, y un conjunto de modelos, uno por concepto $c_i \in \mathscr{C}$. El autómata finito A_c para el lenguaje semántico L_c se estima a partir de las cadenas semánticas $c \in \mathscr{C}^*$ del conjunto de entrenamiento. Un autómata finito A_{c_i} es estimado para cada concepto $c_i \in \mathscr{C}$ a partir del conjunto de segmentos s_i obtenido del conjunto de entrenamiento asociado a cada una de estas unidades semánticas c_i. Estas estimaciones son llevadas a cabo a través de técnicas de aprendizaje automático.

El modelo de comprensión A es un autómata finito que se obtiene con la susti-

tución en el autómata A_c de cada estado que representa un concepto $c_i \in \mathscr{C}$ por el autómata A_{c_i} correspondiente.

Una de las ventajas de esta aproximación, es que podemos escoger la técnica de aprendizaje más adecuada para la estimación de cada modelo: el modelo para lenguaje semántico y los modelos para los conceptos. La única restricción es que la representación de estos modelos debe ser dada en forma de un autómata finito. En este trabajo se han utilizado modelos de n-gramas.

Dada una frase de entrada, su análisis en base al algoritmo de Viterbi en el autómata finito A devuelve el mejor camino. Este camino proporciona, no solo la secuencia de conceptos sino también al segmentación de la frase analizada.

4.2 Aproximación basada en grafos

Esta aproximación se basa en la idea de construir un *grafo de conceptos* de forma que las distintas posibles interpretaciones semánticas de la frase de entrada (de test) estén codificadas como caminos en dicho grafo. En él cada arco estará etiquetado con una secuencia de palabras y el concepto que representa. La construcción de este grafo se lleva a cabo por medio de un algoritmo de programación dinámica que utiliza la información de modelos que representan las distintas formas de expresar léxicamente cada concepto. Estos modelos junto con un modelo semántico que representa las posibles concatenaciones de conceptos, se estiman a partir del conjunto de entrenamiento descrito al inicio de la sección. Al igual que en la aproximación 2-niveles, estos modelos pueden entrenarse siguiendo distintas técnicas de aprendizaje automático, en nuestro caso n-gramas.

Una vez construido el grafo de conceptos se realiza la búsqueda del mejor camino combinando la información de este grafo con la del modelo semántico. Esta búsqueda del mejor camino en el grafo de conceptos da como resultado la mejor secuencia de conceptos junto con la segmentación de la frase de entrada. iberspeechdemo2014 El grafo de conceptos tendrá $N+1$ nodos, donde N es el número de palabras de la cadena de entrada W. Para construir el grafo de conceptos se consideran todas las subcadenas $w_i...w_j$ contenidas en W, y para cada concepto c_k se calcula la probabilidad t que el modelo asociado a c_k asigna a ese segmento. Si la probabilidad t es no nula, entonces se crea un arco entre los nodos i y $j+1$ etiquetado con el par $(w_i...w_j,\ c_k)$ y con peso t^α. Este método se muestra en el Algoritmo 1. La función de α es escalar las probabilidades de los modelos que intervienen en la construcción del grafo de conceptos para su posterior combinación con la del modelo semántico durante la búsqueda del mejor camino en el grafo.

Algoritmo 1 Método para la construcción de un grafo de conceptos.

Entrada: Frase de entrada $W = w_1 w_2 ... w_N$, conjunto de modelos $M = \{M(c_1) \dots M(c_{|\mathscr{C}|})\}$ que permitan estimar las probabilidades $p(w_i...w_j|c_k)$, factor de escala α
Salida: Grafo de conceptos GC

```
Crear N + 1 nodos para GC y numerarlos comenzando en 1
Para i = 1 hasta N − 1 hacer
    Para j = i + 1 hasta N hacer
        Para k = 1 hasta |C| hacer
            t = p(w_i...w_j|c_k)
            Si t > 0 entonces
                Añadir a GC un arco con origen el nodo i y destino el j+1 etiquetado con el par (w_i...w_j, c_k) y con peso t^α
            Fin Si
        Fin Para
    Fin Para
Fin Para
Devolver GC
```

4.3 CRF

Los "Conditional Random Fields" (Lafferty, McCallum, y Pereira, 2001) son modelos "log-lineal" en los que se realiza una normalización a nivel de toda la frase. Los CRF reúnen algunas de las ventajas de los modelos generativos y discriminativos. Con ellos se pueden tener en cuenta muchas características de las entradas que se han aprendido de forma discriminativa, pero también tienen en cuenta las decisiones previas para escoger la mejor etiqueta semántica en cada momento. En estos modelos se representa la probabilidad condicional de una secuencia de etiquetas de conceptos $c_1 \dots c_N$ dada una secuencia de palabras $w_1 \dots w_N$ como:

$$p(c_1^n|w_1^n) = \frac{1}{Z} \quad \prod_{m=1}^{N} \lambda_m \cdot h_m(c_{n-1}, c_n, w_{n-k}^{n+k}) \Bigg)$$

donde λ_m es el vector de parámetros que se aprende a partir de un corpus etiquetado, y $h_m(c_{n-1}, c_n, w_{n-k}^{n+k})$ representa las dependencias entre las entradas (palabras u otras características que pueden aparecer en una ventana alrededor de la palabra a etiquetar) y los conceptos de salida. Por otra parte, el factor de normalización Z viene dado por

$$Z = \sum_{\tilde{c}_1^N} \prod_{n=1}^{N} (\tilde{c}_{n-1}, \tilde{c}_n, w_{n-k}^{n+k})$$

donde $\tilde{c}_{n-1}$ y $\tilde{c}_n$ son los conceptos predichos por las palabras previa y actual.

En nuestro caso hemos usado un conjunto de características básicas como son la información léxica y el etiquetado semántico. Para ello se ha definido una ventana $k = 2$ que incorpora las dos palabras (y su concepto asociado) previas y posteriores.

5 Experimentos

Para evaluar los diferentes sistemas de comprensión expuestos en este trabajo se ha realizado una serie de experimentos con el corpus en castellano DIHANA y el corpus en francés MEDIA. Para ello se definieron dos medidas: el CA (Concept Accuracy), que es el porcentaje de conceptos correctos; y el CSS (Correct Semantic Sequence) que es el porcentaje de secuencias completas semánticas correctas. Además se han incluido el porcentaje de sustituciones (S), borrados (B) e inserciones (I) entre la secuencia semántica correcta y la hipótesis dada por cada sistema para cada una de las frases. En los dos corpus se han evaluado las transcripciones correctas de los diálogos y en el caso del corpus DIHANA también las proporcionadas por un reconocedor de habla basado en HTK.

En las tablas se presentan los resultados para los tres sistemas: 2-niveles, Grafos y CRF. En todas las tablas de resultados se ha añadido la fila Oráculo en la que se elige para cada frase de test la interpretación semántica que más se ajusta a la de referencia, considerando las salidas de los tres sistemas. Estos valores constituyen una cota superior de los resultados alcanzables usando los tres sistemas a la vez.

Los 6.280 turnos del corpus de DIHANA se dividieron en un conjunto de entrenamiento de 4.887 turnos, un conjunto de desarrollo de 340 turnos y un conjunto de test de 1000 turnos. Los resultados obtenidos para las transcripciones correctas se muestran en la Tabla 3.

Sistema	**S**	**B**	**I**	**CSS**	**CA**
2-niveles	4,6	2,1	5,3	76,0	87,9
Grafos	2,1	1,5	1,0	88,2	95,4
CRF	3,9	3,8	2,0	79,6	90,4
Oráculo	1,5	1,2	0,7	91,5	96,6

Tabla 3: Resultados con DIHANA empleando como test las transcripciones correctas.

Los resultados obtenidos para la salida del reconocedor HTK con el corpus DIHANA se muestran en la Tabla 4. Como puede verse los resultados sobre las transcripciones correctas son muy altos, principalmente en el caso de los Grafos. Sin embargo, cuando se utiliza la salida del reconocedor, lógicamente disminuyen aunque en ese caso son los CRFs los que mejor resultado proporcionan. Esto puede deberse a que el método de aprendizaje de los modelos del sistema de Grafos genera un modelo muy adaptado a las muestras de entrenamiento y desarrollo. De este modo si hay una fuerte correlación entre el conjunto de entrenamiento y el de test el sistema funciona muy bien, pero conforme las estructuras sintácticas de las muestras de test se alejan de las de entrenamiento (como es el caso de la salida del reconocedor, y del corpus MEDIA que se verá más adelante) bajan las prestaciones de este sistema en comparación al resto.

Sistema	**S**	**B**	**I**	**CSS**	**CA**
2-niveles	8,0	2,9	11,3	64,0	77,9
Grafos	7,0	2,4	11,1	67,0	79,6
CRF	6,6	4,9	5,3	68,9	83,2
Oráculo	5,6	2,6	5,5	72,9	85,6

Tabla 4: Resultados con DIHANA empleando como test la salida del reconocedor HTK.

A la hora de mostrar los resultados de la segmentación y etiquetado semántico del corpus muchas veces no se evalúan etiquetas que no tienen significado semántico para la tarea como "*nada*", "*cortesía*", etc. Los resultados obtenidos para las transcripciones correctas y la salida del reconocedor HTK sin evaluar este tipo de etiquetas se muestran en la Tabla 5 y Tabla 6.

Como puede verse se mantiene la misma tendencia que en las tablas anteriores, aunque los resultados mejoran, ya que se eliminan los errores producidos por esas etiquetas sin significado semántico.

Sistema	S	B	I	CSS	CA
2-niveles	1,0	2,8	6,1	81,2	90,0
Grafos	0,2	1,6	1,7	91,5	96,5
CRF	1,3	3,3	3,5	83,5	92,0
Oráculo	0,2	0,9	1,4	94,3	97,5

Tabla 5: Resultados con las transcripciones correctas de DIHANA sin evaluar etiquetas sin significado semántico.

Sistema	S	B	I	CSS	CA
2-niveles	4,6	3,4	11,8	68,2	80,2
Grafos	4,3	2,8	11,6	70,2	81,3
CRF	4.2	4.3	7.2	72,2	84,3
Oráculo	3,6	2,5	7,2	75,9	86,7

Tabla 6: Resultados con DIHANA empleando como test la salida del reconocedor HTK sin evaluar etiquetas sin significado semántico.

Los 16.279 turnos de usuario del corpus MEDIA se dividieron en, 12.000 turnos para entrenamiento, 1.279 turnos para desarrollo y por último 3.000 turnos para test. En la Tabla 7 se muestran los resultados obtenidos para las transcripciones correctas del MEDIA.

Sistema	S	B	I	CSS	CA
2-niveles	5,7	7,8	6,9	69,7	79,6
Grafos	4,8	5,0	7,3	73,1	82,9
CRF	2,5	7,7	3,1	76,9	86,6
Oráculo	8,1	9,8	5,6	82,0	90,8

Tabla 7: Resultados con MEDIA empleando como test las transcripciones correctas.

Los resultados obtenidos para las transcripciones correctas sin evaluar etiquetas sin significado semántico como la etiqueta "*null*" se muestran en la Tabla 8.

En general los resultado para el corpus MEDIA son peores que los del corpus DIHANA. Esto puede deberse a que es un corpus con muchas más etiquetas semánticas, además de que algunas de ellas tienen pocas muestras en el corpus de entrenamiento. Por otra parte, ciertas etiquetas son iguales en cuanto a su realización léxica y sólo son distinguibles con el contexto semántico, como números en las etiquetas "*nombre_chambre*" (número de habitaciones), "*nombre_reservation*" (número de reservas) o "*nombre_hotel*" (número de hoteles). Es por tanto una tarea más difícil, aunque los resultados son adecuados a la tarea.

No existen otros trabajos comparables sobre el corpus DIHANA. Sin embargo si que

Sistema	S	B	I	CSS	CA
2-niveles	5.1	7.4	6.9	73,9	80,6
Grafos	4.2	5.6	6.1	77,7	84,1
CRF	2,8	7,8	2,0	80,9	87,3
Oráculo	2,3	4,9	1,8	85,4	91,0

Tabla 8: Resultados con MEDIA empleando como test las transcripciones correctas sin evaluar etiquetas sin significado semántico.

hay sobre el corpus MEDIA (Hahn et al., 2010) donde los mejores resultados reportados hasta el momento son del 89,4 % de CA, lo cual indica que nuestros resultados son competitivos.

6 Conclusiones

Hemos presentado en este artículo tres propuestas para la comprensión del habla, basadas todas ellas en el aprendizaje automático de modelos estadísticos. Los resultados experimentales muestran que este tipo de modelos es adecuado para las tareas de semántica restringida. Aunque existe un cierto deterioro en los resultados cuando se considera la salida de un reconocedor de voz, la capacidad de generalización y suavizado intrínseca a los modelos estadísticos permite mantener un buen resultado de comprensión.

Se ha comprobado que para estas dos tareas, DIHANA y MEDIA, los sistemas que devuelven mejores resultados en general son los basados en el modelo discriminativo CRF. En una comparación entre los dos métodos basados en modelos de autómatas finitos, los resultados muestran que el sistema de grafos funciona mejor que el de 2-niveles. Posiblemente esto ocurre porque en el sistema de grafos se realizan dos etapas. En la primera de seleccionan determinadas estructuras sintácticas asociadas a conceptos y en la segunda se elige la mejor concatenación de estas estructuras. Este proceso es más selectivo que la búsqueda del mejor camino en el modelo integrado del sistema de 2-niveles, que incorpora los autómatas de cada concepto al autómata que representa las concatenaciones de conceptos, y constituye un espacio de búsqueda mucho mayor. Esto supone un aumento de la cobertura del modelo en relación al modelo de grafos, pero puede introducir más confusión.

Bibliografía

Benedí, J.M., E. Lleida, A. Varona, M.J. Castro, I. Galiano, R. Justo, I. López de Letona, y A. Miguel. 2006. Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA. En *LREC*, páginas 1636–1639.

Bonneau-Maynard, H., S. Rosset, C. Ayache, A. Kuhn, y D. Mostefa. 2005. Semantic annotation of the French MEDIA dialog corpus. En *Proc. of InterSpeech 2005*, páginas 3457–3460, Portugal.

Calvo, M., F. García, L.-F. Hurtado, S. Jiménez, y E. Sanchis. 2013. Exploiting multiple hypotheses for multilingual spoken language understanding. En *Proc. of the CoNLL-2013*, páginas 193–201.

Dinarelli, M., A. Moschitti, y G. Riccardi. 2009. Concept Segmentation And Labeling For Conversational Speech. En *Interspeech*, Brighton.

Esteve, Y., C. Raymond, F. Bechet, y R. De Mori. 2003. Conceptual Decoding for Spoken Dialog systems. En *Proc. of EuroSpeech'03*, páginas 617–620.

García, F., L.-F. Hurtado, E. Segarra, E. Sanchis, y G. Riccardi. 2012. Combining multiple translation systems for Spoken Language Understanding portability. En *Proc. of IEEE Workshop on Spoken Language Technology (SLT 2012)*, páginas 282–289, Miami.

Hahn, S., M. Dinarelli, C. Raymond, F. Lefèvre, P. Lehnen, R. De Mori, A. Moschitti, H. Ney, y G. Riccardi. 2010. Comparing stochastic approaches to spoken language understanding in multiple languages. *IEEE Transactions on Audio, Speech, and Language Processing*, 6(99):1569–1583.

He, Y. y S. Young. 2003. A data-driven spoken language understanding system. En *Proc. of ASRU'03*, páginas 583–588.

Hurtado, L., E. Segarra, F. García, y E. Sanchis. 2004. Language understanding using n-multigram models. En *Advances in Natural Language Processing, Proceedings of 4th International Conference EsTAL*, volumen 3230 de *Lecture Notes in Computer Science*. Springer-Verlag, páginas 207–219.

Lafferty, J., A. McCallum, y F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. En *International Conference on Machine Learning*, páginas 282–289. Citeseer.

Laguna, S., M. Giménez, M. Calvo, F. García, E. Segarra, E. Sanchis, y L.-F. Hurtado. 2014. A Multilingual Spoken Language Understanding System. En *Proc. of the Iberspeech*, páginas 348–353, Las Palmas de Gran Canaria.

Lefèvre, F. 2007. Dynamic bayesian networks and discriminative classifiers for multi-stage semantic interpretation. En *ICASSP 2007*, volumen 4, páginas 13–16.

Ortega, L., I. Galiano, L.-F. Hurtado, E. Sanchis, y E. Segarra. 2010. A statistical segment-based approach for spoken language understanding. En *Proc. of InterSpeech 2010*, páginas 1836–1839, Makuhari, Chiba, Japan.

Raymond, C. y G. Riccardi. 2007. Generative and discriminative algorithms for spoken language understanding. *Proc. of Interspeech 2007*, páginas 1605–1608.

Segarra, E., E. Sanchis, M. Galiano, F. García, y L. Hurtado. 2002. Extracting Semantic Information Through Automatic Learning Techniques. *IJPRAI*, 16(3):301–307.

Seneff, S. 1992. TINA: A natural language system for spoken language applications. *Computational Linguistics*, 1(18):61–86.

Servan, C., N. Camelin, C. Raymond, F. Béchet, y R. De Mori. 2010. On the use of Machine Translation for Spoken Language Understanding portability. En *Procs. of ICASSP'10*, páginas 5330–5333.

Ward, W. y S. Issar. 1994. Recent improvements in the CMU spoken language understanding system. En *Proc. of the ARPA Human Language Technology Workshop*, páginas 213–216.

Desarrollo de recursos y herramientas lingüísticas

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 41-48 recibido 26-03-2015 revisado 27-04-2015 aceptado 19-05-2015

EusEduSeg: A Dependency-Based EDU Segmentation for Basque

EusEduSeg: Un Segmentador Discursivo para el Euskera Basado en Dependencias

Mikel Iruskieta, Benat Zapirain
IXA Group. University of the Basque Country
{mikel.iruskieta, benat.zapirain}@ehu.eus

Resumen: Presentamos en este artículo el primer segmentador discursivo para el euskera (EusEduSeg) implementado con heurísticas basadas en dependencias sintácticas y reglas lingüísticas. Experimentos preliminares muestran resultados de más del 85 % F_1 en el etiquetado de EDUs sobre el Basque RST TreeBank.
Palabras clave: Segmentación discursiva, Rhetorical Structure Theory (RST), segmentador, euskera

Abstract: We present the first discursive segmenter for Basque implemented by heuristics based on syntactic dependencies and linguistic rules. Preliminary experiments show F_1 values of more than 85% in automatic EDU segmentation for Basque.
Keywords: Discourse segmentation, Rhetorical Structure Theory (RST), segmenter, Basque

1 *Introduction*

An obligatory first step in the annotation of any discourse parser is to identify the discourse units. This is known as the segmentation phase. The aim of segmentation is to mark the elementary units of the text, or in other words, to establish the basic elements of each language analysis level in order to enable the subsequent identification of the relation that exist between them.

The definition of an Elementary Discourse Unit (EDU) is nowadays controversial in the areas of Discourse Studies, and, as a consequence, several segmentation granularities (van der Vliet, 2010) have been proposed within RST[1].

Although it is hardly ever explicitly stated, segmentation proposals are based on the following three basic concepts:

- Linguistic "form" (or category).
- "Function" (the function of the syntactical components).
- "Meaning" (the coherence relation between propositions).

The possible combinations between these basic concepts used in discourse segmentation and those proposed in RST are underlined in Figure 1.

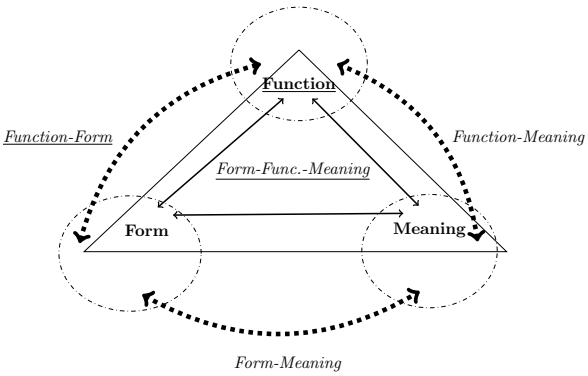

Figure 1: The basic concepts of discourse segmentation: form, function and meaning

Best-known segmentation proposals within RST are:

- The original RST proposal in English (Mann and Thompson, 1987): all clauses are EDUs, except for restrictive relative clauses and clausal subject or object components (syntactical function). This proposal is based solely on syntactical function.
- The first RST-based annotated corpus in English (Carlson and Marcu, 2001): in addition to that outlined in the orig-inal proposal, here both the compo-nents of attribution clauses (criterion based on function and meaning) and those phrases that begin with a discourse marker (e.g. *because of*, *spite of*, *accord-*

[1] A relational discourse structure theory proposed by Mann and Thompson (1987): for discourse coherence.

ISSN 1135-5948

ing to, etc.) are also segmented (criterion based on form and semantics).This proposal uses all three basic concepts: form, function and meaning.

– A segmentation proposal in English that adheres more closely to the original RST proposal (Tofiloski, Brooke, and Taboada, 2009): it segments verb clauses, coordinated clauses, ad-junct clauses and non-restrictive rela-tive clauses marked by a comma (it is a proposal based both on form restriction and syntactical function). Unlike in the proposal tabled by Carlson and Marcu (2001), in this method phrases beginning with discourse markers are not segmented, since they contain no verbs. In the annotation of the Spanish and Basque RST corpus, (da Cunha et al., 2010b; Iruskieta et al., 2013) this segmentation method was followed.

When attempting to define what a "discourse unit" actually is, these three basic concepts (form, function and meaning) pose a number of problems: *a*) If we based our analysis on form alone, many of the segmented elements would not be discourse units. *b*) If we based our analysis on function alone, then we would only be able to give annotators overly generalized definitions and imprecise segmentation criteria, such as adjunct clauses or adverbial clauses. *c*) And finally, if we based our analysis solely on meaning, we would encounter the problem of circularity between the segmentation annotation phase and the rhetorical relation annotation phase. The clearest example of this is that in order to annotate ATTRIBUTION relations, we would first have to segment the attribution clauses in the segmentation phase, resulting in a mixing of the two phases.

Following Thompson, Longacre, and Hwang (1985) we consider discourse units as functionally independent units, where three types of subordinate clauses can be distinguished: *i*) complements (which functions as noun phrases), *ii*) relative clauses (which functions as noun modifiers) and *iii*) adverbial clauses (which functions as modifiers of verb phrases or entire clauses). Blühdorn (2008) stated this subordinated but adverbial clauses can be seen as clause linkages, because it is the adverbial clauses which gives to the main clause a (discourse) thematic role.[2]

[2]More detailed information about adverbial clauses can be read in Liong (2000) and Lehmann (1985).

Clause type	**Example**
Independent sentence	[*Whipple (EW) gaixotasunak hesteei eragiten die bereziki.*]$_1$ GMB0503
Main, part of sentence	[pT1 tumoreko 13 kasuetan ez *zen* gongoila inbasiorik *hauteman*;]$_1$ [aldiz, pT1 101 tumoretatik 19 kasutan (18.6%) inbasioa *hauteman zen*, eta pT1c tumoreen artetik 93 kasutan (32.6%).]$_2$ GMB0703
Finite adjunct	[Haien sailkapena egiteko hormona hartzaileen eta c-erb-B2 onkogenearen gabeziaz baliatu gara,]$_1$ [*ikerketa anatomopatologikoetan erabili ohi diren zehaztapenak direlako.*]$_2$ GMB0702
Non-finite adjunct	[Ohiko tratamendu motek porrot eginez gero,]$_1$ [gizentasun erigarriaren kirurgia da epe luzera egin daitekeen tratamendu bakarra.]$_2$ GMB0502
Non-restrictive relative	[Dublin Hiriko Unibertsitateko atal bat da Fiontar,]$_1$ [zeinak Ekonomia, Informatika eta Enpresa-ikasketetako Lizentziatura ematen baitu, irlanderaren bidez.]$_2$ TERM23

Table 1: Main clause structures.

The segmentation guidelines we have use for Basque conflate all the approach presented before (Tofiloski, Brooke, and Taboada, 2009) and Basque clause combin-ing (Salaburu, 2012). As an example of what an EDU is, we show the main clause struc-tures in Table 1.

In this paper we present EusEduSeg[3] the first segmenter for Basque language, based on form and function rules. We evaluate the segmenter over a hand annotated corpora and we obtain promising results.

The remainder of this paper is structured as follows. Section 2 lays out the related work. Section 3 sets out the description of our system and Section 4 presents the experiment and results. Finally, Section 5 presents

[3]The segmenter EusEduSeg can be tested at http://ixa2.si.ehu.es/EusEduSeg/EusEduSeg.pl.

the discussion and establishes directions for future work.

2 Related Work

Although there are some works in Basque processing which identifies verbal chains, phrases (Aranzabe, 2008) and clauses (Aduriz et al., 2006), to cite some, there is not any discourse segmenter available for comparison in Basque. Iruskieta, Diaz de Ilarraza, and Lersundi (2011) established the bases for Basque discourse segmentation and implemented a prototypical segmenter reusing a statistical and morphological rule based chunk identifier (Arrieta, 2010). Including sentence boundaries, they obtained an F_1 of 66.94 in the experiments they car-ried out.

The evaluation of discourse segmentation it is not a trivial task, and several statistical measures have been used to check the robustness of a segmenter or to determine the reliability between human annotators and system evaluations:

i) Percent agreement was used to evaluate the agreement between human annota-tors by Hearst (1997) and Marcu (1999).

ii) Tofiloski, Brooke, and Taboada (2009) and Afantenos et al. (2010) used precision, recall and F_1 measures to evaluate the reliability and robustness of both automatic systems and human annotators.

iii) *Kappa* (κ) was used in Hearst (1997), Miltsakaki et al. (2004) and Tofiloski, Brooke, and Taboada (2009) to evaluate both automatic systems and human annotators.

Regarding to automatic discourse segmenters in languages others than Basque, Afantenos et al. (2010) presented a dis-course segmenter for French, da Cunha et al.(2010b) for Spanish and Tofiloski, Brooke, and Taboada (2009), Subba and Eugenio (2007) and Soricut and Marcu (2003) for En-glish. Table 2 summarizes the F_1 results pub-lished in those works.

Language	F_1	Reference
English	79	(Tofiloski, Brooke, and Taboada, 2009)
English	83-84	(Soricut and Marcu, 2003)
Spanish	80	(da Cunha et al., 2010a)
French	73	(Afantenos et al., 2010)

Table 2: State of the art in EDU parsing

The approach we followed to build our EDU segmentation system is rule-based and we avoid "same-unit" constructions as in Tofiloski, Brooke, and Taboada (2009). Specifically, as our rules are based on syntactical (dependencies) and morphological information, we follow a form-function approach for building our rule based automatic EDU segmentation.

3 EusEduSeg: System Description

From the syntactic point of view, most EDUs in the Basque RST TreeBank corpus exhibit two characteristic patterns that could be described as follows:[4]

- **Pattern 1**: verb nodes (*ROOT*, *ADI* and *ADT*) in the sentence's dependence tree govern an EDU if any of their recursively projected nodes accomplishes all the following conditions:

  1-*a*) It is the furthest node to the right from the governing head node (not necessarily the furthest one in the tree structure, but in the sentence order).

  1-*b*) It is a punctuation mark.

- **Pattern 2**: If a connector node (examples of LOT node are *edo* 'or', *eta* 'and', or *baina* 'but') has two direct verbal children nodes, then the connector node (LOT) delimits the frontier between two EDUs.

Given the simplicity of these dependency patterns, we developed a straightforward classifier that search for nodes that fulfill the previous conditions and label them as ending EDUs (E-EDU).

In order to better explain the patterns mentioned above, dependency trees in fig-ures 2 and 3 are introduced next. The tree in Figure 2 is a tree fragment (i.e. not the whole sentence's tree) representing an EDU that matches the pattern named as 1 right before. In this case, the node governing an EDU is the top most node in the tree, which is labeled as an verb (ADI) by Maltixa (Diaz de Ilarraza, Gojenola, and Oronoz, 2005), a dependency parser for Basque[5] (*lokalizatu* 'to

[4] Table 6 in Appendix A shows the descriptions of the Basque glosses employed in the paper.

[5] Maltixa can be tested at http://ixa2.si.ehu.es/maltixa/index.jsp.

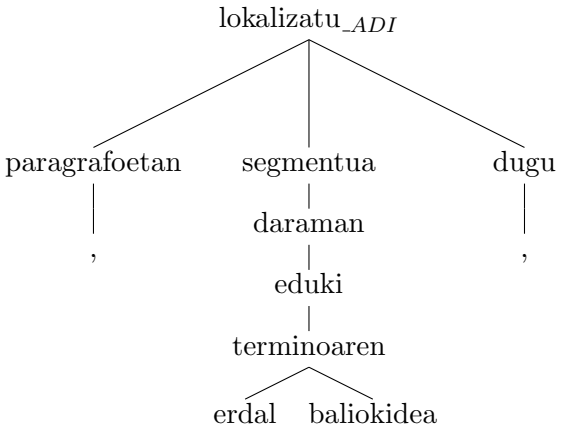

Figure 2: An application example of Pattern 1 (*paragrafoetan, erdal terminoaren eduki baliokidea daraman segmentua lokalizatu dugu,)* TERM28

locate'). As required by pattern 1, there is a punctuation node (a comma) under the auxiliary node *dugu* (auxiliar verb) that fulfills 1-a and 1-b conditions. This punctuation node is delimiting the frontier between the current EDU (represented in Figure 2) and the next one (the rest of the sentence is omitted here for lacking of space) and it should be labeled as an end-EDU (E-EDU) by the segmenter.

Figure 3 shows a tree fragment from the Basque RST TreeBank corpus that exactly matches pattern 2. There are two verbal nodes (ADI and ADT) and both share the same connector (LOT) parent node. As stated in pattern 2, the connector node es-tablishes boundaries between EDUs. In the example of Figure 3 the boundaries (E-EDU and B-EDU) would establish as follows (in bold): *...formal eta osoa lortzea lan neketsua* ***da***$_{E-EDU}$ ***eta***$_{B-EDU}$ *horretan datza atal...*

In order to increase the performance of the classifier, we added a post processing layer consisting of a rule set based on previous observations by Iruskieta, Diaz de Ilarraza, and Lersundi (2011). Target and token sequences that matches the target are underlined in corpus examples below:

- **Rule 1** (temporal): label *ADI (ERL:DENB)* nodes as *E-EDU*.

  (1) *Termino teknikoak hautatzerakoan*][*deklinabide kasua erabakigarria izan daiteke.* TERM31

- **Rule 2** (conditional-I): label *ERL:BALD + ,* sequences as *E-EDU*.

  (2) *Halako tresna bat euskararako garatu nahi badugu,*][*eragozpen gehiago topatuko dugu ondoko hiru arrazoiengatik.* TERM31

- **Rule 3** (conditional-II): tag *ERL:BALD + ere + ,* sequences as E-*EDU*.

  (3) *Emaitzarik ez badugu ere,*][*eredua izen-sintagmarena baino zabalagoa izango dela sumatzen dugu.* TERM31

- **Rule 4** (adjunct): label *ADI + ADB + ,* sequences as *E-EDU*.

  (4) *Ohiko tratamendu motek porrot eginez gero,*][*gizentasun erigarriaren kirurgia da epe luzera egin daitekeen tratamendu bakarra.* GMB0502

- **Rule 5** (reason): label *ERL:KAUS + ,* sequences as *E-EDU*.

  (5) *Hona hemen oin malgua izateagatik*][*kalkaneo-stop teknika erabiliz gure zerbitzuan ebakuntza egin diegun haurrek izandako emaitzak* GMB0601

- **Rule 6** (concessive): label *ERL:KONT* nodes as *E-EDU*.

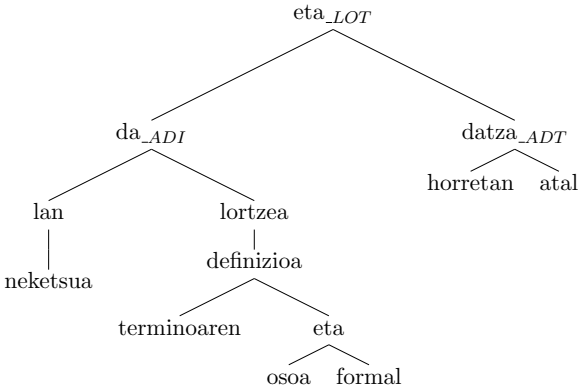

Figure 3: An application example of Pattern 2 (*Terminoaren definizio formal eta osoa lortzea lan neketsua eta horretan datza* [...] *atal* [...]) TERM31

(6) *Prebentzio metodoen eta artoplastiako teknika modernoen laguntzaz horrelako kasuak murriztu diren arren,*][*infekzio hori sendatzea erronka bat da oraindik ere.* GMB0802

- **Rule 7** (purpose): label *ADI(tzeko) + IZE + ,* as *E-EDU*.

(7) ingurunea aldatu ondoren elkarrekintza magnetikoak aztertzeko asmoz,][eta inguru biologikoetan ere erabiltzeko asmoz. ZTF17

4 Experiments and Results

4.1 Datasets

The corpus[6] used in this study consists of manually annotated abstracts from three specialized domains (medicine, terminology and science), and, it comprises 60 documents that contain 15,566 words (803 sentences) that were manually annotated with 1,355 EDUs and 1,292 relations. The corpus was analyzed with Maltixa, and randomly divided into training (50% for rule designing), development (25% for rule tuning) and test (25% for testing) sets.

4.2 EusEduSeg: EDU Segmenter

As mentioned before, the EDU classifier is entirely based on dependency and linguistic rules, as well as on a final consistency layer that checks the resulting EDUs with the aim of removing duplicated and incorrectly built EDUs (e.g: EDUs with no verbs in). In order to determine the influence of each rule set in the EDU segmentation task, we developed three different versions from the main classification system described in Section 3:

- ***EDU-Seg-1***: an EDU segmenter based only on dependency based patterns 1 and 2 described in Section 3.
- ***EDU-Seg-2***: an EDU segmenter based only on linguistic based rules (rules 1-7 from Section 3).
- ***EDU-Seg-3***: an EDU segmenter that takes advantage from both dependency based patterns and linguistic rules.

[6]The RST Basque Treebank (Iruskieta et al., 2013) and it's segmentation can be consulted at: http://ixa2.si.ehu.es/diskurtsoa/en/.

It is worth to remember that segmenter's rules and heuristics were developed manually and based, when needed, on observations made in training or development data.

EusEduSeg gives the possibility to configure several output formatting options that can be used in several tasks: *a*) web format to use in other NLP tasks. *b*) RSTTool format to annotate manually the RS-tree with RSTTool (O'Donnell, 2000). *c*) DiZer format (Pardo, Nunes, and Rino, 2004) to use in an automatic discourse parser.

System architecture is presented in Figure 4.

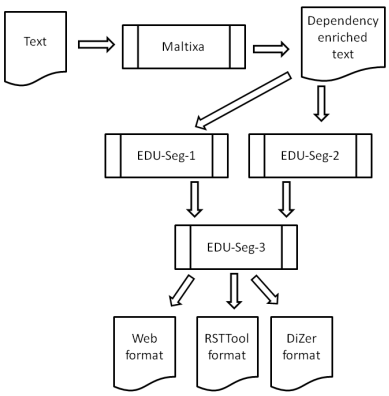

Figure 4: EusEduSeg phases

4.3 Evaluation measures

Performance of EDU segmenters has been reported with the standard precision, recall and F_1 measures, in similar way to many other authors on the task such as Tofiloski, Brooke, and Taboada (2009) and Afantenos et al. (2010). We calculate each of the mea-sures as follows:

$$precision = \frac{correct_{E-EDU}}{correct_{E-EDU} + excess_{E-EDU}}$$

$$recall = \frac{correct_{E-EDU}}{correct_{E-EDU} + missed_{E-EDU}}$$

$$F_1 = \frac{2 * precision * recall}{precision + recall}$$

where $correct_{E-EDU}$ is the number of correct *end-EDU*s, $excess_{E-EDU}$ is the number of overpredicted *end-EDU*s and $missed_{E-EDU}$ is the number of *end-EDU*s the system missed to tag.

Data set	correct	excess	missed	precision	recall	F_1	F_1'
Train	592	49	173	92.35	77.38	84.21	61.72
Dev	237	36	79	86.81	75.00	80.47	48.88
Test	292	25	95	92.11	75.45	82.95	60.52

Table 3: Results for EDU-Seg-1 on train, development and test sets

Data set	correct	excess	missed	precision	recall	F_1	F_1'
Train	548	14	217	97.5	71.63	82.59	53.89
Dev	208	9	108	95.85	65.82	78.04	30.76
Test	259	16	128	94.18	66.92	78.24	45.03

Table 4: Results for EDU-Seg-2 on train, development and test sets

Data set	correct	excess	missed	precision	recall	F_1	F_1'
Train	621	62	144	90.92	81.17	85.71	66.88
Dev	240	43	76	84.80	75.94	80.13	49.36
Test	303	39	84	88.59	78.29	83.12	62.61

Table 5: Results for EDU-Seg-3 on train, development and test sets

4.4 Results

Tables 3, 4 and 5 show the results ob-tained by EDU-Seg-1, EDU-Seg2 and EDU-Seg-3 respectively at the task of automatic segmentation of Basque texts. Correct, excess, missed, precision, recall and F_1 measures are reported, as customary for all data sets. The difference between F_1 and F_1' is that while former refers to classifier's F-score for all EDUs in the data set, latter refers to the F-score for "non trivial" EDUs only (hits on trivially identifiable EDU boundaries that begin or end a sentence are not take into account when computing F_1'). F_1' should be considered as the real indicator of the segmenter's performance.

Results show very high precision values for all segmenters used in the experiments. As already explained in previous sections, the heuristic and rule based engine of the segmenters makes this high precision values likely to be expected.

Regarding to the comparison between dependency based heuristics and linguistic rules (results shown in Table 3 and 4 respectively), linguistic rules are more precise than heuristics, but, on the other hand, higher recall values in Table 3 suggest that dependency based heuristics seem to be more general or better suited for broad spectrum EDU labeling.

Table 5 reports our best results in EDU segmentation experiments. The improvements in F_1 and F_1' with respect to the values in tables 3 and 4, seem to indicate that EDU-Seg-3 is able to successfully combine knowledge bases from EDU-Seg-1 and EDU-Seg-2, as well as that both dependency based heuristics and linguistic rules seem to be relatively complementary.

4.5 Error analysis

A more detailed error analysis, which is not under the scope of this work, will be useful for the future development of the automatic text segmentation of Basque text and also to improve Maltixa the automatic dependency analyzer for Basque.

A complex clause combining, as in Example 8, with three verbs (two coordinated finite verbs *erabakitzen dute* 'they decide it' and *jotzen dute* 'they go to' and one nominalized *jotzea* 'the going'), which can be detected with our system (Pattern 2), was not segmented by our system, due to some errors done by the dependency parser.

(8) Erabiltzaileen % 80ak bere kabuz erabakitzen dute larrialdi zerbitzu batetara jotzea][eta kontsulta hauen % 70a larritasun gutxikotzat jotzen dituzte zerbitzu hauetako medikuek.
*GMB*0401
The 80% of the users go by their own initiative to the emergency department,][and the 70% of the surgeries are

considered slights by the health staff. *TRANSLATION*

5 Conclusions and Future Work

In this paper we have introduced EusEduSeg, the first discourse segmenter for Basque implemented with simple dependency based heuristics and several high precision linguistic rules.

Experiments carried out on the Basque RST TreeBank corpus show competitive and promising results given the simplicity of the proposed solution and, in the same way, they leave enough room for improvement to more sophisticated and machine learning based architectures.

The authors are currently striving to achieve the following aims:

- To increase the performance of the segmenter adding more rules or better tuning the existing ones.
- To integrate a new layer of Constraint Grammar rules from previous work of Iruskieta, Diaz de Ilarraza, and Lersundi (2011).
- To train more sophisticated and robust classifiers by using state–of–the–art machine learning algorithms.
- To export the rule set of EusEduSeg into other languages such as English, Spanish or Portuguese. Given the lexical dependency of rules 1-7 from Section 3, this exportation task could be tough. However, patterns 1 and 2 seem more neutral and, thus, more suitable to be applied to other languages.

References

Aduriz, I., B. Arrieta, J.M. Arriola, A. Diaz de Ilarraza, E. Iza-girre, and A. Ondarra. 2006. Muga Gramatikaren optimizazioa (MuGa). Technical report, EHU.

Afantenos, S. D., P. Denis, P. Muller, and L. Danlos. 2010. Learning recursive segments for discourse parsing. In *Seventh conference on Inter-national Language Resources and Evalua-tion*, pages 3578–3584, Paris, France, 19-21 May.

Aranzabe, M. J. 2008. Dependentzia-ereduan oinarritutako baliabide sintak-tikoak: zuhaitz-bankua eta gramatika konputazionala. Doktore-tesia, Euskal Herriko Unibertsitatea, Donostia.

Arrieta, B. 2010. Azaleko sintaxi-aren tratamendua ikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera koma-zuzentzaile batean. Doktore-tesia, Euskal Herriko Unibertsitatea, Donostia.

Blühdorn, H., 2008. *Subordination and coordination in syntax, semantics and discourse: Evidence from the study of connectives.* 'Subordination' versus 'Coordination' in Sentence and Text. Benjamins, Amsterdam.

Carlson, L. and Daniel M. 2001. Discourse tagging reference manual. Technical report.

da Cunha, I., E. SanJuan, J.M. Torres-Moreno, M. Lloberes, and I. Castellón. 2010a. Discourse segmentation for Spanish based on shallow parsing. In *9th Mexican international conference on Advances in artificial intelligence: Part I*, pages 13–23, Pachuca, Mexico, 8-13 November. Springer-Verlag.

da Cunha, I., E. SanJuan, J.M. Torres-Moreno, M. Lloberes, and I. Castellón. 2010b. Diseg: Un segmentador discursivo automatico para el español. *Procesamiento de Lenguaje Nat-ural*, 45.

Diaz de Ilarraza, A., K. Gojenola, and M. Oronoz. 2005. Design and Development of a System for the Detection of Agreement Errors in Basque. In *Computational Linguistics and Intelligent Text Processing*, pages 793–802. Springer.

Hearst, M. A. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.

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

Iruskieta, M., A. Diaz de Ilarraza, and M. Lersundi. 2011. Bases para la implementación de un segmentador discursivo para el euskera. In *8th Brazilian Symposium in Information and Human Language Technology (STIL 2011),*

OCTOBER 2011.

Lehmann, C. 1985. Towards a typology of clause linkage. In *Conference on Clause Combining*, volume 1, pages 181–248.

Liong, T. 2000. Adverbial clauses, functional grammar, and the change from sentence grammar to discourse-text grammar. *Círculo de lingüística aplicada a la comunicación*, 4(2).

Mann, W. C. and S. A. Thompson. 1987. Rhetorical Structure Theory: A Theory of Text Organization. *Text*, 8(3):243–281.

Marcu, D., 1999. *Discourse trees are good indicators of importance in text*, pages 123–136. Advances in Automatic Text Summarization. MIT, Cambridge.

Miltsakaki, E., R. Prasad, A. Joshi, and B. L. Webber. 2004. Annotating discourse connectives and their arguments. In *HLT/NAACL Workshop on Frontiers in Corpus Annotation*, pages 9–16, Boston, USA.

O'Donnell, M. 2000. RSTTool 2.4: a markup tool for Rhetorical Structure Theory. In *First International Conference on Natural Language Generation INLG '00*, volume 14, pages 253–256, Mitzpe Ramon, June12-16. ACL.

Pardo, T. A. S., M. G. V. Nunes, and L. H. M. Rino. 2004. DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese. *Advances in Artificial Intelligence–SBIA 2004*, pages 224–234.

Salaburu, P. 2012. Menderakuntza eta menderagailuak (Sareko Euskal Gramatika: SEG). http://www.ehu.es/seg/morf/5/2/2/2.

Soricut, R. and D. Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1, pages 149–156. Association for Computational Linguistics.

Subba, R. and B. Di Eugenio. 2007. Automatic discourse segmentation using neural networks. In *11th Workshop on the Semantics and Pragmatics of Dialogue*, page 189–190, Trento, Italy, 30-1 May-June.

Thompson, S. A., R. Longacre, and Shin Ja J. Hwang, 1985. *Adverbial clauses*, volume 2 of *Language Typology and Syntactic Description: Complex Constructions*, pages 171–234. Cambridge University Press, New York.

Tofiloski, M., J. Brooke, and M. Taboada. 2009. A syntactic and lexical-based discourse segmenter. In *47th Annual Meeting of the Association for Computational Linguistics*, pages 77–80, Suntec, Singapore, 2-7 August. ACL.

van der Vliet, N. 2010. Syntax-based discourse segmentation of Dutch text. In *15th Student Session, ESSLLI*, pages 203–210, Ljubljana, Slovenia, 1-12 August.

A Appendix: Glosses employed in the paper

Gloss abbrev.	Description
ADB	Adverb
ADI	Non-finite verb
ADL	Auxiliary finite verb
ADT	Finite verb
AUX	Auxiliary
BALD	Conditional clause
DENB	Temporal clause
ERL	Clause relation function
IZE	Noun
KAUS	Causal clause
LOT	Connector
PUNT	Punctuation
ROOT	Root of sentence

Table 6: Glosses used in examples

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 49-56 recibido 27-03-2015 revisado 27-04-2015 aceptado 22-05-2015

Classification of Grammatical Collocation Errors in the Writings of Learners of Spanish

Clasificación de errores gramaticales colocacionales en textos de estudiantes de español

Sara Rodríguez-Fernández
DTIC, UPF
C/Roc Boronat, 138
08018 Barcelona
sara.rodriguez.fernandez@upf.edu

Roberto Carlini
DTIC, UPF
C/Roc Boronat, 138
08018 Barcelona
roberto.carlini@upf.edu

Leo Wanner
ICREA y DTIC, UPF
C/Roc Boronat, 138
08018 Barcelona
leo.wanner@upf.edu

Resumen: Las combinaciones recurrentes y arbitrarias de palabras (*colocaciones*) son clave para el aprendizaje de lenguas pero presentan dificultades incluso a los estudiantes más avanzados. El uso de herramientas eficientes destinadas al aprendizaje de colocaciones supondría una gran ayuda, sin embargo, las que existen actualmente intentan corregir colocaciones erróneas sin diferenciar entre los distintos tipos de errores ofreciendo, como consecuencia, largas listas de colocaciones de muy diversa naturaleza. Además, sólo se consideran los errores léxicos, dejando de lado los gramaticales que, aunque menos frecuentes, no pueden ignorarse si el objetivo es desarrollar una herramienta capaz de corregir cualquier colocación errónea. En el presente trabajo se propone un método de clasificación automática de errores colocacionales gramaticales cometidos por estudiantes de español estadounidenses, como punto de partida para el diseño de estrategias de corrección específicas para cada tipo de error.
Palabras clave: Aprendizaje de lenguas, colocaciones, tipolgía de errores colocacionales, clasificación de errores gramaticales colocacionales

Abstract: Arbitrary recurrent word combinations (*collocations*) are a key in language learning. However, even advanced students have difficulties when using them. Efficient collocation aiding tools would be of great help. Still, existing "collocation checkers" still struggle to offer corrections to miscollocations. They attempt to correct without making any distinction between the different types of errors, providing, as a consequence, heterogeneous lists of collocations as suggestions. Besides, they focus solely on lexical errors, leaving aside grammatical ones. The former attract more attention, but the latter cannot be ignored either if the goal is to develop a comprehensive collocation aiding tool, able to correct all kinds of miscollocations. We propose an approach to automatically classify grammatical collocation errors made by US learners of Spanish as a starting point for the design of specific correction strategies targeted for each type of error.
Keywords: Second language learning, collocation, collocation error typology, grammatical collocation error classification

1 Introduction

Over the last decades, collocations, i.e., idiosyncratic word co-occurrences such as *spend time*, *take* [*a*] *leave*, *fierce heat*, *deep concern*, and so on have attracted increasing attention of research not only in computational lexicography and lexicology, but also in second language learning (Granger, 1998; Lewis, 2000; Nesselhauf, 2004; Nesselhauf, 2005; Lesniewska, 2006; Alonso Ramos et al., 2010). Studies indicate that collocations are a real challenge for language learners and that they are difficult to master even by advanced students (Nesselhauf, 2003; Bahns and Eldaw, 1993). Wible et al. (2003) show that collocation errors are the most frequent errors found in the writings of students. Orol and Alonso Ramos (2013)' study furthermore reveals that the "collocation density" in learner corpora is nearly the same as in na-

ISSN 1135-5948

tive corpora, i.e., that the use of collocations by learners is as common as it is by native speakers. At the same time, they also find that the collocation error rate in learner corpora is about 32% (compared to about 3% by native speakers). That is, automatic collocation error detection and correction in the context of *Computer Assisted Language Learning* (CALL) could be of great aid to support the learners for better mastering of collocations.

Since the pioneering work by Shei and Pain (2000), several "collocation checkers" have been developed. Most often, these checkers draw upon a collocation list extracted from a reference corpus to compare a collocation used by the student with those in the list (or with variants of those in the list) and thus to detect possible miscollocations (Chang et al., 2008; Park et al., 2008; Östling and Knutsson, 2009; Wu et al., 2010; Dahlmeier and Ng, 2011; Kanashiro Pereira, Manguilimotan, and Matsumoto, 2013) and then potentially offer a list of possible corrections (filtered or ranked according to different metrics).

However, no matter what technique is behind them, state-of-the-art collocation checkers suffer from two main limitations. Firstly, they are able to offer as miscollocation correction suggestions merely large heterogeneous lists of collocations in which one of the words involved in the miscollocation occurs. The learner is thus left with the task of identifying the most appropriate correction by themselves. But this is usually a rather complex task for a language student since selecting a collocation from a list implies that the student knows the meaning of all the collocations in the list, or spends extra time trying to find it. Secondly, they focus only on most common variants of miscollocations.

Both limitations are due to the fact that collocation checkers do not distinguish so far between different types of miscollocations, let alone address all types of miscollocations. Alonso Ramos et al. (2010) argue that collocation errors may be very different in their nature and provide a detailed typology of miscollocations. Comprehensive collocation error type-specific correction techniques would thus most certainly improve the correction performance. However, in order to be able to develop such techniques, we must first be able to classify detected miscollocations, for instance, with respect to Alonso Ramos et al. (2010)'s typlogy. This is the goal of our work.

Alonso Ramos et al. (2010)'s miscollocations typology distinguishes at the first level *grammatical* vs. *lexical* collocation errors. Grammatical collocation errors are more subtle. At the same time, they are also quite common: according to Alonso Ramos et al. (2010), 38% of the miscollocations contain grammatical errors. Therefore, we focus, in what follows, on the automatic classification of grammatical miscollocations.

In the following section, we define in more concrete terms the notion of collocation we use in our work and introduce Alonso Ramos et al. (2010)'s miscollocations typology, which we use in our experiments. Section 3 presents the experiments, and in Section 4, the results of these experiments are discussed. Section 5, finally, outlines the conclusions and our future work in the area of miscollocation classification and correction.

2 Fundamentals on collocations

2.1 The notion of collocation

The term "collocation" as introduced by Firth (1957) and cast into a definition by Halliday (1961) encompasses the statistical distribution of lexical items in context: lexical items that form high probability associations are considered collocations. It is this interpretation that underlies most works on automatic identification of collocations in corpora; (Choueka, 1988; Church and Hanks, 1989; Pecina, 2008; Evert, 2008; Bouma, 2010). However, in contemporary lexicography and lexicology an interpretation that stresses the idiosyncratic nature of collocations prevails. According to Hausmann (1984), Cowie (1994), Mel'čuk (1995) and others, a collocation is a binary idiosyncratic co-occurrence of lexical items between which a direct syntactic dependency holds and where the occurrence of one of the items (the *base*) is subject of the free choice of the speaker, while the occurrence of the other item (the *collocate*) is restricted by the base. Thus, in the case of *take* [*a*] *walk*, *walk* is the base and *take* the collocate, in the case of *high speed*, *speed* is the base and *high* the collocate, etc. It is this understanding of the term "collocation" that we find reflected in general public collocation dictionaries and that we follow since it seems most useful in the context of second language learning.

2.2 Grammatical miscollocation typology

The typology suggested by Alonso Ramos et al. (2010) groups collocation errors according to three parallel dimensions. The first dimension refers to the location of the error, i.e., whether the collocation as a whole is incorrect or whether one of its elements (the base or the collocate) is incorrect. The second dimension presents differentiations of the characterization of the linguistic phenomena that were observed in miscollocations. The most global differentiation level suggests three error types: lexical, grammatical, and register. The third dimension captures the possible reasons why collocation errors are produced, both interlingual and intralingual. As mentioned above, we focus on the grammatical errors of the second dimension.

Grammatical errors are divided into eight different types:

1. *Determination errors*: Errors resulting from the omission of a determiner when it is required by the collocation, or from its use when the collocation does not accept it; cf., e.g.: **terminar escuela* 'to finish school', where the determiner is expected in Spanish, but is missing.

2. *Number errors*: Errors produced when either the plural or the singular form of a lexical unit is required for a particular collocation, but the opposite is chosen; cf., e.g., **estamos en vacación* 'to be on holiday', where the singular form is used when plural is needed.

3. *Gender errors*: Errors resulting from the choice of the incorrect gender form of the base; cf., e.g., **pasar los vacaciones* 'to spend the holidays'.

4. *Government errors*: Errors produced when the governing preposition of the base or the collocate is missing or mistakenly chosen, or when a preposition is used when there should be none; cf., e.g., **ver a la película* 'to watch a movie, lit. to watch at a movie'. In Spanish, the preposition *a* is required for a direct object only when it refers to people.

5. *Governed errors*: Errors resulting from the wrong use or omission of a preposition that governs the whole collocation; cf., e.g., **estar en buen humor* 'to be in a good mood', instead of *estar de buen humor*.

6. *Specification errors*: Errors produced when a modifier of the base is missing; cf., e.g., **hacer un aterrizaje* 'to make a landing', where the modifier *forzoso* is needed.

7. *Pronoun errors*: Errors resulting from the inappropriate use or the absence of the reflexive pronoun of a verbal collocate; cf., e.g., **las plantas mueren*, 'plants die', where apart from the incorrect lexical choice of the collocate, *morir* instead of *secar*, the reflexive particle *se* is missing.

8. *Order errors*: Errors produced when the base and the collocate appear in the wrong order; cf., e.g., **reputación mala* 'bad reputation', instead of *mala reputación*.

We found that types 5 and 6, i.e. *Governed* and *Specification* errors, are very seldom. For this reason we opted not to consider them at this stage of the experiments.

3 Experiments

The examples of grammatical miscollocation types above illustrate that some of the grammatical error types (e.g., the *Gender* errors and *Order* errors), can be considered a problem of a grammar checker rather than of a collocation checker. We address them nonetheless in the context of collocation verification and correction because they make a collocation to be incorrect.

3.1 Methodology

We developed a set of functions. Each function focuses on the identification of one specific type of grammatical error in given miscollocations. Each function has thus been designed taking into account both the specific particular characteristics of the type of error it deals with and the possibility of a collocation being affected by several errors at the same time, either grammatical, lexical, register or any combination of them. All six functions (recall that we neglect two types of grammatical errors for the moment) receive as input miscollocations found in writings by learners of Spanish, and most of them use a reference native corpus of Spanish (henceforth, RC). In what follows, we briefly describe each one of them.

Determination errors. This function queries the RC to look up common occurrences of both the base and the collocate of the miscollocation, including those with the presence of a determiner and those in which no determiner is found. If the number of occurrences with the determiner is significantly higher than the number of occurrences without the determiner, the collocation is consid-

ered to require a determiner. In this case, if the context of the miscollocation does not contain a determiner, a determination error is flagged. Along the same lines, if it is determined that the collocation does not take a determiner, but the learner uses one, again, a determination error is flagged.

Number errors. Number errors can affect both the base and the collocate and are not necessarily manifested in terms of the lack of concordance, as, e.g., in **tener una vacación* 'to have a holiday', **dimos bienvenidas* 'to welcome', **gané pesos* 'to put on weight', etc. In order to check whether a collocation contains a number error, the corresponding function retrieves from the RC combinations of the lemmas of the base, collocate and the prepositions that depend on the dependent element. In other words, given a preposition, all possible combinations of the forms of the base and the collocate with that particular preposition are retrieved. Then, alternative number forms of the base and collocate are generated (i.e., if an element in the miscollocation is in plural, its singular form is generated, and vice versa) and occurrences of their combinations are retrieved from the RC. If the original form is not one of the possible combinations retrieved from the RC, but any of the alternatives is, the miscollocation is assumed to contain a number error.

Gender errors. Only miscollocations that have a noun as their base can contain this kind of error. However, the form of the base is rarely erroneous (cf., e.g., **pasar los vacaciones*). Rather, there is often a lack of concordance between the base and its determiner, or between the base and the collocate (in N–Adj collocations), resulting from the wrong choice of the gender of the determiner respectively collocate. For this reason, the corresponding function checks the gender of the determiner and adjectives of the base of the given miscollocation. Both the frequency of the miscollocation n-gram (i.e., string consisting of the collocate and the base with its determiner) and linguistic information are considered. For each miscollocation, the function retrieves from the RC the frequency of the original n-gram. Then, it generates new alternatives by changing the gender of the determiner (in VN, NN or prepositional collocations) or the adjective (in NAdj collocations) and looks for the frequency of the new combinations. If this happens to be higher than the frequency of the miscollocation, a gender error is assumed. Otherwise, the concordance between the base and the determiner respectively collocate is checked. If no concordance is found, a gender error is assigned.

Government errors. For identifying this kind of error, we take into account the context in which the miscollocation appears. For this purpose, first, syntactic patterns that contain the miscollocation's base and collocate and any preposition governed by either of the two are retrieved from the RC. Then, it is looked up whether the original syntactic miscollocation pattern that involves a governed preposition appears in the retrieved list. If this is not the case, the miscollocation is assumed to contain a government error.

Pronoun errors. In order to identify pronoun errors, a similar approach to the one used for recognizing determination errors is followed. In this case, frequencies of the combinations with and without reflexive pronouns are retrieved and compared to the miscollocation.

Order errors. To identify an order error, the frequency of the given miscollocation in the RC is calculated. Then, the frequencies of all the possible permutations of the elements of the collocation are compared to the frequency of the miscollocation. If any of them is significantly higher, the collocation is considered to contain an order error.

3.2 Experimental setup

For our experiments, we used a fragment of the Spanish learner corpus CEDEL2 (Lozano, 2009). CEDEL2 is composed of writings of native speakers of US English with different levels of proficiency in Spanish, from 'low-intermediate' to 'advanced'. The writings are of different styles and on different topics (opinion essays, accounts of some past experience, descriptions and letters, etc.). In total, we used 517 texts, with an average of 500 words. Each text was annotated with both correct and incorrect collocations. The number of miscollocations ascended to 1145. Table 1 shows the number of annotated instances for all eight grammatical collocation errors. Our reference corpus consisted of 7 million sentences from newspaper material in Spanish, stored and indexed in Solr. To obtain syntactic dependency information used

in some of the error recognition functions, both corpora were processed with Bohnet (2010)'s dependency parser.

Class	#Instances
Determination errors	146
Number errors	44
Gender errors	77
Government errors	225
Governed errors	2
Specification errors	1
Pronoun errors	28
Order errors	28

Table 1: Number of instances of the grammatical collocation errors annotated in CEDEL2

3.3 Results

Table 2 shows the classification accuracy of the individual grammatical error identification functions for both the positive (collocations containing the type of error that is to be identified) and the negative cases (incorrect collocations affected by any kind of error, except the one that is dealt with).

Type of error	(+)	(−)
Determination	0.719	0.793
Number	0.659	0.851
Gender	0.818	0.989
Government	0.68	0.708
Pronoun	0.357	0.99
Order	0.75	0.848

Table 2: Accuracy of the error detection functions

4 Discussion

We carried out an analysis of the misclassified instances for each experiment, both for the positive and negative classes. In what follows, we present some examples that illustrate the most relevant findings for each type of error.

In all functions, the error identification has been negatively influenced by: (i) the presence of multiple errors in collocations, which causes that queries to the RC do not retrieve any information, and (ii) the automatic preprocessing of the CEDEL2 corpus (note that we are dealing with writings by language learners; the sentences are thus often ungrammatical, such that the error rate of the preprocessing tools (lemmatizer, POS-tagger, morphology-tagger and parser) is considerably higher than in native texts).

Determination errors. As illustrated in the examples (1–2), some determination errors are not identified as such because these collocations can be found both with and without determiner, depending on the context. For instance, a determiner can be required by a specifier, as in (1). Also, we find a singular form of the collocation with a determiner, as in (2), where *tener un hijo* 'to have a child' is correct.

(1) **tiene una reputación*, instead of *tiene reputación* 'to have a reputation'

(2) **tiene los hijos*, instead of *tiene hijos* 'to have children'

With regard to the negative case, i.e., the classification of miscollocations that contain other kinds of errors as determination error, the same reasons can be identified as the source of error. In the following examples, the forms including a determiner, i.e., the singular forms, are more frequent than the forms that do not have it, such that they are classified as determination error.

(3) **tengo planes*, instead of *tengo planes (de)* 'to have plans'

(4) **dijo secretos*, instead of *contar, revelar secretos* 'to tell secrets'

(5) **hacer decisiones*, instead of *tomar decisiones* 'to take decisions'

Number errors. Most failures to identify a number error are due to the fact that, because of multiple errors appearing in the collocation, no usable patterns are retrieved from the RC. A number of failures occur when a collocation is *per se* valid in Spanish, but incorrect in the particular context in which it is used by the learner; cf. (6) and (7).

(6) **fuimos a un museo*, compared to *fuimos a museos* 'to go to museums'

(7) **tienen razón*, compared to *tienen razones* 'to have reasons'

The same occurs in miscollocations that contain other types of errors, but are classified as number error; cf. (8) and (9). An additional source of failure in the negative case is the appearance of lexical errors in the miscollocation, as, e.g., in (10), where a wrong selection of an element of the collocation leads to a correct collocation with a different meaning.

(8) **tener los derechos*, compared to *tener el derecho* 'to have the rights'

(9) **tiene opciones*, compared to *tiene opción* 'to have options'

(10) **hacer divisiones*, compared to *causar divisiones* 'to cause separation', lit. 'to make mathematical divisions'

Gender errors. The analysis of the incorrectly classified instances of both 'gender' and 'other' miscollocations shows that the misclassification is mainly due to errors resulting from the automatic processing of the writings of the students. For instance, in the case of (11–13), the first step of our function returns no information, since all three collocations are affected by several errors and therefore, no valid patterns are retrieved from the RC. To account for this case, concordance is checked. In (11), both the determiner and the base have been assigned masculine gender, so no concordance error was found and the collocation was classified as 'other'. Similarly, *canoa* was incorrectly tagged and no concordance error was found either. Finally, in (13) a parsing error is responsible for the incorrect assignation of the class, since the determiner appears as depending on the verb.

(11) **rechazar los metas*, instead of *alcanzar, lograr las metas* 'to reach goals'

(12) **hacer el canoa*, instead of *ir en canoa* 'canoeing'

(13) **la idioma habla*, instead of *hablar un idioma* 'to speak a language'

As already (11–13), the following 'other error type' miscollocations are affected by several kinds of errors at the same time, which means that concordance has to be checked. Thus, (14) and (15) were incorrectly POS-tagged as N-Adj collocations, such that a concordance between the noun and the adjective was looked for. Since none was found, the collocations were judged to have gender errors.

(14) **sentado por sillas*, instead of *sentado en sillas* 'to sit on chairs'

(15) **completo mis clases*, instead of *termino las clases* 'to complete classes'

Government errors. An analysis of the results for this kind of error reveals that, as already with determination errors, there is often a correct version of the collocation, in this case with a different government, and it is the context which requires the selection of one or the other alternative. Thus, in (16), *tiene el poder* (whithout preposition) should not be used when followed by a verb, but is a possible expression on its own. The same occurs in (17).

(16) **tiene el poder + V*, instead of *tiene el poder (de) + V* 'to have the power (to)' + V

(17) **tener idea + V*, instead of *tener idea (de) + V* 'to have idea (of)' + V

Other types of collocation errors classified as 'government error' are usually caused by lexical errors involved in the collocation, as in the following examples. In (18), a correct collocation can be found with the given base and collocate (*resolución de este problema*). The same can be observed in (19) (*cambiar de religión*). In both cases, there is a correct collocation composed by the original base and collocate and a different preposition, which leads the function to classify them as government errors.

(18) **resolución a este problema*, instead of *solución a este problema* 'solution to a problem'

(19) **cambiar a la religión*, instead of *convertirse a la religión* 'to convert to a religion'

Pronoun errors. The lower accuracy rate for the identification of pronoun errors is due to several reasons. Firstly, due to lexical errors in the same miscollocation, almost a third of the queries to the RC does not retrieve any frequencies. Secondly, lexical errors produce combinations in Spanish that are not necessarily collocations. Thus, *sacar una operación a flote/adelante* (cf. 20) is correct, but it is not a binary collocation. Thirdly, multiple grammatical errors also give place to possible occurrences, as in (21). Finally, there are collocations that accept both the pronominal form and the bare verb form (cf. 22), where it is the context that marks one or the other use.

(20) **sacar una operación*, instead of *hacerse una operación* 'to have surgery'

(21) **aprovecharme de la oportunidad*, instead of *aprovechar la oportunidad* 'to take the most of an opportunity'

(22) **volver loco*, instead of *volverse loco* 'to go mad'

On the contrary, very few collocations of the class 'other error type' have been incorrectly classified as pronoun error. These are cases in which both the pronominal form and the bare verb form are possible, as in (23–24), or where a lexical error gives rise to an acceptable combination (25).

(23) **ir de vacaciones*, compared to *irse de vacaciones* 'to go on holidays'

(24) **cambios producido*, instead of *producirse cambios* 'produced changes'

(25) **darnos la idea*, instead of *hacernos una idea* 'to get an idea'

Order errors. Misclassified order errors are often produced when neither the original combination nor the generated alternatives are found in the RC. As seen before, this is due to multiple errors, such as in (27) and (28). Another source of error, however, can be seen in (26): the use of superlatives, which make the combinations less likely to appear in the RC.

(26) **amigas buenísimas*, instead of *buenísimas amigas* 'close friends'

(27) **nativa parlante*, instead of *hablante nativa* 'native speaker'

(28) **sumamente creo*, instead of *creo firmemente* 'to strongly believe'

As far as other types of errors that are classified as 'order error' are concerned, the most frequent reason is the case of an incorrect collocation, a reordering or appearance in the RC with a higher frequency. Thus, *el día en, buscar trabajo por* and *problemas hacen* in the following examples are acceptable combinations within a sentence.

(29) **en el día*, instead of *durante el día* 'in the day'

(30) **buscar por trabajo*, instead of *buscar trabajo* 'look for jobs'

(31) **hacen problemas*, instead of *causan problemas* 'to cause trouble'

5 Conclusions and future work

Our results show that it is possible to identify grammatical collocation errors in incorrect collocations found in the writings of foreign language learners of Spanish. Most failures to do so are due to following three main reasons: (i) errors during the automatic preprocessing of the learner and reference corpora, (ii) multiple lexical and / or grammatical errors involved in the same collocation, and (iii) valid collocations being grammatically incorrect in the given context. While (i) is not within our reach, (ii) can be partially solved by designing correction strategies that first address lexical errors and attempt to identify grammatical errors only when the lexical correction has been carried out. With respect to (iii), further research will be carried out. Our investigations show that the context in which a collocation appears is essential to identify the type of error involved. Therefore, in the future we plan to explore the use of context in more depth.

6 Acknowledgements

This work has been funded by the Spanish Ministry of Science and Competitiveness (MINECO), through a predoctoral grant with reference BES-2012-057036, in the framework of the project HARenES, under the contract number FFI2011-30219-C02-02.

References

Alonso Ramos, M., L. Wanner, O. Vincze, G. Casamayor, N. Vázquez, E. Mosqueira, and S. Prieto. 2010. Towards a Motivated Annotation Schema of Collocation Errors in Learner Corpora. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, pages 3209–3214, La Valetta, Malta.

Bahns, J. and M. Eldaw. 1993. Should we teach efl students collocations? *System*, 21(1):101–114.

Bohnet, B. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 89–97. Association for Computational Linguistics.

Bouma, G. 2010. Collocation extraction beyond the independence assumption. In *Proceedings of the ACL 2010, Short paper track*, pages 109–114, Uppsala.

Chang, Y.C., J.S. Chang, H.J. Chen, and H.C. Liou. 2008. An Automatic Collocation Writing Assistant for Taiwanese EFL learners. A case of Corpus Based NLP Technology. *Computer Assisted Language Learning*, 21(3):283–299.

Choueka, Y. 1988. Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *In Proceedings of the RIAO*, pages 34–38.

Church, K. and P. Hanks. 1989. Word Association Norms, Mutual Information, and Lexicography. In *Proceedings of the 27th Annual Meeting of the ACL*, pages 76–83.

Cowie, A. 1994. Phraseology. In R.E. Asher and J.M.Y. Simpson, editors, *The Encyclopedia of Language and Linguistics, Vol. 6*. Pergamon, Oxford, pages 3168–3171.

Dahlmeier, D. and H.T. Ng. 2011. Correcting semantic collocation errors with l1-induced paraphrases. In *Proceedings of*

the Conference on Empirical Methods in Natural Language Processing, pages 107–117. Association for Computational Linguistics.

Evert, S. 2008. Corpora and collocations. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin, pages 1212–1248.

Firth, J. 1957. Modes of meaning. In J.R. Firth, editor, *Papers in Linguistics, 1934-1951*. Oxford University Press, Oxford, pages 190–215.

Granger, S. 1998. Prefabricated patterns in advanced EFL writing: Collocations and Formulae. In A. Cowie, editor, *Phraseology: Theory, Analysis and Applications*. Oxford University Press, Oxford, pages 145–160.

Halliday, M. 1961. Categories of the theory of grammar. *Word*, 17:241–292.

Hausmann, F.-J. 1984. Wortschatzlernen ist kollokationslernen. zum lehren und lernen französischer wortwendungen. *Praxis des neusprachlichen Unterrichts*, 31(1):395–406.

Kanashiro Pereira, L.W., E. Manguilimotan, and Y. Matsumoto. 2013. Automated collocation suggestion for japanese second language learners. *ACL 2013*, page 52.

Lesniewska, J. 2006. Collocations and second language use. *Studia LingÃ$\frac{1}{4}$Ãstica Universitatis lagellonicae Cracoviensis*, 123:95–105.

Lewis, M. 2000. *Teaching Collocation. Further Developments in the Lexical Approach*. LTP, London.

Lozano, C. 2009. CEDEL2: Corpus escrito del español L2. In C.M. Bretones Callejas, editor, *Applied Linguistics Now: Understanding Language and Mind*. Universidad de Almería, Almería, pages 197–212.

Mel'čuk, I. 1995. Phrasemes in Language and Phraseology in Linguistics. In M. Everaert, E.-J. van der Linden, A. Schenk, and R. Schreuder, editors, *Idioms: Structural and Psychological Perspectives*. Lawrence Erlbaum Associates, Hillsdale, pages 167–232.

Nesselhauf, N. 2003. The use of collocations by advanced learners of english and some implications for teaching. *Applied linguistics*, 24(2):223–242.

Nesselhauf, N. 2004. How learner corpus analysis can contribute to language teaching: A study of support verb constructions. In G. Aston, S. Bernardini, and D. Stewart, editors, *Corpora and language learners*. Benjamins Academic Publishers, Amsterdam, pages 109–124.

Nesselhauf, N. 2005. *Collocations in a Learner Corpus*. Benjamins Academic Publishers, Amsterdam.

Orol, A. and M. Alonso Ramos. 2013. A Comparative Study of Collocations in a Native Corpus and a Learner Corpus of Spanish. *Procedia–Social and Behavioural Sciences*, 96:563–570.

Östling, R. and O. Knutsson. 2009. A corpus-based tool for helping writers with swedish collocations. In *Proceedings of the Workshop on Extracting and Using Constructions in NLP*, pages 28–33.

Park, T., E. Lank, P. Poupart, and M. Terry. 2008. Is the sky pure today? awkchecker: an assistive tool for detecting and correcting collocation errors. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*, pages 121–130. ACM.

Pecina, P. 2008. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 54–57, Marrakech.

Shei, C.-C. and H. Pain. 2000. An ESL writer's collocational aid. *Computer Assisted Language Learning*, 13(2):167–182.

Wible, D., C.-H. Kuo, N.-L. Tsao, A. Liu, and H.-L. Lin. 2003. Bootstrapping in a language learning environment. *Journal of Computer Assisted Learning*, 19(1):90–102.

Wu, J.-C., Y.-C. Chang, T. Mitamura, and J.S. Chang. 2010. Automatic collocation suggestion in academic writing. In *Proceedings of the ACL Conference, Short paper track*, pages 115–119, Uppsala.

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 57-64 recibido 27-03-2015 revisado 28-04-2015 aceptado 19-05-2015

P. S. Post Scriptum: Dos corpus diacrónicos de escritura cotidiana*

P. S. Post Scriptum: Two Diachronic Corpora of Ordinary Writing

Gael Vaamonde
Centro de Linguística da Universidade de Lisboa (CLUL)
Av. Prof. Gama Pinto, 2. 1649-003 Lisboa - Portugal
gaelvmnd@gmail.com

Resumen: En este trabajo se da a conocer el proyecto de investigación *P. S. Post Scriptum*, que tiene por objeto la búsqueda sistemática, edición y estudio histórico-lingüístico de cartas privadas escritas en España y Portugal durante la Edad Moderna. Estas cartas constituyen manuscritos inéditos escritos por personas de muy diferente condición social y suelen presentar una retórica cercana a la oralidad, tematizando asuntos de lo cotidiano. Son, por tanto, de gran interés para la investigación en lingüística diacrónica. La finalidad del proyecto es publicar y estudiar 7000 de estas cartas, ofreciendo una edición crítica digital del manuscrito y, simultáneamente, convirtiendo el contenido de las cartas en dos corpus anotados de un millón de palabras cada uno: uno para el español y otro para el portugués.
Palabras clave: Lingüística de corpus, lingüística histórica, español, portugués, cartas.

Abstract: In this paper, we present an overall description of *P. S. Post Scriptum*. Within this research project, systematic research will be developed, along with the publishing and historical-linguistic study of private letters written in Portugal and Spain along the Modern Ages. The letters included in *P. S. Post Scriptum* are unpublished manuscripts, written by authors from different social backgrounds. In addition, these textual resources often present an (almost) oral rhetoric, treating everyday issues of past centuries. They are, therefore, of great interest for research in Diachronic Linguistics. We aim to publish and study 7,000 of those letters. For this purpose, we are preparing a scholarly digital edition of the manuscripts and, simultaneously, converting the content of the letters into two annotated corpora of a million words each, one containing the Portuguese letters, the other the Spanish.
Keywords: Corpus Linguistics, Diachronic Linguistics, Spanish, Portuguese, Letters.

1 Introducción

La investigación en lingüística histórica no ha sido ajena al desarrollo de las nuevas tecnologías informáticas, beneficiándose en las últimas décadas –como no podía ser de otro modo– de las enormes ventajas que ofrecen los corpus en formato electrónico, que permiten almacenar y procesar grandes cantidades de datos lingüísticos de manera rápida y eficaz.

En este sentido, se puede afirmar que el español es una lengua privilegiada, ya que cuenta con dos grandes corpus históricos de acceso libre en red: el *Corpus Diacrónico del Español* (CORDE) y el *Corpus del Español* de Mark Davies (CdE). El gran volumen de texto recopilado –250 millones de palabras y 100 millones de palabras, respectivamente– los convierte en verdaderas herramientas de referencia para la investigación diacrónica en esta lengua.

* El proyecto de investigación *P. S. Post Scriptum* está siendo financiado por el Consejo Europeo de Investigación (7FP/ERC Advanced Grant – GA 295562)

ISSN 1135-5948

Por otro lado, en el ámbito hispánico han ido apareciendo recientemente otros corpus diacrónicos más especializados que, a expensas de reducir el tamaño de la muestra, permiten mejorar algunos aspectos, como son transcripciones paleográficas uniformes o la posibilidad de acceso a los facsímiles. Entre ellos cabe citar el proyecto *Biblia Medieval*[1] (Enrique-Arias, 2010), un corpus paralelo de cinco millones de palabras con las traducciones de la Biblia al castellano producidas durante la Edad Media, y el corpus CODEA[2] (Sánchez-Prieto et al., 2009), que consta de 1500 documentos anteriores al siglo XVIII editados según la triple presentación propuesta por la red CHARTA[3]: transcripción paleográfica, presentación crítica y facsímil.

Siguiendo esta línea de corpus diacrónicos especializados se sitúa el proyecto que presentamos en este trabajo: *P. S. Post Scriptum. Archivo digital de escritura cotidiana en la Edad Moderna*[4]. El objetivo de este proyecto es la creación de dos corpus compuestos por cartas privadas, uno para el español y otro para el portugués, junto con su edición crítica digital. El marco cronológico estudiado comprende desde el siglo XVI hasta el primer tercio del siglo XIX y el tamaño del corpus alcanza un total de 3500 cartas (un millón de palabras, aproximadamente) para cada lengua.

La idea que motivó la creación de *P. S. Post Scriptum* partió de una posibilidad excepcional para recuperar este tipo de material epistolar. Los tribunales de la Edad Moderna, tanto civiles como inquisitoriales, utilizaban la correspondencia privada como una prueba instrumental para condenar o exonerar a sus autores, a sus destinatarios o a otras personas relacionadas o mencionadas en el contenido de las misivas. Por tanto, buena parte de esta documentación se conservó hasta nuestros días archivada en el interior de procesos judiciales de la época.

Las cartas, en su mayoría inéditas, fueron escritas por gente de muy diversa índole, generalmente manos poco instruidas, y suelen reflejar una retórica cercana a la oralidad, ofreciendo así una ventana a variedades lingüísticas del español y del portugués que no suelen tener cabida en los corpus de corte diacrónico, compuestos predominantemente por textos de carácter literario o notarial. En otras palabras, la naturaleza dialógica y coloquial de estas misivas permite compensar, en su justa medida, la carencia de fuentes orales.

P. S. Post Scriptum es un proyecto interdisciplinar formado por lingüistas e historiadores españoles y portugueses. En este trabajo explicamos la metodología de trabajo, desde la búsqueda de los manuscritos hasta la publicación en línea de los textos, y ofrecemos el estado actual del proyecto, que finalizará en 2017.

2 Antecedentes

P. S. Post Scriptum constituye una continuación de un proyecto anterior, llamado *CARDS. Cartas Desconhecidas*. Este proyecto se centró en la recopilación y edición electrónica de cartas privadas portuguesas anteriores a 1900. El corpus pretendido en CARDS ascendía a 2000 cartas. En términos cuantitativos, por tanto, el objetivo de *P. S. Post Scriptum* es completar el corpus portugués con 1500 cartas y crear desde el inicio el corpus epistolar español.

3 Búsqueda en archivos

El primer paso en *P. S. Post scriptum* consistió en la localización, recopilación y digitalización de los manuscritos. Esta tarea fue central en los primeros años del proyecto (2012-2014) y está prácticamente concluida en el momento de redactar estas líneas. Para la localización de las cartas, se han consultado –y se están consultando– fondos judiciales (civiles y criminales), eclesiásticos e inquisitoriales a lo largo de toda la Península Ibérica.

En el caso del español, se han examinado fondos en el Archivo Histórico de Asturias, el Archivo de la Real Chancillería de Valladolid, el Archivo General de Simancas, el Archivo de la Real Chancillería de Granada, el Archivo Histórico Nacional, el Archivo General de la Corona de Aragón, el Archivo Histórico del Reino de Galicia y el Archivo General de Indias, además de varios archivos provinciales y diocesanos (Murcia, Pontevedra, Orense, Toledo, Barcelona, Guadalajara, Cuenca, Sevilla y Zaragoza).

En el caso del portugués, la documentación inquisitorial está concentrada en el Arquivo Nacional Torre do Tombo. Además del trabajo

[1] http://www.bibliamedieval.es/
[2] http://demos.bitext.com/codea/
[3] http://www.charta.es/
[4] http://ps.clul.ul.pt/index.php

continuado en este archivo, también se han consultado fondos en el Arquivo Distrital do Porto, el Arquivo Histórico Militar, El Arquivo Distrital de Braga o el Arquivo Histórico Ultramarino, entre otros.

La tarea de archivo no solo consistió en identificar las cartas, sino también en extraer toda una serie de metadatos relacionados con la producción del texto (fecha, lugar de origen y destino, descripción física del manuscrito, etc). Generalmente, la lectura atenta del proceso permite contextualizar la situación comunicativa de la carta, así como trazar un perfil biográfico de autores y destinatarios. Estas fichas biográficas están siendo almacenadas en una base de datos independiente, cuya información es posible cruzar con los datos del corpus.

El número de archivos y fondos que se han consultado es amplio y variado. En términos históricos y culturales, esta variedad permite obtener un panorama más completo de las sociedades tradicionales y de las relaciones interpersonales en la Edad Moderna, reflejadas en los contextos históricos que acompañan a cada carta o conjunto de cartas relacionadas. En términos lingüísticos, supone el control de un espacio más ampio y, por tanto, la posibilidad de incluir autores de diversa procedencia geográfica, lo que se traduce en un corpus dialectalemente más rico y representativo.

4 Transcripción en XML-TEI

Una vez localizadas las cartas, el siguiente paso es transcribirlas con el objeto de ofrecer una edición crítica digital del manuscrito, esto es, una transcripción paleográfica del texto en edición electrónica que conserve rigor filológico. Para tal fin, se han tomado algunas decisiones de carácter técnico.

Se ha utilizado el lenguaje de marcación XML (eXtensible Markup Language). Los ficheros XML son legibles, sin pérdida de información, por todos los procesadores de texto, lo que facilita su conversión para otros formatos y evita problemas de procesamiento electrónico. Por otro lado, y en consonancia con las prácticas actuales en el campo de las Humanidades Digitales, se han adoptado los estándares de codificación propuestos por el consorcio TEI (*Text Encoding Initiative*) para la edición de textos en formato digital[5]. El consorcio TEI es una convención ya consolidada en la edición virtual de fuentes primarias, lo que garantiza la integración con otros corpus electrónicos de naturaleza similar.

Conviene apuntar que al inicio del proyecto, en 2012, el consorcio TEI todavía no había proporcionado estándares de codificación para la publicación digital de material epistolar. Por este motivo, en un primer momento se adoptó la propuesta de codificación del proyecto DALF (*Digital Archive of Letters in Flanders*), que está a su vez basada en una versión no estándar del consorcio TEI. Actualmente, el modelo XML-TEI que se ofrece en *P. S. Post Scriptum* está basado en dos fuentes: la propuesta de la Red CHARTA (*Corpus Hispánico y Americano en la Red: Textos Antiguos*) y la propuesta del módulo TEI-CORRESP-SIG para material epistolar creada por Peter Stadler, Marcel illetschko y Sabine Seifert. Ambas fuentes toman como referencia la versión más actual y estandarizada del consorcio TEI.

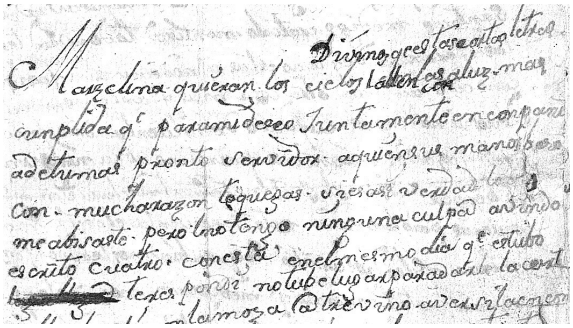

Figura 1: Facsímile de un fragmento de carta

```
<salute>Marzelina</salute> quieran los cielos <add hand="VF3"
place="supralinear">Divinos <abbr>q<expan>u</expan>e</abbr> estas cortas
letras</add> t allen <add hand="VF3" place="underlinear">con</add> la saluz mas
<lb/> cunplida <abbr>q<expan>u</expan>e</abbr> para mi deseo juntamente en
conpañi<lb n="false"/>a de tu mas pronto servidor a quien sus manos besa <lb/>
con mucha razon tu quegas si es asi verdad lo <abbr>q<expan>u</expan>e</abbr>
<lb/> me abisaste pero no tengo ninguna culpa avindo <lb/> escrito cuatro con
esta en el mesmo dia <abbr>q<expan>u</expan>e</abbr> estubo <lb/> <del
hand="VF3">la gallega</del> te respondi no tube lugar para darle la carta
```

Figura 2. Transcripción XML-TEI

Para la transcripción del manuscrito se ha adoptado una actitud conservadora. Tan solo se ha normalizado la segmentación de palabras y el uso de las grafías «i», «j», «u» y «v». Los cambios de línea, la ortografía, las abreviaturas, los tachones, las correcciones del autor, los accidentes del soporte o la orientación de la

[5] http://www.tei-c.org/index.xml

escritura, entre otros aspectos, se han respetado en la edición digital. Como ejemplo de los dicho, sirva el ejemplo recogido en las figuras 1 y 2, que permite comparar el facsímile con la transcripción en XML. En el ejemplo de la figura 2, los elementos XML </lb>, <del>, <add> y <abbr> permiten marcar cambios de línea, tachones, añadidos autoriales fuera de línea y abreviaturas, respectivamente.

5 Tratamiento lingüístico del corpus

Las tareas de transcripión y edición crítica digital forman parte del objetivo filológico del proyecto *P. S. Post Scriptum*. El otro objetivo fundamental es de carácter lingüístico: se trata de ofrecer dos corpus anotados, uno para el español y otro para el portugués.

Este segundo objetivo consta de tres tareas fundamentales: la tokenización del texto, la normalización de la grafía y la anotación lingüística de cada token normalizado. En principio, se ha contemplado para la finalización del proyecto la estandarización y anotación morfosintáctica de todo el corpus (i.e. etiquetado de clases de palabras) y la anotación sintáctica de, al menos, un subconjunto de los datos.

Desde finales de 2014, todas las tareas de tratamiento lingüístico del corpus están centralizadas en TEITOK, un sistema en línea creada por Maarten Janssen[6]. TEITOK fue diseñado para poder compatibilizar en un mismo conjunto de datos XML tanto la transcripción paleográfica como la anotación lingüística, respondiendo así a las demandas de *P. S. Post Scriptum*; cumple además con un doble objetivo: para los miembros del proyecto, funciona como ambiente de trabajo, permitiendo insertar o modificar cualquier información en los diferentes niveles de edición del texto; para el usuario, funciona como interfaz de consulta, facilitando la búsqueda cruzada de los datos que ya hallan sido almacenados[7]. A continuación, se explican brevemente cada una de las tareas de edición del texto para su procesamiento lingüístico.

5.1 Tokenización

Una vez que las cartas son transcritas mediante XML, se importan a la interfaz de TEITOK, en donde se procede al tratamiento lingüístico del texto. El primer paso es la tokenización, que se realiza de manera automática. Durante el proceso de tokenización, cada forma original de la palabra es marcada dentro de un elemento <tok>, al que se le asigna una identificación única también de manera automática.

Esta estructura inicial permite separar cada token para su posterior edición lingüística y permite salvaguardar además los diferentes niveles de edición, que se van almacenando en forma de atributos dentro de cada elemento <tok>. Por ejemplo, la forma *Otbre* como abreviatura de *octubre* en el manuscrito original sería procesada en TEITOK del modo siguiente:

```
<tok id="w-144" form="Otbre" fform="Otubre" nform="octubre">Otbre</tok>
```

Figura 3. Ejemplo de token en TEITOK

Los atributos "form", "fform" y "nform" señalan la forma original, la forma expandida y la forma normalizada de la palabra, respectivamente. Otros niveles de edición, como pueden ser variantes dialectales, información metalingüística, lemas o etiquetas morfosintácticas, también son añadidos de forma correlativa mediante atributos dentro de <tok>. Esta estrategia permite mantener siempre una vinculación entre los diferentes niveles para su posterior recuperación a través del motor de búsqueda de la interfaz.

5.2 Normalización ortográfica

Es obvio que los manuscritos originales de las cartas presentan una gran variedad ortográfica. Así, una misma palabra (p. ej. *vergüenza*) puede aparecer escrita de muy diversas formas (p. ej. *berguensa*, *verguensa*, *berguenza*, *vergüenza*, *berguença*, *verguença*, etc.). Esta diversidad tiene un interés filológico y lingüístico, principalmente para llevar a cabo estudios de carácter fonético o gráfico. Por eso, la forma original es respetada escrupulosamente y conservada en uno de los niveles de edición, como se explicó anteriormente. Esta diversidad gráfica, no obstante, constituye un problema central para la anotación automática de textos históricos (Sánchez-Marco et al., 2010). Esa es la razón principal por la que se decidió realizar una normalización ortográfica de los textos en *P. S. Post Scriptum*, que sirva como archivo de entrada para el anotador automático y maximice

[6] http://maarten.janssenweb.net

[7] Véase el apartado "Búsqueda" en la dirección electrónica del proyecto (nota 4).

su porcentaje de acierto; otra razón secundaria, además, es la posibilidad de ofrecer al público lego una edición que facilite la lectura de los textos.

En este nivel de edición, se ha normalizado la grafía y la acentuación de todas las formas originales y se ha introducido la puntuación propia de la lengua contemporánea, aunque la separación de párrafos se ha mantenido fiel al original. Conviene precisar que las modificaciones realizadas sobre el texto primario se ciñen únicamente al nivel ortográfico, por lo que no se eliminó ni se añadió ninguna palabra respecto del contenido original de la carta. Tampoco se ha intervenido sobre el nivel léxico: se han conservado los regionalismos y los arcaísmos léxicos, así como cualquier otra forma no estándar, si bien se han tratado en un nivel independiente para facilitar su recuperación.

A modo de ejemplo de normalización ortográfica, recurrimos de nuevo al fragmento ofrecido en la figura 1:

> *Marcelina, quieran los cielos divinos que estas cortas letras t' hallen con la salud más cumplida que para mí deseo, juntamente en compañía de tu más pronto servidor a quien sus manos beso. Con mucha razón te quejas, si es verdad lo que me avisaste, pero no tengo ninguna culpa habiendo escrito cuatro con esta. En el mismo día que estuvo te respondí. No tuve lugar para darle la cart' a ella.*

La edición ortográfica se está realizando de manera manual es decir, seleccionando y modificando palabra por palabra todas aquellas formas que son objeto de normalización. Aunque la interfaz de TEITOK ofrece algunas posibilidades para agilizar este proceso, se trata de una tarea que consume bastante tiempo. Por eso, en *P. S. Post Scriptum* se está trabajando actualmente en un procesamiento semiautomático de normalización. De momento, se están haciendo pruebas con la herramienta VARD 2 para el portugués (Hendrickx y Marquilhas, 2011), aunque su aplicación al proyecto todavía se encuentra en fase experimental.

5.3 Anotación lingüística

La tarea de anotación morfosintáctica ha sido objeto de un cambio de estrategia. En un primer momento, se llevó a cabo recurriendo a herramientas diferentes en función de la lengua tratada. Para el español se hizo uso del analizador automático de Freeling 3.0 (Padró y Stalinovsky, 2012) y para el portugués se utilizó la herramienta eDictor (Faria et al., 2010). En el caso de eDictor, el código de etiquetas está basado en el sistema de anotación manual utilizado por los *Penn Corpora of Historical English* (Kroch et al., 2010), ligeramente revisado para adecuarse a las características de la gramática portuguesa. En cuanto al analizador de FreeLing, el etiquetario se basa en la propuesta del grupo EAGLES para la anotación morfosintáctica de lexicones y corpus para todas las lenguas europeas. Siguiendo esta metodología, se llegaron a anotar y revisar manualmente unas 90 cartas del corpus español y unas 890 cartas del corpus portugués. Todo ese conjunto está disponible y puede ser descargado en formato TXT desde la sección “Descargas” de la página electrónica del proyecto.

Actualmente, se está utilizando FreeLing para todo el conjunto de datos, tanto españoles como portugueses, debido a sus posibilidades de configuración, a los buenos resultados que ofrece su adaptación al portugués (García Marcos y Gamallo, 2010) y a las ventajas que conlleva el empleo de una única herramienta para las dos lenguas. Además, esta decisión coincide en el tiempo con la nueva metodología de trabajo de *P. S. Post Scriptum* a través del sistema TEITOK. Así, una vez anotado cada texto con FreeLing, el resultado es importado a esta plataforma para proceder a su revisión manual. Esta importación, que se realiza automáticamente, consiste en la adición para cada elemento <tok> de dos nuevos atributos, uno para el lema y otro para la etiqueta lingüística que sugiere FreeLing.

Finalmente, una pequeña parte del corpus portugués ya ha sido anotada sintácticamente, siguiendo el sistema de anotación de los *Penn Parsed Corpora of Historical English* (Kroch et al., 2004). Se trata de un subconjunto de 260 cartas, lo que equivale a unos 90000 tokens. Este subconjunto también está disponible para descargar desde la página electrónica de *P. S. Post Scriptum*.

Téngase en cuenta que tanto la anotación morfosintáctica como sintáctica se aplican únicamente sobre las partes no formulares del texto. Es decir, en el análisis se excluye el contenido de las abertura y cierres de las cartas así como el de los segmentos formulares que aparezcan en el cuerpo del texto (arengas y peroraciones). El objetivo es conservar únicamente aquel contenido lingüístico que haya sido lo más espontáneo posible.

6 *La base de datos biográfica*

Además del tratamiento textual en sus diferentes niveles de edición, *P. S. Post Scriptum* ofrece información extratextual de diferente naturaleza. Fundamentalmente, se está recogiendo información sobre los aspectos siguientes:

- Datos contextuales de la carta: fecha, lugar de origen y destino, resumen del contenido, contextualización.
- Datos físicos del manuscrito: descripción del soporte, medidas, grafismo, estado de conservación.
- Datos biográficos de los participantes: fecha y lugar de nacimiento, ocupación, parentesco, estado civil, religión, categoría social, etc.

Los detalles biográficos de los participantes son organizados y almacenados en una base de datos XML-TEI. Esta base de datos es en principio independiente del contenido XML de las cartas; no obstante, todos los datos almacenados se pueden relacionar entre sí a través del sistema de consulta incluido en la página electrónica del proyecto *P. S. Post Scriptum*.

Respecto al autor del manuscrito de la figura 1, sabemos que se llamaba Vicente Fernández, que era vecino de Asturias, que era labrador y que fue acusado de estupro en 1789 por el padre de la destinataria, a quien había dejado embarazada. Toda esta información, obtenida a partir del proceso o de la propia carta y debidamente catalogada, puede ser usada a voluntad del usuario, ya sea con un interés histórico y cultural, ya sea para cruzarla con los datos lingüísticos del corpus. Variables como el sexo, la edad, la categoría social o la procedencia geográfica resultan de indiscutible interés para estudios sobre dialectología o sociolingüística históricas.

7 *Resultados*

Se ofrecen a continuación los resultados alcanzados en *P. S. Post Scriptum* desde el inico del proyecto en 2012 hasta el momento actual. Por lo que respecta a la localización de los manuscritos, se decidió establecer una distribución temporal que tuviese en cuenta la realidad demográfica de cada época. Esa distribución es la que sigue:

- siglo XVI: 500 cartas
- siglo XVII: 1000 cartas
- siglo XVIII: 1500 cartas
- siglo XIX: 500 cartas[8]

Teniendo en cuenta esta referencia, los resultados obtenidos hasta la fecha son los que se muestran en la tabla 1:

	español	portugués
XVI	452	283
XVII	1172	770
XVIII	1519	1016
XIX	526	784
Total	3668	2853

Tabla 1: Cartas encontradas

Como se puede apreciar, la tarea de recopilación de las misivas está ya prácticamente rematada para el corpus español. De hecho, el número total de cartas sobrepasa el límite pretendido, aunque todavía es preciso realizar una revisión general del material para descartar documentos no originales[9]. Sin lugar a dudas, la mayor dificultad para obtener fuentes se sitúa en siglo XVI. Basándonos en nuestra experiencia en archivos históricos, podemos constatar que la documentación judicial quinientista que ha sobrevivido hasta el presente es bastante inferior a la producida en siglos posteriores, lo que reduce considerablemente la posibilidad de encontrar material epistolar.

Por lo que se refiere a la transcripción en XML-TEI y a la normalización ortográfica, los datos alcanzados hasta la fecha son los recogidos en la tabla 2:

[8] La fecha extrema con la que se trabaja es 1830, de ahí que el siglo XIX se limite a 500 cartas.

[9] Algunas copias también se transcriben si se consideran especialmente interesantes como fuentes históricas, pero nunca integran el corpus anotado.

	español	portugués
transcritas	1832	1677
normalizadas	893	1042

Tabla 2. Cartas transcritas y normalizadas

Son estas dos tareas, transcripción y normalización, las que están recibiendo mayor atención actualmente, y se espera poder finalizarlas a finales del presente año.

8 *Conclusiones y trabajo futuro*

P. S. Post Scriptum tiene como objetivo crear un recurso digital para el estudio de la escritura cotidiana en España y Portugal durante la Edad Moderna (1500-1830) que responda a los intereses de varias disciplinas: la crítica textual, la lingüística histórica y la historia cultural. Con esa pretensión, se propone reunir una amplia colección de cartas privadas, ofreciéndolas en dos formatos preparados para la búsqueda: edición crítica digital y corpus anotado lingüísticamente.

El recurso es de libre acceso y ofrece al usuario toda una serie de información textual y extratextual. Actualmente, están disponibles para su consulta los siguientes aspectos:

- Digitalización del facsímil
- Edición crítica digital
- Edición normalizada del texto
- Contextualización de las cartas
- Descripción del manuscrito
- Fichas biográficas de autores y destinatarios

Toda esta información se integra en una interfaz que facilita no solo la consulta de cualquiera de los aspectos mencionados, sino también la búsqueda cruzada de los datos. Además de la opción de consulta, el usuario puede descargar libremente los archivos XML con la transcripción (<body>) y el extratexto (<header)> de cada carta, así como los achivos TXT con la parte del corpus que ya ha sido anotada.

Entre las tareas de futuro más inmediatas hay que señalar en primer lugar la necesidad de avanzar en la anotación morfosintáctica de las dos lenguas. Esto nos permitirá contar con un conjunto de datos cada vez más amplio que pueda ser reutilizado como corpus de entrenamiento. Además, téngase en cuenta que el analizador de FreeLing fue entrenado con corpus de época contemporána, por lo que progresar en la revisión manual de nuestros datos será de gran utilidad para poder evaluar el comportamiento de esta herramienta en textos históricos normalizados. Otras tareas que merecerán nuestra atención en el futuro serán el desarrollo semiautomático de la normalización ortográfica, la traducción al inglés de al menos una parte de los textos recopilados y la anotación sintáctica de un subconjunto de los datos en ambos corpus.

Bibliografía

Davies, M. 2002. *Corpus del Español: 100 million words, 1200s-1900s*. Disponible en línea en <http://www.corpusdelespanol.org>

Enrique-Arias, A. 2010. Una nueva herramienta para la investigación de fuentes bíblicas en la Edad Media: el corpus Biblia medieval. En *Actas del XII Congreso Internacional de la Asociación Hispánica de Literatura Medieva*, páginas 85-94, Cáceres, septiembre de 2007.

Faria, P., F. Kepler y M. C. de Sousa. 2010. An Integrated Tool for Annotating Historical Corpora. En *Proceedings of the Fourth Linguistic Annotation Workshop*, páginas 217-221.

García, M. y P. Gamallo. 2010. Análise Morfossintáctica para o Português Europeu e Galego: Problemas, Soluções e Avaliação. *Linguamática*, 2(2): 59-67.

Hendrickx, I. y R. Marquilhas. 2012. From old texts to modern spellings: an experiment in automatic normalisation. *Journal for Language Technology and Computational Linguistics (JLCL)*, 26(2). 65-76.

Kroch, A., B. Santorini y L. Delfs. 2004. *The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME)*. Department of Linguistics, University of Pennsylvania. CD-ROM, first edition.

Kroch, A., B. Santorini y A. Diertani. 2010. *The Penn-Helsinki Parsed Corpus of Modern British English (PPCMBE)*. Department of Linguistics, University of Pennsylvania. CD-ROM, first edition.

Padró Ll. y E. Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. En

Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA, páginas 2473-1479, Estambul (Turquía), mayo de 2012.

Real Academia Española. Banco de datos (CORDE) [en línea]. *Corpus diacrónico del español.* <http://www.rae.es>

Sánchez-Marco, C., G. Boleda, J. M. Fontana y J. Domingo. 2010. Annotation and Representation of a Diachronic Corpus of Spanish. En *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, páginas 2713-2718, Malta.

Sánchez-Prieto B., F. Paredes García, R. Martínez Sánchez, R. Miguel Franco, M. Simón Parra e I. Vicente Miguel. 2009. El Corpus de Documentos Españoles Anteriores a 1700 (CODEA). En A. Enrique-Arias (ed.). *Diacronía de las lenguas iberorrománicas: nuevas aportaciones desde la lingüística de corpus*. Madrid. Frankfurt am Main: Iberoamericana-Vervuert, páginas 25-38.

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 65-72 recibido 30-03-2015 revisado 27-04-2015 aceptado 18-05-2015

Estudio de fiabilidad y viabilidad de la Web 2.0 y la Web semántica para enriquecer lexicones en el dominio farmacológico*

Web 2.0 and Semantic Web Reliability and Viability Study to Enhance Lexicons for the Pharmacological Domain

Isabel Moreno
Dpt. Leng. y Sist. Inf.
Univ. de Alicante
Apdo. de correos, 99
E-03080 Alicante
imoreno@dlsi.ua.es

Paloma Moreda
Dpt. Leng. y Sist. Inf.
Univ. de Alicante
Apdo. de correos, 99
E-03080 Alicante
moreda@dlsi.ua.es

M. Teresa Romá-Ferri
Dpt. Enf.
Univ. de Alicante
Apdo. de correos, 99
E-03080 Alicante
mtr.ferri@ua.es

Resumen: Los actuales sistemas de Reconocimiento de Entidades en el dominio farmacológico, necesarios como apoyo para el personal sanitario en el proceso de prescripción de un tratamiento farmacológico, sufren limitaciones relacionadas con la falta de cobertura de las bases de datos oficiales. Parece por tanto necesario analizar la fiabilidad de los recursos actuales existentes, tanto en la Web Semántica como en la Web 2.0, y determinar si es o no viable utilizar dichos recursos como fuentes de información complementarias que permitan generar y/o enriquecer lexicones empleados por sistemas de Reconocimiento de Entidades. Por ello, en este trabajo se analizan las principales fuentes de información relativas al dominio farmacológico disponibles en Internet. Este análisis permite concluir que existe información fiable y que dicha información permitiría enriquecer los lexicones existentes con sinónimos y otras variaciones léxicas o incluso con información histórica no recogida ni mantenida en las bases de datos oficiales.
Palabras clave: Reconocimiento de Entidades Nombradas; Farmacología; Lexicones; Enriquecimiento; Web 2.0; Web Semántica

Abstract: Nowadays Named Entity Recognition systems in the pharmacological domain, which are needed to help healthcare professional during pharmacological treatment prescription, suffer limitations related to the lack of coverage in official databases. Therefore, it seems necessary to analyse the reliability of existing resources, both in the Semantic Web and Web 2.0, and determine whether it is feasible or not to use these resources for additional information to generate and/or enhance lexicons used by Named Entity Recognition systems. For this reason, this paper analyses the main sources of information related to the pharmacological domain available on the Internet. This analysis leads to the conclusion that there is reliable information and it would enhance existing lexicons with synonyms, variations and even historical information not collected or maintained in official databases.
Keywords: Named Entity Recognition; Pharmacology; Lexicons; Enhancement; Web 2.0; Semantic Web

1 Introducción

Hoy en día disponemos de una gran cantidad de información digital relativa a la salud. Dicha información, en su mayoría textual, se encuentra disponible en fuentes de información heterogéneas como bases de datos o enciclopedias. Emplear toda esta información resulta crítico en el ámbito sanitario (Friedman, Rindflesch, y Corn, 2013). Por ejemplo, en varios estudios se pone de manifiesto que, para el personal sanitario, la prescripción de un tratamiento farmacológico es una situación crítica y frecuente (Ely et al., 1999; Gonzalez-Gonzalez et al., 2007). La prescripción está relacionada con la selección adecuada de los medicamentos (nombre identificativo con el que se comercializan) y de sus principios activos (los componen-

* Este trabajo ha sido financiado parcialmente por la Secretaria de Estado de Investigación, Desarrollo e Innovación - Ministerio de Economía y Competitividad (TIN2012-38536-C03-03 y TIN2012-31224)

ISSN 1135-5948

tes que aportan las cualidades al medicamento). El hecho de poder consultar diferentes fuentes de información ayudaría a los profesionales en este proceso de toma de decisiones. Sin embargo, acceder y analizar toda la información textual disponible resulta: (i) inmanejable para los profesionales sanitarios (Gonzalez-Gonzalez et al., 2006); y (ii) difícil de procesar por procesos automáticos (Meystre et al., 2010; Friedman, Rindflesch, y Corn, 2013). Además, es importante destacar que cualquier fuente de información digital no oficial, requiere un proceso de validación y verificación que determine su fiabilidad.

Una línea de investigación que aborda los obstáculos aquí expuestos es el Procesamiento del Lenguaje Natural (PLN). Su finalidad es proporcionar los mecanismos necesarios para convertir la información textual, fácil de comprender por humanos, en una representación comprensible para procesos computacionales, sin importar su volumen (Friedman, Rindflesch, y Corn, 2013). Para nuestros fines, entre las diferentes tareas que se engloban dentro del PLN destaca la tarea denominada Reconocimiento de Entidades Nombradas (REN). Dicha tarea tiene como objetivo identificar aquellos elementos de información relevantes en un texto y asignarles una categoría, de entre un conjunto predefinido, para su clasificación (Feldman y Sanger, 2007). En el ámbito sanitario, y en concreto durante la prescripción de un tratamiento farmacológico, ejemplos de estas categorías podrían ser los medicamentos y los principios activos.

Como se mostró en Moreno, Moreda, y Romá-Ferri (2015), para lograr su finalidad muchos sistemas REN se apoyan en lexicones especializados, los cuales se componen de un listado de términos que representan el vocabulario habitual para cada una de las categorías predefinidas del sistema. Para que estos sistemas resulten de ayuda en el dominio farmacológico, es muy importante que los términos incluidos en estos repositorios se obtengan de fuentes fidedignas. Un ejemplo de fuente fiable para lengua castellana es la base de datos Nomenclator de Prescripción[1], donde podemos encontrar todos los medicamentos autorizados, suspendidos y revocados en España a partir de mayo de 2013, así como los principios activos que los componen.

Una particularidad de este dominio es su evolución constante, causada por el descarte o la introducción de nuevos medicamentos autorizados en cada país. Por ello, aunque la fuente de información empleada para crear el lexicón sea fiable, los sistemas REN se encuentran con una serie de obstáculos relacionados, principalmente, con su cobertura. Uno de los problemas más importantes es la carencia de sinónimos y variantes léxicas (como plurales o abreviaturas), así como una cobertura temática reducida a los términos empleados en España, en el caso de Nomenclator. Estas limitaciones influyen en los resultados que puede alcanzar un sistema REN farmacológico diseñado para el procesamiento de información en castellano (Moreno, Moreda, y Romá-Ferri, 2015). Como consecuencia, es necesario buscar fuentes de información complementarias que nos permitan superar estos problemas, sin afectar a la fiabilidad de los lexicones. Actualmente Internet proporcion una gran variedad de recursos con información farmacológica de interés. El trabajo pendiente es analizar tales recursos y determinar o no su fiabilidad a la hora de ser utilizados en tareas de REN. Por ello, el objetivo de este artículo es estudiar fuentes de información alternativas y disponibles en la red, y averiguar si permiten ampliar los lexicones creados a partir de la información disponible en Nomenclator, sin perder por ello fiabilidad.

El resto del artículo está organizado como sigue. La sección 2 describe y caracteriza las fuentes de información analizadas en este trabajo. A continuación, la sección 3 detalla cómo se ha obtenido la información de la fuente seleccionada. Después, en la sección 4, proponemos un método de validación automático para ayudar a un experto en el análisis manual. Seguidamente, en la sección 5, se describe dicho análisis manual. Terminamos con las conclusiones y el trabajo futuro en la sección 6.

2 Recursos de la Web 2.0 y la Web Semántica

En la red podemos encontrar diversos recursos o bases de conocimiento con información farmacológica de interés. Estas fuentes de información siguen dos filosofías diferentes: (i) la Web 2.0, destinada a los humanos, está organizada de forma semiestructurada, es de-

[1] `http://www.aemps.gob.es/cima/pestanias.do?metodo=nomenclator` (Último acceso: 13 Febrero 2015)

Nombre	Estructura	Tipo	Creación	Fuentes	NPA
Wikcionario	SE	D	MC	-	9
Wikipedia	SE	EN	MC	-	35
BabelNet	E	B	A	WordNet, Wikipedia, etc.	717
Wikidata	E	B	MC	-	492
DBpedia	E	B	SA	Wikipedia	921

Acrónimos: (i) A: Automático; (ii) B: Base de datos; (iii) D: Diccionario; (iv) E: Estructurado; (v) EN: Enciclopedia; (vi) MC:Manual y Colaborativo; (vii) NPA: Número de instancias clasificadas como Principo Activo; (viii) SA: Semi-Automático; (ix) SE: Semi-Estructurado.

Tabla 1: Fuentes de información de la Web 2.0 y la Web Semántica en castellano

cir, contiene información textual sin una estructura bien definida; mientras que (ii) la Web Semántica, destinada tanto a usuarios como a procesos automáticos, organiza su conocimiento de forma estructurada, es decir, la información textual contiene metadatos que facilitan los procesos automáticos. Concretamente, hemos analizado si las cinco fuentes de información más utilizadas en el dominio general, nos permitirían obtener términos que identifiquen medicamentos y principios activos, para su posterior inclusión en un lexicón para lengua castellana. En la tabla 1 encontramos un resumen de este análisis.

- Wikcionario[2]: diccionario multilinngüe colaborativo. Es un recurso semiestructurado, donde cada palabra tiene varias secciones. Para este estudio, destaca la sección de variantes pues incluye otras formas léxicas para cada entrada. Además, cada entrada puede estar asignada a una o varias categorías. A su vez, cada categoría puede dividirse en otras categorías más específicas. Estas categorías permiten seleccionar aquellas entradas del diccionario clasificadas como un principio activo: "*Categoría:ES:Fármacos*"[3], formada por 9 términos en castellano. No incluye entradas sobre medicamentos.
- Wikipedia[4]: enciclopedia multilingüe colaborativa con una gran cobertura. Cada artículo en Wikipedia está parcialmente estructurado, es decir, en su mayoría es texto libre que consta de información estructurada como las categorías y las cajas de información (ver Figura 1). En concreto, los principios activos se encuentran en la "*Categoría:Fármacos*", la cual se compone de 35 artículos y 15 subcategorías, que a su vez pueden contener más artículos y más subcategorías, en castellano. De cada caja de información, se puede obtener datos como: su nombre químico normalizado establecido por IUPAC[5] (International Union of Pure and Applied Chemistry), o su clasificación ATC[6] (Anatomical Therapeutic Chemical classification, sistema europeo de codificación de sustancias farmacéuticas y medicamentos). Dichos datos permitirían contrastar la fiabilidad de la información disponible. De nuevo, no existe una categoría que represente a los medicamentos.
- BabelNet[7](Navigli y Ponzetto, 2012): red semántica multilingüe cuyo propósito es ofrecer un diccionario enciclopédico combinando WordNet(Miller et al., 1990) y Wikipedia. Cada entrada en esta red contiene información estructurada, incluyendo un conjunto de definiciones en varios idiomas y tanto su categoría gramatical como su categoría en Wikipedia. No incluye medicamentos pero sí principios activos en español. Sin embargo, no se clasifican con la misma categoría descrita para Wikipedia, sino que usan va-

[2] `https://es.wiktionary.org` (Último acceso: 9 Marzo 2015)

[3] `https://es.wiktionary.org/wiki/Categor\%C3\%ADa:ES:F\%C3\%A1rmacos` (Último acceso: 21 Marzo 2015)

[4] `https://es.wikipedia.org` (Último acceso: 9 Marzo 2015)

[5] `http://www.iupac.org/` (Último acceso: 21 Marzo 2015)

[6] `http://www.whocc.no/atc/structure_and_principles/` (Último acceso: 21 Marzo 2015)

[7] `http://babelnet.org/` (Último acceso: 21 Marzo 2015)

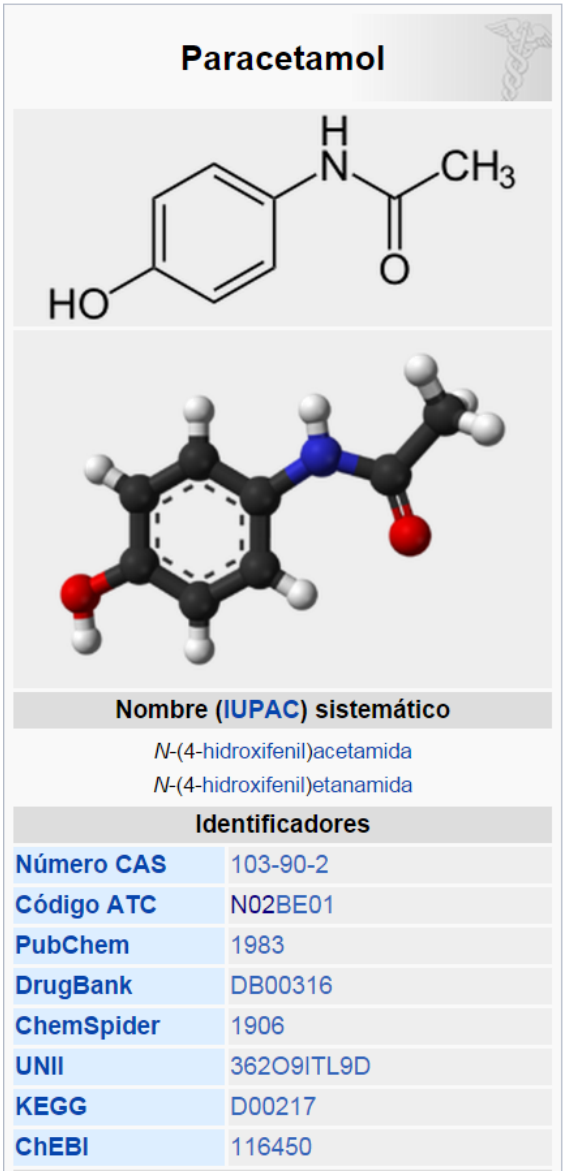

Figura 1: Fragmento de una caja de información de Wikipedia

rias subcategorias relacionadas con el nivel ATC al que el principio activo pertenece (por ejemplo:'*Categoría:Código_ATC_A*"). Contiene 717 principos activos en castellano.

- Wikidata[8](Vrandečić y Krötzsch, 2014): base de datos colaborativa, cuyas propiedades pueden ser leídas y editadas tanto por humanos como por máquinas. Proporciona una fuente común para datos factuales en Wikipedia,esto es cada caja de información en una página de Wikipedia es una entrada en Wikidata. En particular, los principios activos se encuentran en la categoría de Wikipedia "*Compuesto químico*" y se incluyen 492 instancias, independientemente del idioma. Contiene información complementaria como la vía de administración y su fórmula química. Los nombres de los medicamentos disponibles se tratan como sinónimos del principio activo. Por ello no hay manera de distinguir si Wikidata nos puede ofrecer principios activos o medicamentos.

- DBpedia[9](Lehmann et al., 2012): base de conocimiento multilingüe que extrae la información estructurada de Wikipedia. Cada recurso se mapea a una página de Wikipedia basándose en su título. Esta base de conocimiento se construye usando varios procedimientos de extracción. Uno de ellos es definir manualmente mapeos que relacionen las cajas de información de Wikipedia con la ontología DBpedia, produciendo así datos de mayor calidad. La versión española no incluye una clasificación para principios activos ni para medicamentos. Por el contrario, la versión inglesa dispone de una categoría para principios activos, "*dbpedia-owl:Drug*", aunque no de una categoría para medicamentos. Además, esta versión incluye los nombres de principios activos tanto en castellano como en inglés. De esta ontología, al igual que de Wikipedia, se pueden consultar una serie de propiedades, como el nombre de la IUPAC y su clasificación ATC, lo que permitiría contrastar su fiabilidad para los principios activos candidatos que incluye. Contiene 921 recursos clasificados como principio activo tanto en inglés como en castellano.

Tras esta revisión observamos que: (i) la mayoría de estos recursos se basan en la Wikipedia, por lo que la información facilitada puede ser fácilmente contrastada a través de los códigos ATC; (ii) todos ellos incluyen una categoría para el concepto de principio activo pero ninguno contempla el concepto medicamento; (iii) la fuente de información que contiene más principios activos actualmente es DBpedia; (iv) el acceso a DBpedia es sencillo

[8]`http://www.wikidata.org/?uselang=es` (Último acceso: 9 Marzo 2015)

[9]`http://dbpedia.org/` (Último acceso: 9 Marzo 2015)

puesto que la información está estructurada; y (v) aunque para poder acceder a la información sobre principios activos es necesario hacerlo a través de la versión en inglés de DBpedia, ésta aporta la información también en castellano. Por ello, en este trabajo nos centraremos en analizar si los principios activos incluidos en DBpedia son fiables y si es viable o no emplearlos como fuente única o bien como fuente complementaria en la creación de lexicones.

3 Obtención de principios activos de DBpedia

Como se ha comentado previamente, para obtener los principios activos de DBpedia es necesario acceder a la versión inglesa pues dispone de la categoría semántica principio activo ("*dbpedia-owl:Drug*") y, además, éstos pueden obtenerse en castellano mediante una etiqueta identificativa del idioma (séptima línea de la Figura 2). Dado que la información está estructurada, puede obtenerse fácilmente una lista de principios activos realizando una consulta sobre la base de conocimiento. Como se observa en la Figura 2, se ha empleado el lenguaje de consulta SPARQL[10] para recuperar: (i) la URI del recurso de DBpedia que representa un principio activo (por ejemplo, el principio activo "paracetamol" tiene la URI: "`http://dbpedia.org/resource/Paracetamol`"); (ii) su nombre ("*rdfs:label*") en español (ver las líneas sexta y séptima de la Figura 2); junto con (iii) su código ATC, el cual está divido en dos partes, la parte inicial del código ("*dbpedia-owl:atcPrefix*") y la parte final ("*dbpedia-owl:atcSuffix*"). Se han incluido aquellos principios activos cuyo código de la ATC ("*dbpedia-owl:atcPrefix*") comenzaba entre las letras A y V (ver la octava línea de la Figura 2) por dos razones: (i) son sustancias que en algún momento se han empleado para la prevención, curación o mejora de una enfermedad sufrida por un humano; y (ii) forman parte de la ATC de la Organización Mundial de la Salud (OMS). Ambas, nos aportan la confianza de un vocabulario estándar, lo que nos permitirá comprobar la validez de los candidatos obtenidos.

Como resultado de esta consulta SPARQL, hemos recuperado 921 nombres de principios activos en español con sus correspondientes códigos ATC. El siguiente paso será determinar la fiabilidad de la información obtenida. Este proceso de validación se detalla en la Sección 4.

[10] `http://dbpedia.org/sparql` (Último acceso: 14 Marzo 2015)

```
select ?ppio, (CONCAT(?prefijo, ?sufijo) AS ?codigo), ?nombre where
 {
   ?ppio rdf:type dbpedia-owl:Drug .
   ?ppio dbpedia-owl:atcPrefix ?prefijo.
   ?ppio dbpedia-owl:atcSuffix ?sufijo.
   ?ppio rdfs:label ?nombre.

   FILTER(langMatches(lang(?nombre), "ES"))
   FILTER(regex(?prefijo,"^[A-V]"))
 }
order by ?prefijo ?sufijo
```

Figura 2: Consulta SPARQL para recuperar principios activos de DBpedia

4 Validación automática de los principios activos recuperados de DBpedia

Cada una de los 921 instancias relativas a principios activos obtenidas de DBpedia está compuesta por su código ATC y su nombre. Para contrastar la validez de estos principios activos hemos establecido un procedimiento de comparación automático en dos fases (Secciones 4.1 y 4.2). El objetivo de dicha comparación es seleccionar aquellos principios activos o variaciones sobre ellos no existentes en el lexicón ActILex (Moreno, Moreda, y Romá-Ferri, 2015), generado a partir de Nomenclator, y por tanto, candidatos a ser incluidos como complemento. En la primera fase, denominada filtrado de códigos, se comparan los códigos ATC obtenidos de DBpedia con los de fuentes fiables con el fin de identificar principios activos no incluidos en las bases de datos oficiales. En la segunda, denominada comparación de términos, se intentan comparar nombres de principios activos para aquellos casos en los que no existe coincidencia en el código. Como resultado de este proceso se obtendrá la lista de principios activos candidatos a enriquecer ActILex y que habrán por tanto de ser validados por un experto del dominio.

4.1 Filtrado de códigos

En este paso se ha realizado una comparación entre los códigos ATC recuperados de DBpedia y los códigos ATC del lexicón ActILex (Moreno, Moreda, y Romá-Ferri, 2015). Este lexicón se genera a partir de Nomenclator, una base de conocimiento oficial. Incluye los principios activos (nombre y código) de la versión 03 2011 eliminando aquellas entradas cuyos códigos ATC comenzaban por W, X,

Y y Z, dado que especifican productos sanitarios de uso exclusivo en el sistema de atención español. A la versión actual, además, se le añadieron nuevos códigos y nombres de principios activos de la versión Nomenclator 20-02-2015, siguiendo así la misma orientación establecida en la sección 3 para DBpedia.

Cuando los códigos ATC son iguales en ambos recursos, nos encontramos ante un principio activo válido que no requiere de ningún proceso de verificación por parte de expertos. En concreto, obtenemos 789 códigos ATC coincidentes, de los 921.

4.2 Comparación de términos

Los 132 principios activos para los que no se encontró coincidencia en la fase anterior a través del código ATC, son considerados ahora por su nombre. Para ello, primero se emplea el buscador del Centro de Información online de Medicamentos[11](CIMA) de la AEMPS para verificar automáticamente tanto el nombre como el código del principio activo. Cuando los códigos eran coincidentes, se ha verificado si el nombre en ambos recursos era exactamente el mismo. En caso de encontrar coincidencia el principio activo era considerado válido y eliminado de la lista de elementos a analizar manualmente por un experto tal y como se detalla en la Sección 5.

Asimismo, un proceso automático ha confirmado si el mismo nombre de principio activo de DBpedia se encontraba en ActILex, pero con otro código. Cuando los nombres eran coincidentes, el principio activo era considerado válido y se eliminaba de la lista de elementos a verificar por un experto.

Como resultado de este proceso sólo 69 del conjunto inicial de principios activos quedaron en la lista de elementos a validar por un experto.

5 Análisis de fiabilidad de los principos activos extraídos de DBpedia

El proceso de comparación automático presentado en la sección anterior dio lugar a un conjunto de principios activos cuyo código y término de identificación no coincidía con ninguna de las fuentes de referencia consultadas, por lo que se determinó la comprobación manual por un experto. Dicha revisión se centro en determinar:

(i) Grupo de principios activos que no coincide ni con el código ni con el término de identificación. Tras la comprobación manual se encontró que 9 principios activos localizados en DBpedia no deberían ser considerados como tales. Uno de ellos por tratarse de una sustancia no reconocida por la OMS para uso generalizado (el caso de Picamilón, http://dbpedia.org/resource/Picamilon).
Los restantes 8 principios activos se descartaron por ser sustancias pertenecientes a la clasificación ATC veterinaria[12]. Cabe mencionar que estas sustancias en la versión inglesa de Wikipedia, fuente de origen para DBpedia, se identifican como tales principios activos veterinarios con su código característico, iniciado por la letra "Q". Sin embargo, en la carga de información estructurada de DBpedia esta parte inicial del código no es incluida. Un ejemplo de esta situación es el principio activo Sulfadoxina (http://dbpedia.org/resource/Sulfadoxine), en DBpedia se le asigna el código J01EQ13 mientras que su código real es el QJ01EQ13, disponible en Wikipedia.
No obstante, el experto confirmó que 4 de los principios activos sí que eran válidos, para ello utilizó otras fuentes de información, tanto de caracter nacional (Vademecum) como internacional (PubChem Classification Browser). Estos principios activos son: nafcilina, fenibut, carfilzomib y lorcaserina.

(ii) En la comprobación de principios activos con códigos iguales se detectaron aquellos con un término de identificación diferente al proporcionado por CIMA. A este nivel se localizaron 25 principios activos de DBpedia que eran o bien sinónimos o variaciones léxicas de los nombres contenidos en el lexicón propio. En concreto, el 64 % eran sinónimos (ejemplo: "Ácido glicirrícico", en CIMA y "glicirricina", en DBpedia) y el 36 % eran variantes léxicas (ejemplo de género: "cinoxacinO", en CIMA, y "cinoxacinA", en DBpedia).

(iii) El último conjunto de resultados corresponde a 24 principios activos cuyo código ATC no coincide con ActILex. En este caso se comprobó la coincidencia del término de identificación del principio activo con los términos del lexicón ActILex. La premisa que

[11] `http://www.aemps.gob.es/cima/pestanias.do?metodo=accesoAplicacion` (Último acceso: 14 Marzo 2015)

[12] `http://www.whocc.no/atcvet/atcvet/` (Último acceso: 14 Marzo 2015)

sustenta este tipo de comprobación se basa en que un término de un principio activo es asociado a uno o varios códigos, de acuerdo a su funcionalidad terapéutica. No obstante, cuando se demuestra que una funcionalidad terapéutica no es adecuada el código es eliminado; en este caso, el término sigue en vigor pero perdiendo uno de sus código de identificación. La comprobación confirmó que los 24 términos coincidían de forma completa con términos acumulados en nuestro lexicón de principios activos.

Atendiendo al análisis realizado, se puede concluir que la información facilitada por DBpedia tiene utilidad para enriquecer un lexicón especializado en principios activos en castellano empleados por sistemas REN, puesto que incluiría sinónimos, variantes léxicas y códigos obsoletos. Sin embargo, DBpedia no debería ser empleado como la única fuente para un sistema REN basado en diccionarios debido a las limitaciones que presenta, ya que por ejemplo contiene principos activos de uso veterinario.

6 Conclusiones y trabajos futuros

Este trabajo ha realizado un análisis relativo a la viabilidad de utilizar o no fuentes de información disponibles en la red y procedentes de fuentes no oficiales, para la generación y/o el enriquecimiento de lexicones que ayuden en la tarea de REN en el dominio farmacológico. Dicho análisis ha permitido establecer que si bien la generación no es fiable, puesto que contiene sustancias de uso no generalizado por la OMS y principios activos de uso veterinario; el enriquecimiento sí que sería deseable, ya que aporta sinónimos y variaciones léxicas, así como de términos de identificación de principios activos obsoletos. Estos últimos no deben de ser ignorados (Cimino, 1998), sino que deben incluirse como datos históricos para permitir así su detección, pues en algún momento fueron usadas para los tratamientos farmacoterapéuticos. El estudio ha quedado reducido a los principios activos puesto que ninguno de los recursos considerados ofrecía información sobre medicamentos.

A pesar de ello, la mejora relativa a principios activos, justifica el estudio y plantea como trabajo futuro la necesidad de definir el proceso que permita incorporar de forma automática los elementos no contenidos en el lexicón ActILex, generado a partir de Nomenclator, y enriquecerlo con sinónimos, variantes léxicas y entidades obsoletas procedentes de DBpedia. Así como cuantificar con métodos estadísticos la aportación de estas nuevas entradas en la efectividad de un sistema REN basado en diccionarios, lo que permitirá confirmar si el aumento de cobertura es significativo al emplear la versión enriquecida de ActILex.

Bibliografía

Cimino, J. J. 1998. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine - Author manuscript; available in PubMed Central 2012 August 10*, 37(4-5):394–403.

Ely, J. W, J. A. Osheroff, M. H. Ebell, G. R. Bergus, B. T Levy, M. Chambliss, y E. R. Evans. 1999. Analysis of questions asked by family doctors regarding patient care. *British Medical Journal*, 319:358–361.

Feldman, R. y J. Sanger. 2007. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press, New York, 2009 edición.

Friedman, C., T. C. Rindflesch, y M. Corn. 2013. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. *Journal of biomedical informatics*, 46(5):765–73, Octubre.

Gonzalez-Gonzalez, A. I., M. Dawes, J. Sanchez-Mateos, R. Riesgo-Fuertes, E. Escortell-Mayor, T. Sanz-Cuesta, y T. Hernandez-Fernandez. 2007. Information Needs and Information-Seeking Behavior of Primary Care Physicians. *Annals of Family Medicine*, 5:345–352.

Gonzalez-Gonzalez, A. I., J.F Sanchez Mateos, T. Sanz Cuesta, R. Riesgo Fuertes, E. Escortell Mayor, y T. Hernandez Fernandez. 2006. Estudio de las necesidadesde información generadas por los médicos de atención primaria (proyecto ENIGMA)*. *Atención primaria*, 38(4):219–224.

Lehmann, J., R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mende, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, y C. Bizer. 2012. DBpedia - A Large-scale, Multilingual Knowledge Base Ex-

tracted from Wikipedia. *Semantic Web*, 1:1–5.

Meystre, S. M., J. Thibault, S. Shen, J. F. Hurdle, y B. R. South. 2010. Textractor: a hybrid system for medications and reason for their prescription extraction from clinical text documents. *Journal of the American Medical Informatics Association*, 17(5):559–562.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, y K.J. Miller. 1990. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244.

Moreno, I., P. Moreda, y M. T. Romá-Ferri. 2015. MaNER: a MedicAl Named Entity Recogniser for Spanish. En *Proceedings of the 20th International Conference on Applications of Natural Language to Information Systems, NLDB 2015*.

Navigli, R. y S. P. Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Vrandečić, D. y M. Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM*, 57(10):78–85.

Extracción y recuperación de información monolingüe y multilingüe

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 75-82 recibido 26-02-2015 revisado 26-04-2015 aceptado 08-05-2015

Explorando Twitter mediante la integración de información estructurada y no estructurada

Exploring Twitter by Combining Structured and Unstructured Information

Juan M. Cotelo Universidad de Sevilla jcotelo@us.es
Fermín Cruz Universidad de Sevilla fcruz@us.es
F. Javier Ortega Universidad de Sevilla javierortega@us.es
José A. Troyano Universidad de Sevilla troyano@us.es

Resumen: En este artículo mostramos cómo es posible sacar partido de la información estructurada que proporciona la red social Twitter. Los textos escritos en Twitter son cortos y de baja calidad, lo que dificulta la aplicación de técnicas y herramientas que tradicionalmente se han venido usando para procesar textos en lenguaje natural. Sin embargo, Twitter ofrece mucho más que los 140 caracteres de sus mensajes para trabajar. En el ecosistema Twitter hay muchos objetos (*tweets*, *hashtags*, usuarios, palabras, ...) y relaciones entre ellos (co-ocurrencia, menciones, re-tuiteos, ...) que ofrecen innumerables posiblidades de procesado alternativo a las técnicas clásicas de PLN. En este trabajo hemos puesto nuestra atención en la tarea de clasificación de *tweets*. Sólo usando la información de la relación *Follow* hemos conseguido un clasificador que iguala los resultados de un clasificador basado en bolsas de palabras. Cuando usamos las *features* de los dos modelos, el resultado de la clasificación mejora en más de 13 puntos porcentuales con respecto a los modelos originales lo que demuestra que ambos clasificadores aportan informaciones complementarias. También hemos aplicado la misma filosofía a la tarea de recopilación del corpus con el que hemos trabajado, usando una técnica de recuperación dinámica basada en relaciones entre entidades Twitter que nos ha permitido construir una colección de *tweets* más representativa.
Palabras clave: Recuperación de tweets, clasificación de tweets, información estructurada y no estructurada

Abstract: In this paper we show how it is possible to extract useful knowledge from Twitter structured information that can improve the results of a NLP task. Tweets are short and low quality and this makes it difficult to apply classical NLP techniques to this kind of texts. However, Twitter offers more than 140 characters in their messages to work with. In Twitter ecosystem there are many objects (tweets, hashtags, users, words, ...) and relationships between them (co-occurrence, mentions, re-tweets, ...) that allow us to experiment with alternative processing techniques. In this paper we have worked with a tweet classification task. If we only use knowledge extracted from the relationship *Follow* we achieve similar results to those of a classifier based on bags of words. When we combine the knowledge from both sources we improve the results in more than 13 percentual points with respect to the original models. This shows that structured information is not only a good source of knowledge but is also complementary to the content of the messages. We also have applied the same philosophy to the task of collecting the corpus for our classification task. In this case we have use a dynamic retrieval technique based on relationships between Twitter entities that allows us to build a collection of more representative tweets.
Keywords: Tweets retrieval, tweets categorization, structured and unstructured information

ISSN 1135-5948

1 Introducción

Desde su aparición en 2006 Twitter se ha convertido, además de en un fenómeno social, en un proveedor de material de experimentación para la comunidad del Procesamiento del Lenguaje Natural. Hay infinidad de trabajos que aprovechan los escasos y de baja calidad 140 caracteres para múltiples tareas de tratamiento de textos. Entre estas tareas se encuentran la clasificación de textos (Vitale, Ferragina, y Scaiella, 2012; Schulz et al., 2014), en especial para determinar la polaridad de las opiniones siendo ésta una de las tareas sobre Twitter más estudiadas por la comunidad científica (Agarwal et al., 2011; Montejo-Ráez et al., 2014; Fernández et al., 2014; Pla y Hurtado, 2014) la extracción de *topics* (Lau, Collier, y Baldwin, 2012; Chen et al., 2013), la identificación de perfiles (Abel et al., 2011), la geolocalización (Han, Cook, y Baldwin, 2014), y muchas otras tareas cuyo objetivo es sacar información en claro desde textos escritos en lenguaje natural. Sin embargo, cuando uno se enfrenta al trabajo de leer y etiquetar *tweets* para conseguir un recurso de entrenamiento para una tarea PLN la pregunta que recurrentemente se viene a la cabeza es: ¿realmente se puede hacer PLN sobre Twitter con los problemas de cantidad y calidad que presentan sus textos? Lo cierto es que no se puede hacer un PLN de calidad si los textos con los que se trabajan son cortos, con una estructura gramatical en ocasiones poco definida, llenos de errores ortográficos o de elementos extraños (como *emoticonos* o *ascii-art*). Hay intentos de mejorar la calidad de los textos mediante técnicas de normalización (Han y Baldwin, 2011; Villena Román et al., 2013) que consiguen limpiar un poco los *tweets* de algunos fenómenos, pero estas técnicas tienen un límite y en muchos casos hay que tratar con textos que directamente "no tienen arreglo".

Estos problemas de calidad son claramente un handicap, pero afortunadamente hay maneras de resolver tareas sobre textos sin prestar mucha atención a los textos en sí. Por ejemplo, cuando Google irrumpió en 1998 con su buscador ofreciendo una solución rápida y eficaz al problema de recuperación de documentos en Internet no lo hizo porque su sistema incluyese un tratamiento de textos especialmente bueno, sino porque aprovechó los hipervínculos para evaluar la calidad de las páginas independientemente de lo que contuviesen. Es decir, usó información estructurada (los hipervínculos) como clave para resolver un problema que tenía que ver con información no estructurada (recuperar textos relacionados con una consulta). La pregunta que ha motivado este trabajo es ¿se podría hacer algo parecido con ciertas tareas sobre Twitter? La respuesta es claramente sí. Twitter, además de sus defectos en cuanto a la calidad de los textos, tiene una gran virtud: los textos están acompañados de mucha información estructurada que relaciona a múltiples entidades de distinta naturaleza. Y es posible diseñar soluciones a muchas tareas que contengan un componente no estructural (mediante el análisis del contenido de los mensajes) y otro componente estructural (mediante el análisis de los datos y relaciones asociados a los mensajes).

La Figura 1 muestra algunos de los objetos y relaciones más importantes que podemos extraer de Twitter. No están todos los objetos, y ni siquiera están todas las posibles relaciones entre los cuatro objetos destacados, pero aún así da una idea del potencial de esta información si se utiliza convenientemente.

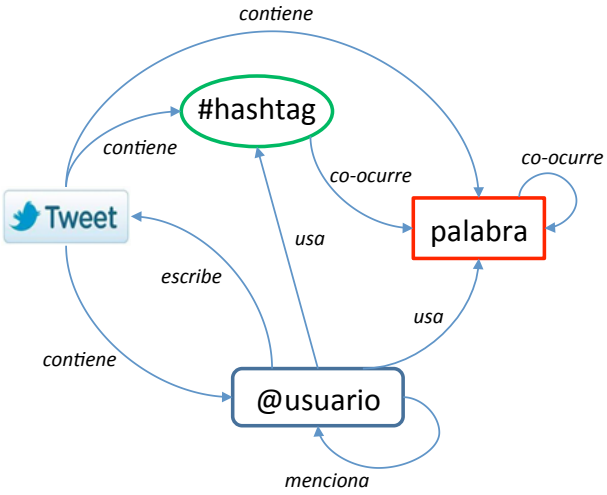

Figura 1: Algunas de las relaciones de Twitter.

En este trabajo pretendemos mostrar cómo el uso combinado de información estructurada y no estructurada beneficia la resolución de dos tareas relacionadas con Twitter: la recuperación de *tweets* y la clasificación de *tweets*. En ambos casos se usan técnicas que hacen uso tanto del contenido de los mensajes como de toda la estructura que los rodea.

El resto del trabajo se organiza de la siguiente manera. En la sección 2 se presenta el marco de trabajo en el que vamos a realizar las experimentaciones, se trata del análisis de afinidades políticas en Twitter. En la sección 3 se explica cómo se aprovecha un grafo de relaciones entre elementos de Twitter para recopilar el corpus con el que trabajaremos. En la sección 4 se muestra cómo se pueden mejorar los resultados de una clasificación de *tweets* haciendo uso de información sobre relaciones entre los autores. Por último, en la sección 5 se extraen las conclusiones.

2 El marco de trabajo: análisis de afinidades políticas en Twitter

El ámbito político es uno de los más utilizados por los estudios que usan Twitter como fuente de datos. La política siempre da que hablar y eso también se traslada a las redes sociales. Tanto los eventos importantes en el calendario político (p.e. las elecciones) como las noticias que ocurren día a día, provocan un aluvión de mensajes de personas, más o menos influyentes en la red, que se posicionan y opinan con respecto a la actualidad política. Hay trabajos que usan Twitter para resolver distintas tareas relacionadas con la política como la predicción de resultados (Tumasjan et al., 2010), la clasificación de usuarios (Pennacchiotti y Popescu, 2011) o la aplicación de técnicas de análisis de sentimientos con respecto a partidos políticos (Mejova, Srinivasan, y Boynton, 2013).

Para este trabajo hemos elegido, dentro del dominio político, una tarea de clasificación múltiple que permita determinar la postura de los mensajes con respecto a los dos grandes partidos del panorama político español: PP y PSOE. Para cada uno de los dos partidos hemos definido tres posibles categorías: a favor, en contra y neutral. Con esta configuración, cada *tweet* puede ser clasificado en una de las nueve categorías que se corresponden con el producto cartesiano de los ejes correspondientes a cada partido.

Para realizar nuestros experimentos necesitamos, en primer lugar, un corpus representativo de textos referentes a estos dos partidos políticos, que utilizaremos para evaluar distintas aproximaciones a la tarea de clasificación. Tanto el proceso de construcción del corpus, como la solución final planteada para la clasificación, nos brindan oportunidades de verificar la hipótesis principal del trabajo: el beneficio de integrar información estructurada y no estructurada a la hora de analizar los contenidos publicados en una plataforma como Twitter. En ambos casos (recuperación y clasificación de *tweets*) acabaremos apoyándonos en sendos grafos construidos a partir de ciertas relaciones de Twitter para mejorar el resultado de cada tarea. En el problema de la recuperación usaremos un grafo de usuarios y términos que nos ayudará a diseñar consultas flexibles con las que podremos adaptarnos a los nuevos temas y tendencias que continuamente aparecen en Twitter. En el caso de la clasificación demostraremos que con la ayuda de un grafo de usuarios se puede extraer información adicional que permite mejorar los resultados de la tarea.

3 Recuperación adaptativa de tweets

La manera más habitual de recopilar corpus de *tweets* es decidir un conjunto de términos (palabras, *hashtags* o nombres de usuarios) y usar la API de Twitter para lanzar consultas con esos términos. Este método es directo y se consigue recuperar todos los mensajes que se vayan escribiendo y que contengan esas palabras clave. Si esos términos son frecuentes, en poco tiempo se puede conseguir una buena colección de mensajes con la que trabajar. El problema de esta aproximación es que no tiene en cuenta una de las principales características de Twitter: la espontaneidad. Fijar de antemano el conjunto de palabras clave de las consultas nos limita sólo a los mensajes que contengan estos términos y nos podemos dejar fuera mensajes que tengan que ver con nuevos eventos que pueden aparecer en cualquier momento. En el caso de la política, si usamos un conjunto de términos predeterminados perdemos la posibilidad de capturar al vuelo términos que en un determinado momento captan la atención de la comunidad de usuarios interesada.

Nuestra aproximación a la recuperación pasa por tener consultas dinámicas, que evolucionan y que se adaptan a los temas que en cada momento toman más protagonismo en la conversación colectiva. Nuestras consultas estarán compuestas de dos tipos de términos, usuarios y etiquetas. Para no perder el foco se definen una serie de términos semilla que en nuestro caso serán las etiquetas #PP y #PSOE. A partir de ellos, y cada cierto tiempo, se construirán las consultas que incluirán

también aquellos términos que se consideren más relevantes en función del análisis de los mensajes recuperados en la consulta anterior. Este proceso contempla dos fases: construcción de un grafo de relaciones entre usuarios y etiquetas a partir de los textos recuperados en la consulta anterior, y aplicación de un algoritmo de *ranking* a dicho grafo para determinar los términos más relevantes en ese momento.

Las cuatro relaciones contempladas a la hora de construir el grafo de etiquetas y usuarios son:

- Usa: Relaciona un autor con una etiqueta. El autor del *tweet* usa la etiqueta en el texto del mensaje.
- Menciona: Relaciona un autor con otro autor. El primero de ellos usa el nombre del segundo en el mensaje.
- Re-tuitea: Relaciona un autor con otro autor. El primero de ellos reenvía un mensaje del segundo.
- Co-ocurre: Relaciona dos etiquetas entre sí. Ambas aparecen en un mismo mensaje.

Si se intenta añadir un arco ya existente al grafo, se incrementa en uno el peso de esa relación. Esta información será utilizada por el algoritmo de *ranking* para dar más importancia a las relaciones con frecuencias de aparición más altas.

La figura 2 muestra un sencillo ejemplo del tipo de grafo que se construye al procesar una secuencia de *tweets*. Todas las relaciones identificadas en cada uno de los mensajes son registradas mediante la creación de los correspondientes nodos o arcos del grafo.

Una vez que se ha construido el grafo, se le aplica una adaptación del algoritmo PageRank (Page et al., 1999) que tiene en cuenta los pesos de los arcos a la hora de calcular la relevancia de los nodos. Cada cierto período de tiempo se calcula un nuevo grafo a partir de los mensajes de la consulta anterior y se lanza una nueva consulta con los términos más relevantes según el *ranking*. El período de tiempo dependerá de la actividad de la comunidad relacionada con los términos semilla, en el caso de PP y PSOE observamos que 60 minutos era un buen período. En cuanto al número de términos del *ranking* con los que quedarnos, tras hacer varias pruebas, vimos que entre 5 y 15 se conseguía una captura de

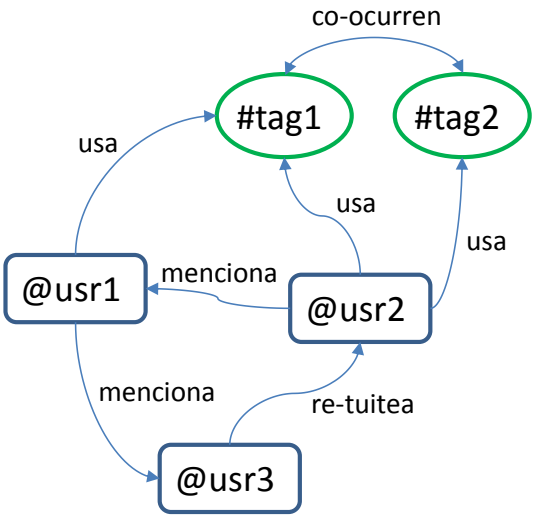

Figura 2: Ejemplo sencillo del tipo de grafo de etiquetas y usuarios usado en el proceso de recuperación.

buena calidad sin la introducción de mensajes no relacionados con las semillas y elegimos como umbral de corte el 10.

El objetivo final de esta tarea es disponer de una colección de *tweets* anotados que nos permita evaluar la siguiente tarea de clasificación. Para ello se lanzó el proceso de recuperación durante una semana y del conjunto total de mensajes se extrajo una muestra aleatoria de 3000 *tweets*. Dichos mensajes fueron anotados manualmente, registrando en cada caso la categoría a la que pertenecía de las nueve definidas en nuestro esquema.

Con idea de evaluar también el proceso de recuperación, se anotó si los *tweets* de la muestra eran relevantes, o no, para la temática PP/PSOE. El resultado de esta anotación fue que un 92,25 % de los mensajes estaban relacionados con la temática. Hay que destacar que muchos de los *tweets* no relevantes se deben a la utilización de términos de búsqueda, a priori fiables, que resultan ser ambiguos (por ejemplo la etiqueta #PP es usada por el programa de televisión chileno *Primer Plano*). Por tanto, la pérdida de precisión no es achacable en su totalidad al posible ruido introducido por el método de recuperación dinámico. Por ejemplo, en el caso de la etiqueta #PP sólo el 96,60 % de los *tweets* recuperados de forma directa con la consulta estática #PP son relevantes.

Esta pérdida de precisión del método dinámico de recuperación está acompañada con una mayor capacidad de recuperación de *tweets* interesantes, cifrada en un incremen-

to en volumen del 125,93 % con respecto a los mensajes que se hubiesen recuperado sólo con los términos semilla #PP y #PSOE mediante consultas estáticas.

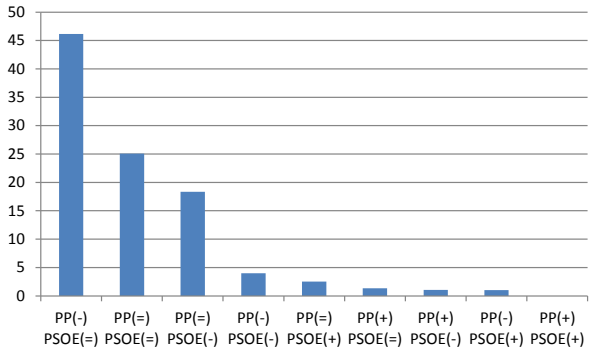

Figura 3: Distribución de porcentajes en el corpus anotado para cada una de las nueve categorías de la tarea de clasificación.

4 Clasificación de la afinidad política

En esta sección explicaremos las diferentes aproximaciones que hemos seguido a la hora de abordar la tarea de clasificación. En total hemos probado tres modelos distintos: uno basado en el contenido de los mensajes, otro basado en la estructura y otro que integra ambas informaciones.

4.1 Clasificación basada en el contenido

En la figura 3 se observa cómo se distribuye la muestra de *tweets* anotados en las nueve categorías de la tarea de clasificación. Los resultados muestran en general que la percepción de la política por parte de los usuarios de Twitter es bastante negativa. Por destacar sólo algunos datos, el 69,49 % de los mensajes son negativos con respecto a alguno de los dos partidos (el 4 % son negativos para los dos), mientras que sólo el 5,96 % son positivos con respecto a uno de los dos partidos (y no hay ningún mensaje que sea positivo para los dos).

Dado que la colección está bastante sesgada hacia las tres categorías mayoritarias (que suman el 89,54 % de los mensajes, hemos realizado dos tipos de experimentos: con todas las categorías (lo hemos denominado *Tarea-9*) y sólo con los mensajes de las tres categorías mayoritarias (*Tarea-3*).

La primera aproximación probada es la de la clasificación basada en el contenido. Con un modelo clásico de bolsa de palabras obtuvimos un resultado del 61,97 % de *accuracy* para la *Tarea-9* y de un 68,36 % para la *Tarea-3*. Los resultados se obtuvieron con un clasificador SVM y aplicando validación cruzada sobre el corpus de entrenamiento. Son resultados bastante bajos incluso para la tarea reducida en la que sólo hay que decidir entre tres categorías, lo que da idea de la dificultad del corpus. Esta dificultad de decidir en base al contenido se debe, entre otros factores, a los fenómenos usados para expresar la afinidad o rechazo a un partido, la falta de calidad de los textos y el hecho de que en muchos mensajes no haya menciones explícitas a los partidos PP y PSOE a pesar de que sí se refieren a ellos.

4.2 Clasificación basada en la estructura

Una vez que tenemos los resultados de la clasificación basada en contenidos el siguiente paso es intentar verificar la hipótesis de que la información estructurada que proporciona Twitter es de utilidad para una tarea como ésta. Analizando las distintas informaciones y relaciones disponibles, nos decidimos por aquellas relacionadas con los autores. La idea es intentar obtener un perfil de los autores de los mensajes que refleje la tendencia política del mismo. Si se asume que esta tendencia política va a ser relativamente constante, esta información puede ser de gran ayuda para complementar la que aporta el propio mensaje. Para determinar esa tendencia nos apoyamos en la relación *Follow* de Twitter, que nos da una información muy valiosa sobre los intereses de los usuarios.

A partir de los autores de los mensajes de nuestro corpus, recuperamos a todos los usuarios seguidos por éstos, lo que nos da como resultado un grafo de seguidores. Nuestra intuición era que en ese grafo hay suficiente información como para agrupar a los usuarios de nuestro corpus en grupos con la misma tendencia política.

El problema recuerda al de detección de comunidades en una red (Girvan y Newman, 2002; Shen et al., 2009) aunque las técnicas clásicas para resolver esta tarea no son directamente aplicables por la particular estructura de nuestro grafo. Por lo general estas técnicas identifican grupos de nodos densamente conectados (Ball, Karrer, y Newman, 2011) y en nuestro caso lo que necesitamos es

agrupar los nodos de los autores que siguen a un grupo similar de usuarios (llamaremos a estos usuarios *referentes*). Esta diferenciación entre los usuarios autores de mensajes y usuarios referentes nos da una estructura de grafo bipartito como la que se ilustra en el esquema de la figura 4.

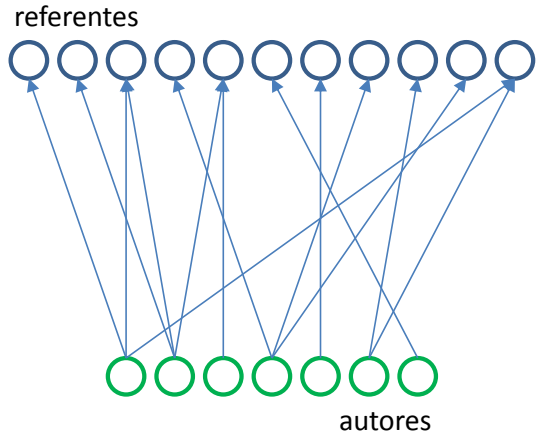

Figura 4: Ejemplo sencillo del grafo bipartito de usuarios usado en el proceso de clasificación.

Con esta topología, más que identificar si los usuarios están bien conectados, lo que nos interesa es identificar si existen patrones en la forma de seguir a los usuarios referentes. En este sentido hemos usado la matriz de adyacencia de autores y consumidores para agrupar a los autores similares en los mismos grupos. Usando técnicas de clustering hemos obtenido un total de 5 grupos de usuarios, de modo que para cada autor podemos calcular su grado de pertenencia a cada uno de estos grupos. Usando los grados de pertenencia del autor de un *tweet* como *features* para nuestra tarea de clasificación obtenemos un 59,97 % de precisión para la tarea completa y un 68,75 % para la tarea reducida. Es decir, sólo usando información sobre las personas a las que sigue el autor de un texto somos capaces de obtener resultados similares a los que obtiene el clasificador basado en bolsa de palabras.

4.3 Integración de ambos modelos

Una vez que hemos observado que la capacidad de clasificación del contenido y de las relaciones *Follow* es similar, la pregunta natural es, ¿serán complementarias? Para responderla hemos realizado dos experimentos de combinación. En el primero de ellos hemos construido un *dataset* en el que se incluye, para cada mensaje, las *features* provenientes del modelo de bolsa de palabras (experimento *BOW*) y las de la afinidad a cada una de las comunidades detectadas (experimento *Grafo*). En el segundo esquema de combinación hemos aplicado la técnica de *Stacking* (Wolpert, 1992) que permite aplicar un clasificador de segundo nivel sobre las salidas de una serie de clasificadores base. Para ello hemos usado las *features* de los modelos *BOW* y *Grafo* para entrenar a su vez a cuatro clasificadores base (SVM, *Naive Bayes*, Máxima Entropía y *Random Forests*). En la tabla 1 se muestran los resultados de los dos esquemas de combinación, junto con los de los modelos originales y los de un *baseline* que consiste en elegir la categoría más frecuente en ambas tareas de clasificación. Los resultados de todos los experimentos se han obtenido mediante validación cruzada de 10 iteraciones.

Modelo	**Tarea-9**	**Tarea-3**
Baseline	46,37 %	51,53 %
BOW	61,97 %	68,36 %
Grafo	59,97 %	68,75 %
BOW+Grafo	64,79 %	71,98 %
Stacking	75,10 %	81,14 %

Tabla 1: Resultados de *accuracy* de los diferentes modelos entrenados para la clasificación de *tweets*.

Tal y como se observa en la tabla, la información estructurada no sólo es de la "misma calidad" que el contenido de los mensajes, al obtenerse con ella resultados similares, sino que además es complementaria. Cuando se integran en un clasificador *features* de ambos modelos el resultado de la clasificación mejora a los dos clasificadores base. Para la *Tarea-9* esta mejora es de casi 3 puntos porcentuales si se integran directamente las *features* y en más de 13 puntos si se usa una técnica más sofisticada de combinación que permite sacar mayor ventaja de las visiones complementarias que aporta cada tipo de información.

5 Conclusiones y trabajo futuro

En este trabajo hemos mostrado la utilidad de la información estructurada proporcionada por un medio social como Twitter a la hora de procesar los mensajes de los usuarios. Este tipo de aproximaciones permiten que las técnicas PLN usadas a la hora de resolver

ciertas tareas sobre textos puedan ser complementadas con otras informaciones e indicios. Este apoyo es especialmente interesante a la hora de procesar contenidos escritos por usuarios que por lo general suelen ser de una pobre calidad lingüística.

Hemos elegido la tarea de clasificación de *tweets* en el ámbito político para verificar esta hipótesis. Aplicando técnicas clásicas de modelo vectorial obtuvimos una precisión del 61,97 % en la tarea de clasificar los mensajes según su afinidad o rechazo con respecto a los partidos PP y PSOE. Este resultado es prácticamente igualado (con un 59,97 %) con un clasificador que usa sólo como *features* información extraída de las relaciones entre usuarios de Twitter. La primera conclusión, por tanto, es que para una tarea de clasificación de contenidos la información estructurada de la red es casi tan útil como los mensajes en sí.

Tras experimentar con técnicas de combinación, conseguimos alcanzar el 75,10 %, superando ampliamente los resultados de los clasificadores iniciales. La segunda conclusión de nuestro trabajo es, por tanto, que el conocimiento aportado por la vía estructural resulta ser muy complementario con respecto al que se puede extraer mediante el análisis de los contenidos.

Esta misma filosofía de integración de información estructurada y no estructurada ha sido también puesta en práctica en el proceso de recuperación de *tweets* que hemos desarrollado para construir el corpus para la tarea de clasificación. En este caso hemos utilizado un método dinámico que hace uso de relaciones entre distintas entidades de Twitter para crear consultas que en cada momento se adaptan a los términos de interés con respecto a una determinada temática.

Estamos convencidos de que este tipo de aproximaciones será de utilidad en muchas tareas, no sólo sobre Twitter, sino en cualquier medio en el que se disponga informaciones de distinta naturaleza. Como línea de trabajo futuro tenemos pensado explorar este espacio de nuevas tareas, medios sociales y tipos de informaciones estructuradas disponibles en estos medios.

Agradecimientos

Este trabajo ha sido financiado a través de los proyectos ATTOS-ACOGEUS (TIN2012-38536-C03- 02) y AORESCU (P11-TIC-7684 MO).

Bibliografía

Abel, F., Q. Gao, G.J. Houben, y K. Tao. 2011. Semantic enrichment of twitter posts for user profile construction on the social web. En *The Semanic Web: Research and Applications*. Springer, pági-nas 375–389.

Agarwal, A., B. Xie, I. Vovsha, O. Rambow, y R. Passonneau. 2011. Sentiment analysis of twitter data. En *Proceedings of the Workshop on Languages in Social Media*, páginas 30–38. Association for Computational Linguistics.

Ball, B., B. Karrer, y M. Newman. 2011. Efficient and principled method for detecting communities in networks. *Physical Review E*, 84(3):36–103.

Chen, Y., H. Amiri, Z. Li, y T. Chua. 2013. Emerging topic detection for organizations from microblogs. En *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, páginas 43–52. ACM.

Fernández, J., Y. Gutiérrez, J.M. Gómez, y P. Martınez-Barco. 2014. Gplsi: Supervised sentiment analysis in twitter using skipgrams. *SemEval 2014*, páginas 294–298.

Girvan, M. y M. EJ Newman. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

Han, B. y T. Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a# twitter. En *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, páginas 368–378. Association for Computational Linguistics.

Han, B., P. Cook, y T. Baldwin. 2014. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, páginas 451–500.

Lau, J.H., N. Collier, y T. Bald-win. 2012. On-line trend analysis with topic models: \# twitter trends detection

topic model online. En *COLING*, páginas 1519–1534. Citeseer.

Mejova, Y., P. Srinivasan, y B. Boynton. 2013. Gop primary season on twitter: popular political sentiment in social media. En *Proceedings of the sixth ACM international conference on Web search and data mining*, páginas 517–526. ACM.

Montejo-Ráez, A., E. Martínez-Cámara, M. T. Martín-Valdivia, y L. A. Ureña-López. 2014. Ranked wordnet graph for sentiment polarity classification in twitter. *Computer Speech & Language*, 28(1):93–107.

Page, L., S. Brin, R. Motwani, y T. Winograd. 1999. The page-rank citation ranking: Bringing order to the web.

Pennacchiotti, M. y A.M. Popescu. 2011 Democrats, republicans and starbucks afficionados: user classification in twitter. En *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, páginas 430–438. ACM.

Pla, F. y L.F. Hurtado. 2014. Sentiment analysis in twitter for spanish. En *Natural Language Processing and Information Systems*. Springer, páginas 208–213.

Schulz, A., E. Loza Mencía, T. T. Dang, y B. Schmidt. 2014. Evaluating multi-label classifica-tion of incident-related tweets. *Making Sense of Microposts (# Microposts2014)*, páginas 26–33.

Shen, H., X. Cheng, K. Cai, y M. Hu. 2009. Detect overlapping and hierarchical community structure in net-works. *Physica A: Statistical Mechanics and its Applications*, 388(8):1706–1712.

Tumasjan, A., T. O Sprenger, P. G Sandner, y I. M. Welpe. 2010. Election forecasts with twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*, páginas 402–418.

Villena R., J., S. L. Serrano, E. Martínez Cámara, y J. C. González Cristóbal. 2013. Tass-workshop on sentiment analysis at sepln. *Procesamiento del Lenguaje Natural,* 50:37–44.

Vitale, D., P. Ferragina, y U. Scaiella. 2012. Classification of short texts by deploying topical annotations. En *Advances in Information Retrieval*. Springer, páginas 376–387.

Wolpert, D. H. 1992. Stacked generalization *Neural networks*, 5(2):241–259.

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 83-90 recibido 25-03-2015 revisado 24-04-2015 aceptado 19-05-2015

Extracción no supervisada de relaciones entre medicamentos y efectos adversos*

Unsupervised extraction of adverse drug reaction relationships

Andrés Duque
NLP Group at UNED
28040 Madrid, Spain
aduque@lsi.uned.es

Juan Martínez-Romo
NLP Group at UNED
28040 Madrid, Spain
juaner@lsi.uned.es

Lourdes Araujo
NLP Group at UNED
28040 Madrid, Spain
lurdes@lsi.uned.es

Resumen: En este trabajo se presentan los resultados preliminares de una nueva técnica no supervisada para la extracción de relaciones entre medicamentos y efectos adversos. La identificación de relaciones se consigue a partir de un modelo de representación de conocimiento que extrae pares de entidades con un peso determinado, en función de la significatividad estadística de su coaparición en un mismo documento. Dicho modelo puede ser posteriormente convertido en un grafo. El sistema ha sido evaluado sobre un corpus de referencia, denominado ADE corpus, consiguiendo resultados prometedores al obtener una eficacia muy por encima de un *baseline* estándar. Las primeras pruebas también muestran un alto potencial para inducir conocimiento nuevo.
Palabras clave: Extracción de información, Dominio médico, Extracción de relaciones, Reacciones adversas a medicamentos

Abstract: In this work we present preliminary results obtained by a new unsupervised technique for extracting relations between drugs and adverse drug reactions. The identification of those relations is achieved using a knowledge representation model that generates pairs of entities and assigns them a specific weight, depending on the statistical significance of their co-occurrence in the same document. This model may subsequently be transformed into a graph. The system has been evaluated over the reference ADE corpus, obtaining promising results, since its effectiveness is quite higher than that obtained by a standard basline. First tests also show a high potential for inducing new knowledge.
Keywords: Information extraction, Medical domain, Relation extraction, Adverse drug effect

1. Introducción

La identificación de reacciones adversas a medicamentos (*ADR: Adverse Drug Reaction*) es una problemática de gran relevancia en la práctica médica. ADR (Edwards y Aronson, 2000) se define como cualquier forma nociva, no intencionada, de reacción no deseada o desagradable que resulta del uso de una dosis de un medicamento para el propósito de la profilaxis, el diagnóstico o la terapia. Predice el peligro de la futura administración y establece la necesidad del cambio de la dosis o de la retirada del producto. Habitualmente los efectos adversos de los medicamentos no se conocen por completo en el momento de su aprobación ya que los ensayos clínicos previos son de tamaño limitado y se realizan en un período de tiempo corto. Por ello es frecuente que posteriormente aparezcan efectos adversos adicionales, en algunos casos graves. Esto hace que sea de vital importancia monitorizarlos y reportarlos en el menor tiempo posible. La extracción automática de información es por tanto de gran ayuda en este proceso, ya que puede aliviar notablemente el trabajo manual, y se está explorando, tanto en documentos científicos (Gurulingappa et al., 2012) e informes clínicos (Aramaki et al., 2010), como en información extraída de sitios web (Segura-Bedmar, de la Peña González, y Martínez, 2014).

La extracción de relaciones en general y en este caso en particular requiere realizar dos tareas. En primer lugar es necesario identificar en el texto las entidades entre las que se pueden dar las relaciones buscadas. Posteriormente se trata de identificar los casos en los que se cumple la relación entre dos entidades. En los últimos años se han aplicado técnicas de procesamiento del lenguaje natural a ambos aspectos del problema, aunque nosotros nos centramos en el segundo, en el contexto de la documentación científica.

La identificación de relaciones se ha abordado con diversas técnicas. Algunas propuestas se ba-

* Trabajo financiado parcialmente por los proyectos EXTRECM (TIN2013-46616-C2-2-R), y TwiSE (2013-025-UNED-PROY).

ISSN 1135-5948

san en la coaparición de las entidades de interés (Pyysalo et al., 2008; Kandula y Zeng-Treitler, 2010). En estos trabajos se supone que dos entidades que se mencionan en la misma frase o en el mismo resumen pueden estar relacionadas. Lógicamente, este enfoque proporciona una cobertura alta, pero una precisión muy baja. Debido a su sencillez se suele adoptar como *baseline* para hacer comparativas con otros métodos. En Wang et al. (2009) el sistema MedLEE es aplicado para identificar potenciales ADRs en resúmenes. Realizan pruebas basadas en la distribución χ^2 para seleccionar las asociaciones. Por su parte, el sistema descrito en Kang et al. (2014) utiliza una base de conocimiento para la identificación de relaciones. Concretamente se usa una representación en forma de grafo de la información contenida en el metatesauro UMLS (Lindberg, Humphreys, y McCray, 1993). UMLS define términos y conceptos, así como relaciones entre los conceptos. Estos autores utilizan distancias entre conceptos para seleccionar relaciones. Otros autores han utilizado un enfoque de aprendizaje automático (Gurulingappa et al., 2011; Gurulingappa, Mateen-Rajput, y Toldo, 2012). En (Eltyeb y Salim, 2015) se aplica un sistema basado en patrones para identificar las asociaciones. Los patrones se identifican automáticamente y después se pueden utilizar para aumentar la base de datos de relaciones. Sin embargo, no se han encontrado trabajos en la literatura que utilicen aproximaciones no supervisadas al problema.

En este trabajo nos centramos en la segunda fase de extracción de relaciones, en la que las entidades a relacionar ya han sido anotadas. En concreto, se aplica un refinamiento del modelo basado en co-ocurrencia. Como otros trabajos, suponemos que la coaparición de dos entidades en el mismo documento (en este caso, el resumen de un artículo médico) puede considerarse una indicación de una posible relación entre ellas. Sin embargo, sólo consideramos que la coaparición de dos entidades es representativa si su frecuencia es estadísticamente significativa respecto a la aparición de las entidades por separado.

El resto del artículo se organiza de la siguiente forma: en la Sección 2 se describe el corpus y la técnica utilizada para etiquetar los resúmenes. En la Sección 3 se detalla el proceso completo de extracción de relaciones, a través del modelo de representación del conocimiento propuesto. Las Secciones 4 y 5 se centran en la experimentación y el análisis de resultados. Finalmente, en la Sección 6 se extraen las principales conclusiones y se exponen las líneas de trabajo futuro.

2. Materiales y Métodos

2.1. Preparación del Corpus

Como base de conocimiento para nuestro sistema, se ha seleccionado el corpus ADE (Gurulingappa et al., 2012), que detalla relaciones entre medicamentos y efectos adversos extraídas a partir de un conjunto de 2972 resúmenes de artículos, almacenados en Medline. Del corpus inicial, construido con una base de 5063 medicamentos y 5776 condiciones médicas, son públicos un total de 1644 resúmenes, aquéllos que presentan al menos una frase describiendo un efecto adverso. En total el corpus contiene un total de 6821 relaciones de efecto adverso a un medicamento. El archivo que almacena dichas relaciones se va a utilizar en el presente trabajo como *Gold Standard* de relaciones entre medicamentos y efectos adversos, tras eliminar las relaciones repetidas. En dicho archivo, cada relación se expresa con una serie de campos, separados por el carácter "|" , que contienen respectivamente el identificador del resumen en la base de datos de Medline, la frase de la que se extrae la relación, el efecto adverso, las posiciones de inicio y final del efecto adverso en la frase, el medicamento, y las posiciones de inicio y final del medicamento en la frase. Por ejemplo, la línea *"3159106 |Allopurinol hypersensitivity. |hypersensitivity |21 |37 |Allopurinol |9 |20"* define el efecto adverso *"hypersensitivity"* provocado por el medicamento *"allopurinol"*, que puede ser encontrado en el resumen con identificador "3159106".

La Tabla 1 contiene los datos del corpus ADE original, así como los datos útiles del mismo, utilizados para construir el *Gold Standard* que se utiliza en el presente trabajo.

	Corpus	**Gold Standard**
Resúmenes	2972	1644
Medicamentos	5063	1049
Efectos Adversos	5776	2983
Relaciones	6821	5098

Tabla 1: Estadísticas del corpus ADE, en su versión original (columna **Original**), y tras extraer los elementos útiles de los resúmenes que contienen al menos una relación entre medicamento y efecto adverso.

Además del número de resúmenes, en la tabla se puede observar el número de medicamentos, efectos adversos, y relaciones no repetidas que se encuentran en el *Gold Standard*.

2.2. Anotación de los resúmenes

El corpus ADE ofrece los identificadores de Medline de los resúmenes que contienen relaciones entre medicamentos y efectos adversos. Sin embargo, para la aplicación de nuestro algoritmo es necesario etiquetar dichos resúmenes con todos los posibles medicamentos y efectos adversos que aparezcan en los mismos. Por tanto, se accede a los resúmenes vía PubMed, y se identifican sobre el texto de cada uno de dichos resúmenes aquellas entidades (medicamentos o efectos adversos) susceptibles de aparecer en una relación. Para ello, se utilizan las listas de medicamentos y efectos adversos extraídas del corpus ADE. Una vez hecho esto, cada documento quedará representado por una bolsa de entidades etiquetadas, que serán las que nos permitan elaborar el grafo de coaparición. El número total de entidades en el *Gold Standard* es de 13642, mientras que el número total de entidades etiquetadas es de 25687. Esto nos da una idea de la dificultad de encontrar las relaciones correctas de entre todas las posibles combinaciones entre entidades dentro de los documentos etiquetados.

Tal y como se ha adelantado en la Sección 1, es importante destacar que el objetivo fundamental de este trabajo es analizar la utilidad de la técnica basada en coaparición de entidades para la extracción de relaciones entre medicamentos y efectos adversos. Es decir, no nos centramos en la eficacia de una técnica de etiquetado concreta, sino que consideramos un etiquetado hipotéticamente perfecto, en el que se anotan todas las entidades que nos interesan, para analizar el comportamiento del sistema propuesto. Una técnica de etiquetado diferente introduciría un sesgo que es el que se pretende evitar a través del etiquetado propuesto.

3. Modelo de extracción de relaciones (significatividad estadística)

El siguiente paso consiste en el análisis de coaparición de entidades, a partir de los documentos etiquetados.

Para comprobar si la coaparición de dos entidades en un documento es significativa, se define un modelo nulo en el que las entidades se distribuyen aleatoria e independientemente entre un conjunto de documentos de un corpus. Concretamente, se calcula la probabilidad de que dos entidades coincidan por puro azar. Este valor nos permite determinar un p-valor p para la coaparición de dos entidades. Si $p \ll 1$ se puede considerar que la aparición de las dos entidades en el mismo documento es significativa, y por lo tanto, es probable que su significado esté relacionado.

Concretamente, si dos entidades se encuentran respectivamente en n_1 y n_2 documentos, de entre los N que componen el corpus, para contar cuantos casos existen en los que dos entidades coincidan en exactamente k documentos, debemos tener en cuenta que hay cuatro tipos de documentos: k documentos que contienen ambas entidades, $n_1 - k$ documentos que contienen sólo la primera entidad, $n_2 - k$ documentos que contienen sólo la segunda entidad, y $N - n_1 - n_2 + k$ documentos (siempre que este número no sea cero) que no contienen ninguna de las dos entidades. Por lo tanto, el número de disposiciones que buscamos viene dado por el coeficiente multinomial:

$$\binom{N}{k, n_1 - k, n_2 - k} \tag{1}$$

Así, la probabilidad de que dos entidades que aparecen en los documentos n_1 y n_2 respectivamente y que están distribuidas de forma aleatoria e independiente entre N documentos, coincidan en exactamente k de ellos viene dada por:

$$p(k) = \binom{N}{n_1}^{-1} \binom{N}{n_2}^{-1} \binom{N}{k, n_1 - k, n_2 - k}, \tag{2}$$

si $\max\{0, n_1 + n_2 - N\} \leq k \leq \min\{n_1, n_2\}$ y cero en otro caso.

Podemos escribir la ecuación (2) de una forma más fácil de tratar computacionalmente. Para ello introducimos la notación $(a)_b \equiv a(a-1)\cdots(a-b+1)$, para cualquier $a \geq b$, y sin pérdida de generalidad suponemos que la primera entidad es la más frecuente, es decir $n_1 \geq n_2 \geq k$. Entonces:

$$\begin{aligned} p(k) &= \frac{(n_1)_k (n_2)_k (N - n_1)_{n_2 - k}}{(N)_{n_2} (k)_k} \\ &= \frac{(n_1)_k (n_2)_k (N - n_1)_{n_2 - k}}{(N)_{n_2 - k} (N - n_2 + k)_k (k)_k}, \end{aligned} \tag{3}$$

donde en la segunda forma se ha usado la identidad $(a)_b = (a)_c (a-c)_{b-c}$ valida para $a \geq b \geq c$. La ecuación (3) se puede reescribir como

$$\begin{aligned} p(k) = &\prod_{j=0}^{n_2 - k - 1} \left(1 - \frac{n_1}{N - j}\right) \\ &\times \prod_{j=0}^{k-1} \frac{(n_1 - j)(n_2 - j)}{(N - n_2 + k - j)(k - j)} \end{aligned} \tag{4}$$

Esto nos permite determinar un p-valor para la coaparición de dos entidades como

$$p = \sum_{k \geq r} p(k), \tag{5}$$

donde r es el número de documentos en el corpus en el que coaparecen las dos entidades.

La forma de proceder a partir de este punto es la usual en la comprobación de hipótesis estadísticas: se establece un nivel de confianza p_0 (habitualmente $p_0 \leq 0{,}05$, es decir, la hipótesis nula es incorrecta con un nivel de confianza del 95 % o superior) de manera que la coaparición es significativa sólo si $p < p_0$. De acuerdo con esto se define un par de entidades i y j relacionadas sólo si coaparecen de acuerdo con este criterio. Pero cuanto más bajo sea el valor de p más significativa es la coaparición, por lo que tiene sentido asignar un peso a esta relación. Esta significatividad se puede cuantificar tomando la mediana (correspondiente a $p = 1/2$) como una referencia y calculando el peso como $\ell = -\log(2p)$, es decir una medida de cuanto se desvía de la mediana el valor real de r (número de documentos en el corpus en los que coaparecen las dos entidades).

Mediante esta técnica se extrae un modelo de representación del conocimiento en el que se almacenan los pares de entidades que coaparecen de forma significativa. Estos pares de entidades se conectan con un peso que mide la significatividad de su coaparición. En este caso estamos interesados únicamente en aquellos pares de entidades que conecten medicamentos (entidades almacenadas con la etiqueta "MED") con efectos adversos (entidades almacenadas con la etiqueta "DIS"). De las entidades que están formadas por varias palabras se eliminan las denominadas *stopwords*, o palabras del propio idioma (en este caso inglés) que no aportan información.

4. Experimentación y resultados

La evaluación se ha realizado en términos de Precisión, Cobertura y Medida-F (*Precision*, *Recall* y *F-Measure*).

4.1. Baseline

Dentro de la evaluación de nuestro sistema, se ha desarrollado un *baseline*, obtenido mediante una técnica simple, que define un umbral de resultados a superar por el algoritmo propuesto. Dicho *baseline* se obtiene considerando que todas las entidades que aparecen en un mismo resumen están relacionadas, siempre que una de las entidades sea un medicamento y la otra un efecto adverso. El *Gold Standard* se construye a partir de frases que se encuentran en los resúmenes, es decir, toda relación que se encuentra en el *Gold Standard* se extrae de uno o varios documentos concretos. El *baseline* ofrece por tanto una cobertura perfecta, aunque su precisión es muy baja ya que contiene muchas relaciones incorrectas.

4.2. Resultados iniciales

La Tabla 2 muestra el *baseline*, así como los resultados obtenidos por la primera aproximación de nuestro algoritmo, descrito en la Sección 3.

Sistema	P	C	F
Baseline	25,20	**100,00**	40,25
Propuesto	**42,33**	59,67	**49,53**

Tabla 2: Resultados en función de la Medida-F (F), Precisión (P) y Cobertura (C) obtenidos por nuestro algoritmo, en comparación con el *baseline*. Los campos en negrita indican el mayor valor de cada medida.

Tal y como se indicaba anteriormente, el *baseline* consigue una cobertura perfecta del problema, sin embargo, la precisión (número de aciertos partido por el número total de relaciones propuestas) de nuestro algoritmo es mayor, lo que redunda en una mayor Medida-F. Es decir, aunque nuestro sistema no encuentra todas las relaciones posibles (encuentra alrededor de un 60 % de ellas), la proporción de aciertos es mayor, lo cuál nos indica que nuestro algoritmo está encontrando con mayor facilidad aquellas relaciones más significativas, desechando con efectividad otras que una técnica como la que implementa el *baseline* asignaría por el simple hecho de aparecer en un mismo documento. El umbral de significatividad estadística utilizado en nuestro algoritmo es de $P_0 = 0{,}01$, es decir, se aplica un nivel de confianza del 99 %.

4.3. Análisis de distancias

Tras los primeros resultados se realizó un análisis detallado de las relaciones obtenidas por el sistema. La primera impresión fue que existían numerosas relaciones redundantes, en comparación con aquéllas contenidas en el *Gold Standard*. Si consideramos las relaciones $R_1(M_1, E_1)$ y $R_2(M_2, E_2)$, donde M_i y E_i representan, respectivamente, el medicamento y el efecto adverso contenidos en la relación i, existen casos para los que E_1 y E_2 son dos formas diferentes de definir el mismo efecto adverso, es

decir, contienen palabras muy similares, aunque no son exactamente iguales. Por ejemplo, uno de los efectos adversos del medicamento *"itraconazole"* se define como *"vanishing bile duct"*, así como *"vanishing bile duct syndrome"*, es decir, se contabilizan como dos relaciones pero únicamente difieren en una palabra y representan el mismo efecto adverso. Si observamos el *Gold Standard* nos damos cuenta de que la relación correcta sería entre el medicamento *"itraconazole"* y el efecto adverso *"vanishing bile duct syndrome"*.

La consecuencia de que se produzca esta situación, en términos de la evaluación del sistema, es que la precisión (y por tanto la Medida-F) disminuye debido al número de relaciones extraídas por el sistema, que aunque no existan en el *Gold Standard*, su efecto adverso es equivalente al de una relación que sí se encuentra en él (con el mismo medicamento). Para solucionar estos casos, se consideran aquéllas relaciones $R_1(M_1, E_1)$ y $R_2(M_2, E_2)$, obtenidas por el sistema, para las cuales $E_1 \equiv E_2$. Puesto que E_1 y E_2 son términos compuestos por una o más palabras, basamos esta equivalencia en la cercanía entre ambos números, según la medida de similitud de Jaccard, la cual, dados dos conjuntos, se expresa mediante la siguiente fórmula:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}, \qquad (6)$$

donde, en nuestro caso A y B son las dos entidades (efectos adversos) que queremos comparar. Tras una serie de pruebas, se ha optado por establecer un valor mínimo de $J = 0{,}6$ para considerar que dos entidades se refieren al mismo efecto adverso. Este valor es lo suficientemente elevado como para que las equivalencias que introduce sean correctas y la precisión no disminuya. Una vez que se analiza la distancia entre efectos adverso a un mismo medicamento, se unen en una sola relación aquéllos efectos adversos con similitud superior al umbral. La Figura 1 muestra el resultado de aplicar el algoritmo de similitud sobre un subconjunto de efectos adversos provocados por el mismo medicamento ("Dalteparin").

Como podemos observar, los efectos adversos que se han unido en una sola entidad utilizan diferentes definiciones para representar a dicha entidad. En este caso, se considera que una relación del *Gold Standard* ha sido encontrada por nuestro sistema, para un medicamento concreto, si existe un efecto adverso dentro del conjunto de efectos adversos similares para ese medicamento, según la distancia Jaccard, cuya definición coincide exactamente con el efecto adverso de la relación del *Gold Standard*. Cada relación encontrada por el sistema, se considera a efectos de evaluación como una sola instancia.

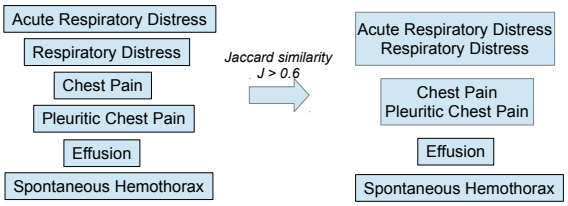

Figura 1: Aplicación de la medida de similitud Jaccard sobre los efectos adversos del medicamento *"Dalteparin"*.

La Tabla 3 muestra los resultados una vez aplicada la similitud de Jaccard entre efectos adversos, tanto a las relaciones obtenidas anteriormente como a las obtenidas por el *baseline*.

Sistema	P	C	F
BaseJac	27,11	**100,00**	42,65
PropJac	**45,46**	59,67	**51,60**

Tabla 3: Resultados tras aplicar la similitud de Jaccard sobre los efectos adversos, en función de la Medida-F (F), Precisión (P) y Cobertura (C). Comparación entre nuestro algoritmo (**PropJac**) y el *baseline* (**BaseJac**). Los campos en negrita indican el mayor valor de cada medida.

La tabla muestra un aumento de la Precisión y la Medida-F, gracias a la reducción de relaciones propuestas por el algoritmo, mientras que la Cobertura se mantiene constante, es decir, se sigue encontrando el mismo número de relaciones correctas que en los resultados iniciales.

5. *Inducción de conocimiento*

Uno de los aspectos más importantes de un sistema que busca relaciones dentro de textos, ya sea entre medicamentos y efectos adversos, como es el caso, o entre cualquier otro tipo de entidades, es su capacidad de descubrir nuevas relaciones que no se conociesen anteriormente, es decir inducir conocimiento nuevo. En el presente trabajo, es importante conocer si el sistema propuesto sería capaz de encontrar relaciones "nuevas", que no estén directamente basadas en evidencias que se puedan encontrar en los textos de partida. Dichas evidencias serían resúmenes concretos en los que aparezcan frases relacionando directamente medicamentos con efectos adver-

sos. En esta sección detallaremos los experimentos llevados a cabo para comprobar si nuestro sistema es capaz de realizar esta inducción de conocimiento.

El modelo de representación de conocimiento descrito en la Sección 3 nos permite obtener un conjunto de pares de entidades, relacionadas entre sí con un determinado peso. A partir de estos datos es sencillo construir un grafo en el que cada nodo sea una entidad (medicamento o efecto adverso), y un enlace entre dos nodos represente la significatividad estadística de coaparición de dichas entidades. El valor del enlace será el del peso almacenado para el par de entidades. Visualizando el modelo de representación de conocimiento como un grafo, los resultados obtenidos en la Sección 4.2 vendrían dados por las relaciones extraídas al recorrer sólo los enlaces directos entre medicamentos y efectos adversos del grafo, es decir, aquellos caminos de distancia $d = 1$ entre un medicamento y un efecto adverso. Sin embargo, es posible que si recorremos el grafo con mayor profundidad, encontremos nuevas relaciones. Sin embargo, es indispensable establecer condiciones para evitar que el número de relaciones propuestas aumente demasiado, comprometiendo la precisión del sistema (aunque se aumente su cobertura), y por tanto, la Medida-F. De acuerdo a esto último, en este caso se extraen las relaciones que se extraían anteriormente, y además, se añaden aquéllas en las que el medicamento y el efecto adverso se encuentran a dos pasos (enlaces) de distancia dentro del grafo. Además, consideramos que la entidad intermedia en dicho camino ha de ser otro efecto adverso: $M \Rightarrow E_1 \Rightarrow E_2$. Esta restricción parece lógica, basándonos en la hipótesis de que si E_1 y E_2 coaparecen con frecuencia, hay una probabilidad alta de que si M provoca E_1, pueda provocar también E_2.

Una vez que se generan nuevas relaciones gracias al grafo de coaparición, queremos saber si alguna de esas nuevas relaciones podría representar un caso de inducción de conocimiento. Como la única manera de saber si una relación es correcta sin recurrir a expertos del dominio es el *Gold Standard*, el proceso que se sigue para determinar si existe inducción de conocimiento es el siguiente:

1. Se extraen las nuevas relaciones, recorriendo el grafo con $d = 2$.
2. Se selecciona una relación concreta $R(M, E)$, que sea correcta según el *Gold Standard*, y se buscan aquéllos resúmenes que apoyen dicha relación. Es decir, se seleccionan los resúmenes en los que coaparezcan directamente el medicamento M y el efecto adverso E.
3. Se eliminan los resúmenes que apoyan la relación, y se vuelve a construir el grafo de coaparición.
4. Se vuelven a extraer las relaciones, con $d = 2$, del nuevo grafo de coaparición.
5. Si se vuelve a encontrar la relación $R(M, E)$, el conocimiento representado por dicha relación se ha inducido, sin que haya un resumen específico que apoye la existencia de dicha relación.

Estos pasos se han seguido para analizar varias relaciones positivas. Algunas de ellas se volvían a encontrar al final del proceso (se inducía el conocimiento), mientras que otras no. En este punto, nos interesa conocer si el peso de las relaciones juega un papel importante a la hora de inducir el conocimiento. Nuestra hipótesis se basa en que si el peso normalizado de una relación (el peso directo en el grafo, dividido entre la suma de los pesos de todos los enlaces que parten del nodo) es alto, dicha relación será más fuerte y por tanto el conocimiento que representa será más fácil de inducir. La forma de obtener el peso normalizado de un enlace entre el nodo i y el j, $P(i, j)$, se muestra en la fórmula 7:

$$P(i,j) = \frac{D(i,j)}{\sum_{k=1}^{O(i)} D(i,k)}, \qquad (7)$$

donde $D(i, j)$ es el peso directo (sin normalizar) entre los nodos, y $O(i)$ es el número de enlaces que parten del nodo i (su *"Outdegree"*).

Por ejemplo, a partir del medicamento *"methotrexate"* encontramos tres relaciones diferentes, con distancia $d = 2$, que se vuelven a encontrar al final del proceso anterior (su conocimiento se induce):

- $methotrexate \Rightarrow toxicity(0{,}05) \Rightarrow renal\ toxicity(0{,}48)$
- $methotrexate \Rightarrow sarcoma(0{,}05) \Rightarrow nodules(0{,}03)$
- $methotrexate \Rightarrow arthritis(0{,}31) \Rightarrow nephropathy(0{,}21)$

Los números situados a la derecha de los efectos adversos nos indican el peso normalizado de la relación entre la entidad anterior y el efecto adverso. De las tres relaciones mostradas, únicamente la última relación presenta ambos pesos

normalizados relativamente elevados (uno representa el 30 % de los pesos y el otro el 20 %, aproximadamente). Por tanto, podemos establecer como condición, que si se cumplen las restricciones $P(M_1, E_1) > 0{,}3$ y $P(E_1, E_2) > 0{,}2$ entonces existe $R(M_1, E_2)$ (el sistema considera la relación como correcta).

Una vez determinados estos umbrales, se vuelven a considerar todos los resúmenes para construir el grafo de coaparición. Esta configuración del sistema se ha denominado "Normalizada 1" o "N1".

Nos interesa, igualmente, realizar una prueba en el que se restrinja al máximo uno de los dos pesos, en este caso el peso normalizado entre E_1 y E_2, obligando a que el peso de dicha relación sea igual a 1. La interpretación de este peso sería una relación directa entre dos efectos adversos, en la que E_1 únicamente está conectado con E_2, y por tanto podemos suponer que la ocurrencia del efecto E_1 provoca en todos los casos que ocurra también el efecto E_2. Esta configuración del sistema se denomina "Normalizada 2" o "N2".

Por último, para comprobar el comportamiento base del grafo de coaparición con $d = 2$, consideramos la configuración del sistema "Normalizada 0" o "N0". En esta configuración no se restringen los pesos normalizados, sino que se generan todas las posibles relaciones $R(M, E_2)$ y se añaden a las obtenidas en los resultados iniciales.

La Tabla 4 contiene los resultados de todas las configuraciones del sistema, en términos de Precisión, Cobertura y Medida-F. Se ha añadido el número de relaciones correctas encontradas (*True Positives*) para ilustrar en términos absolutos el comportamiento de los sistemas.

Se observa que la configuración **N0** (sin restricciones en los pesos) obtiene el mayor valor de cobertura posible, encontrando más relaciones que ninguna otra. Sin embargo, el número de relaciones totales que obtiene es demasiado elevado, por lo que su precisión y Medida-F finales son muy pequeñas. La configuración **N1** consigue un compromiso entre precisión y cobertura que provoca que la Medida-F se mantenga similar a la conseguida por la configuración inicial. El aspecto importante de esta configuración es la inducción de conocimiento que se produce, tal y como se ha mostrado anteriormente. Finalmente, la configuración **N2** presenta el mejor valor de la Medida-F, aunque en este caso no se han encontrado casos en los que se produzca inducción de conocimiento.

En la tabla también se incluyen como referencia los valores de precisión, cobertura y Medida-F obtenidos por otros sistemas: un sistema (Gurulingappa, Mateen-Rajput, y Toldo, 2012) basado en máquinas de soporte vectorial (**JSRE**), otro sistema supervisado, aunque con tamaños pequeños del conjunto de entrenamiento, que utiliza bases de conocimiento (Kang et al., 2014), identificado en la tabla como **KB**, y un sistema basado en correspondencia de patrones (Eltyeb y Salim, 2015), también supervisado e identificado en la tabla como **PB**. Aunque los resultados ofrecidos por nuestro sistema quedan lejos de los obtenidos por dichos sistemas supervisados, estas técnicas requieren de unos recursos determinados para la fase de entrenamiento que el sistema propuesto en este trabajo no necesita.

Sistema	P	C	F	TP
Ini	**45,46**	59,67	51,60	3042
N0	10,45	**80,25**	18,48	**4091(*)**
N1	42,26	61,46	50,08	3133(*)
N2	45,12	60,34	**51,63**	3076
JSRE	86,00	89,00	87,00	—
KB	91,80	86,10	88,80	—
PB	93,60	72,80	81,70	—

Tabla 4: Resultados finales del sistema y comparación con otros sistemas. Se comparan los valores de cobertura (C), precisión (P) y Medida-F (F), expresados en porcentaje, así como el total de relaciones correctas encontradas (TP), sobre las cuatro configuraciones de nuestro sistema: inicial (**Ini**), normalizada 0 (**N0**), normalizada 1 (**N1**) y normalizada 2 (**N2**). Los campos en negrita indican el mayor valor de cada medida; el asterisco indica que se ha producido inducción de conocimiento. Se comparan nuestros resultados con otros sistemas, en este caso supervisados (ver texto).

6. Conclusiones y Trabajo Futuro

La técnica de extracción de relaciones descrita en este trabajo ofrece mejoras significativas en relación al *baseline* propuesto, lo cual nos indica que nuestro sistema discrimina correctamente aquéllas coapariciones de medicamentos y efectos adversos susceptibles de convertirse en una relación. La eficacia del sistema tiene un elevado margen de mejora, como se puede observar en la cobertura potencial que se podría alcanzar utilizando el grafo de coaparición (Tabla 4). El análisis de las relaciones que se podrían extraer a través de exploraciones más profundas del grafo (con valores de d mayores que 2) es una de las

cuestiones más inmediatas a analizar. También se analizarán patrones que permitan diferenciar las relaciones que nos interesan, de aquéllas que representan medicamentos aplicados como tratamiento a enfermedades específicas. En lo relativo al análisis de distancias, se explorarán otras técnicas orientadas a la extracción de equivalencias entre enfermedades, como por ejemplo el uso de variaciones léxicas propuestas por bases de datos médicas como SNOMED (Donelly, 2006).

Es importante destacar que las relaciones que se extraen en este trabajo se cotejan con un *Gold Standard* determinado, extraído a partir de un corpus que obviamente, no contiene todas las posibles relaciones existentes entre medicamentos y efectos adversos. Por tanto, es posible que exista conocimiento inducido que responde a relaciones correctas, pero que no se pueden evaluar como tales por falta de documentos médicos que las apoyen. En este sentido, sería útil realizar búsquedas en diversas fuentes, como bases de datos que proporcionen artículos médicos, para comprobar si existen evidencias de que una relación encontrada en el grafo pero no presente en el corpus ADE, puede ser igualmente correcta.

Bibliografía

Aramaki, E., Y. Miura, M. Tonoike, T. Ohkuma, H. Masuichi, K. Waki, y K. Ohe. 2010. Extraction of adverse drug effects from clinical records. *Studies in health technology and informatics*, 160(Pt 1):739–743.

Donelly, K. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, 121:279–290.

Edwards, I. R. y J. K. Aronson. 2000. Adverse drug reactions: definitions, diagnosis, and management. *The Lancet*, 356(9237):1255 – 1259.

Eltyeb, S. y N. Salim. 2015. Pattern-based system to detect the adverse drug effect sentences in medical case reports. *Journal of Theoretical and Applied Information Technology*, 71(1):137–143.

Gurulingappa, H., J. Fluck, M. Hofmann-Apitius, y L. Toldo. 2011. Identification of adverse drug event assertive sentences in medical case reports. En *Proceedings of First international workshop on knowledge discovery and health care management (KD-HCM)*, páginas 16–27, Athens.

Gurulingappa, H., A. Mateen-Rajput, y L. Toldo. 2012. Extraction of potential adverse drug events from medical case reports. *Journal of biomedical semantics*, 3(1):15.

Gurulingappa, H., A. Mateen Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, y L. Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5):885 – 892.

Kandula, S. y Q. Zeng-Treitler. 2010. Exploring relations among semantic groups: a comparison of concept co-occurrence in biomedical sources. *Studies in health technology and informatics*, 160(2):995–999.

Kang, N., E.M. van Mulligen, B. Singh, C. Bui, Z. Afzal, y J. Kors. 2014. Knowledge-based extraction of adverse drug events from biomedical text. *BMC bioinformatics*, 15(1):64.

Lindberg, D., B. Humphreys, y A. McCray. 1993. The unified medical language system. *Methods of Information in Medicine*, 32(4):281–291.

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

Segura-Bedmar, I., S. de la Peña González, y P. Martínez. 2014. Extracting drug indications and adverse drug reactions from spanish health social media. En *Proceedings of BioNLP 2014*, páginas 98–106, Baltimore, Maryland, June. Association for Computational Linguistics.

Wang, X., G. Hripcsak, M. Markatou, y C. Friedman. 2009. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: A feasibility study. *JAMIA*, 16(3):328–337.

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 91-98 recibido 29-03-2015 revisado 28-04-2015 aceptado 19-05-2015

Una aproximación a la recomendación de artículos científicos según su grado de especificidad*

An approach to the recommendation of scientific articles according to their degree of specificity

Antonio Hernández, David Tomás, Borja Navarro
Universidad de Alicante
Carretera San Vicente del Raspeig s/n - 03690 Alicante (Spain)
antoniojhb@gmail.com, {dtomas, borja}@dlsi.ua.es

Resumen: En este artículo se presenta un método para recomendar artículos científicos teniendo en cuenta su grado de generalidad o especificidad. Este enfoque se basa en la idea de que personas menos expertas en un tema preferirán leer artículos más generales para introducirse en el mismo, mientras que personas más expertas preferirán artículos más específicos. Frente a otras técnicas de recomendación que se centran en el análisis de perfiles de usuario, nuestra propuesta se basa puramente en el análisis del contenido. Presentamos dos aproximaciones para recomendar artículos basados en el modelado de tópicos (*Topic Modelling*). El primero de ellos se basa en la divergencia de tópicos que se dan en los documentos, mientras que el segundo se basa en la similitud que se dan entre estos tópicos. Con ambas medidas se consiguió determinar lo general o específico de un artículo para su recomendación, superando en ambos casos a un sistema de recuperación de información tradicional.
Palabras clave: recuperación de información, modelado de tópicos, sistemas de recomendación

Abstract: This article presents a method for recommending scientific articles taking into consideration their degree of generality or specificity. This approach is based on the idea that less expert people in a specific topic prefer to read more general articles to be introduced into it, while people with more expertise prefer to read more specific articles. Compared to other recommendation techniques that focus on the analysis of user profiles, our proposal is purely based on content analysis. We present two methods for recommending articles, based on Topic Modelling. The first one is based on the divergence of topics given in the documents, while the second uses the similarities that exist between these topics. By using the proposed methods it was possible to determine the degree of specificity of an article, and the results obtained with them overcame those produced by an information retrieval traditional system.
Keywords: information retrieval, topic modelling, recommender systems

1 Introducción

En los últimos años, se ha producido un aumento exponencial de la información digital que se genera y distribuye a través del World Wide Web, produciendo un incremento en la dificultad de encontrar información inmediatamente acorde a las necesidades de los usuarios. Ante la necesidad de escudriñar este maremágnum de información digitalizada, es que surgen los sistemas de recuperación de información (RI) (Baeza-Yates y Ribeiro-Neto, 1999).

Estos sistemas reciben una consulta por parte del usuario. Como resultado obtienen una lista de documentos relevantes y ordenados siguiendo criterios que intenta reflejar en qué medida se corresponden con dicha pe-

* Queremos agradecer a los revisores sus valiosas sugerencias y comentarios. Este trabajo ha sido parcialmente financiado por los siguientes proyectos: ATTOS (TIN2012-38536-C03-03), LEGOLANG-UAGE (TIN2012-31224), FIRST (FP7-287607), DIIM2.0 (PROMETEOII/2014/001) y por el Programa Nacional de Movilidad de Recursos Humanos del Plan Nacional de I+D+i (CAS12/00113).

ISSN 1135-5948

tición. Para ponderar esta relevancia se han propuesto numerosas medidas basadas en el contenido de los documentos, como TF-IDF (Salton y Buckley, 1988) y (Salton, 1991), los modelos de lenguaje (Ponte y Croft, 1998) y Okapi BM25 (Robertson et al., 1994).

Si bien estos sistemas han demostrado su efectividad para encontrar documentos relevantes que permitan satisfacer las necesidades de información expresadas por los usuarios mediante un conjunto de términos claves. Hay numerosas situaciones en las que la necesidad de información de un usuario no puede ser expresada ni satisfecha por estas vías. Un ejemplo de este tipo de situaciones se produce cuando un usuario pretende iniciarse en un nuevo campo del conocimiento o profundizar en un campo ya conocido.

Las medidas tradicionales de RI no resultan válidas por sí solas para determinar cómo de bueno es un documento a la hora de iniciarse en un tema. Hay aspectos como la necesidad de conocimientos previos, la diversidad de temas tratados en el documento, o la complejidad del discurso que no son tenidos en cuenta.

En este artículo presentamos una aproximación a la recomendación de artículos científicos según su grado de generalidad o especificidad. Para esta aproximación nos basamos en la noción que los usuarios más noveles en un tema preferirán leer artículos más generales para introducirse en el mismo, mientras que otros con un nivel más experto preferirán artículos más específicos.

Nuestras aproximaciones analizan el contenido de los artículos para determinar el grado de especificidad. Partiendo de la salida proporcionada por un sistema tradicional de RI, nuestra propuesta realiza una reordenación para ofrecer los artículos de más general a más específico. Utilizamos el modelo *Latent Dirichlet Allocation* (LDA) propuesto por Blei, YNg, y Jordan (2003) para el modelado de tópicos y planteamos dos aproximaciones. La primera se basa en la divergencia de tópicos entre artículos, mientras que la segunda se basa en la similitud de los mismos.

El resto del artículo está organizado de la siguiente manera: en la Sección 2 se aborda el estado de la cuestión de los sistemas de recomendación; la Sección 3 describe la aproximación propuesta; en la Sección 4 se describen los experimentos realizados y sus resultados; en la Sección 5 se detallan las conclusiones y las propuestas de trabajo futuro.

2 Estado de la cuestión

Como estado de la cuestión asumimos los estudios sobre la recomendación de artículos científicos que se están desarrollando. En este estudio nos centramos en las aproximaciones, técnicas de representación del contenido y el análisis de la especificidad o generalidad de los artículos.

En la recomendación de artículos científicos, McNee et al. (2002) después de analizar más de 170 algoritmos basados en Filtrado Colaborativo (CF) y Basados en el Contenido (CB), obtiene los mejores resultados con los de CF para determinar la importancia de un artículo con sus citas a partir de las citaciones web de los artículos de Research.net.

Ekstrand et al. (2010), siguiendo la idea anterior, aplica los algoritmos de CF y los híbridos para generar una lista automática de artículos, evaluando estos algoritmos sobre los artículos del *ACM Computing Surveys*, que dan una visión general sobre un tema, concluyendo también que los CF son mejores en este tipo de recomendación.

El perfil de usuario es una de las características que más se analizan para recomendar artículos. Se puede crear a través de la valoración que le proporciona el usuario a un documento, señalando su interés en el mismo (Giugni y León, 2011). Así como examinando los perfiles de usuarios existentes para detectar temas en común entre ellos (Wang y Blei, 2011).

Para determinar el carácter general o específico de los artículos, hasta donde nosotros sabemos, Candan et al. (2009) aplica la teoría de conjunto para detectar cuando los temas de los foros de discusión son nuevos, generales o específicos. Siguiendo está misma línea, Wang et al. (2010) aplica también la teoría de conjuntos a perfiles de tópicos. Estos perfiles son creados extrayendo las palabras claves del título y las 10 primeras oraciones de las noticias originales, incluyendo también los comentarios de los lectores.

Las aproximaciones descritas previamente determinan lo general o específico de un artículos de noticias de los foros de discusión. Estas aproximaciones únicamente aplican un proceso de extracción de tópicos basado en TF-IDF.

A diferencia de los trabajos anteriores. En nuestra aproximación obtenemos un listado

de artículos mediante el uso de un sistema tradicional de RI. Al contenido de los artículos se le aplica el modelo LDA para extraer los tópicos. Estos tópicos son analizados mediante su similitud y divergencia. Y finalmente obtenemos un listado reordenado de más general a más específico.

En la siguiente sección explicaremos en detalle nuestra aproximación.

3 Descripción de la aproximación

3.1 Modelado de tópicos

El modelado de tópicos tiene como fin encontrar a través de algoritmos estadísticos, los principales temas de colecciones de documentos.

Los recientes avances en este campo van más allá de analizar las colecciones masivas de documentos al análisis de colecciones de *streaming* de vídeos, de imágenes, así como encontrar patrones en datos genéticos, imágenes, redes sociales (Blei, Carin, y Dunson, 2010).

Existen diferentes tipos de modelos de tópicos, entre los que podemos encontrar, *Explicit Semantic Analysis, Latent Semantic Analysis, Latent Dirichlet Allocation, Hierarchical Dirichlet Process, Non-negative Matrix Factorization.* Como ya se mencionó en la introducción nosotros usamos el *Latent Dirichlet Allocation* (LDA).

LDA es un modelo probabilístico generativo de tópicos. Los documentos de un corpus se representan como una combinación aleatoria sobre los tópicos latentes, donde cada tópico es caracterizado por una distribución de probabilidades sobre un vocabulario fijo de palabras. Para cada documento del corpus se generan palabras aleatoriamente siguiendo las siguientes etapas:

1. Seleccionar aleatoriamente una distribución de tópicos.
2. Para cada palabra del documento:
   - Seleccionar un tópico aleatoriamente sobre los tópicos generados en el paso 1.
   - Seleccionar una palabra aleatoriamente sobre su correspondiente vocabulario.

Al final de proceso obtendremos la probabilidad de pertenencia de cada palabra de cada documento a cada tópico (Blei, YNg, y Jordan, 2003) y (Blei, Carin, y Dunson, 2010).

3.2 Descripción

Vamos a utilizar dos métodos para analizar el contenido de los artículos, uno basado en la divergencia de los tópicos de un artículo y otro basado en la similitud de tópicos entre artículos.

Para analizar la divergencia de los tópicos calculamos la desviación estándar entre ellos, siguiendo la intuición de que aquellos artículos cuya desviación sea menor tendrán una distribución de tópicos más equilibrado y por lo tanto tendrán un carácter más general, ya que no tendrán tópicos que destaquen marcadamente sobre el resto. Por otra parte, aquellos cuya desviación típica sea más elevada será porque existen tópicos más representativos que otros en el artículo, y que por lo tanto se trata de un artículo más específico.

En el caso de la similitud de tópicos entre artículos, emplearemos la similitud del coseno siguiendo la idea de que los artículos que presentan una similitud promedio más alta con otros artículos serán más generales, ya que tienen más tópicos comunes con otros artículos.

Los que tienen una similitud baja es porque tienen tópicos en común con sólo unos pocos y seguramente serán artículos más específicos.

Para determinar cuándo un artículo científico es general o específico. Seguimos los siguientes pasos que se detallan a continuación:

- Consulta: es el tema que se desea buscar en el corpus de artículos científicos.
- Indexación y recuperación de los artículos: usamos Lucene [1] que es un motor de búsqueda de alto rendimiento escrito en Java y de código abierto. Entre sus características permite la indexación incremental de documentos, páginas web y contenidos de bases de datos, entre otras (McCandless, Hatcher, y Gospodnetic, 2010).
  Los artículos recuperados se denotan por $P = (a_1, a_2, a_3, a_4, \ldots, a_n)$ y están ordenados de mayor a menor relevancia. Esta relevancia viene dada por los parámetros por defecto con que se configura Lucene.

[1] `http://lucene.apache.org/core/`

Por lo tanto, este sistema de RI nos sirve como un primer filtro para detectar dentro de un corpus los documentos más relevantes al área de estudio.

- Análisis del contenido de los artículos: para ello aplicamos la herramienta Mallet (McCallum, 2002). Esta herramienta está basada en el algoritmo LDA para el modelado de tópicos, permitiendo detectar y extraer tópicos del corpus de artículos. Los tópicos se almacenan en diferentes formatos y varios ficheros como resultado final.
  Nuestro análisis se centra en el fichero de pesos de los tópicos. Este fichero está conformado por una matriz donde los artículos están en la filas y los pesos de los tópicos de cada artículo en las columnas. En otras palabras, cada artículo estaría representado por los pesos de sus tópicos.
- Resultados: el listado de artículos recomendados se muestran según su generalidad (de más general a específico). También se puede configurar para que se muestre de lo más específico a general, es decir según su especificidad.

3.3 Aproximación #1. Divergencia

En esta aproximación analizamos como difieren entre sí los tópicos en un mismo artículo. Este análisis lo hacemos calculando la desviación estándar mediante la fórmula 1:

$$\sigma(P) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{a})^2} \qquad (1)$$

Como resultado obtenemos un valor ponderado para cada artículo entre 0 y 1. Estos valores se interpretan de la siguiente manera:

- $\sigma(P) \approx 1$, se trata de un artículo específico ya que la importancia de los distintos tópicos en el documento varía mucho de unos a otros.
- $\sigma(P) \approx 0$, se trata de uno general ya que todos los tópicos tienen aproximadamente la misma relevancia en el artículo.

3.4 Aproximación #2. Similitud

En esta segunda aproximación, analizamos la similitud entre artículos. La similitud se calcula determinado el ángulo del coseno que forman entre sí los vectores de tópicos de los artículos a través de la fórmula 2:

$$Sim(a_i, a_{i+1}) = \frac{\sum_k^n a_{ik} \cdot a_{i+1k}}{\sqrt{\sum_k^n a_{ik}^2} \cdot \sqrt{\sum_k^n a_{i+1}^2}} \qquad (2)$$

Como resultado obtenemos un valor ponderado para cada artículo entre 0 y 1. Estos valores se interpretan de la siguiente manera:

- $\sigma(P) \approx 1$, se trata de un artículo general ya que los tópicos del artículo aparecen en muchos otros artículos.
- $\sigma(P) \approx 0$, se trata de uno específico ya que todos los tópicos del artículo aparecen en pocos.

En la siguiente sección se explican mediante experimentos los resultados de ambas aproximaciones.

4 Experimentos y resultados

4.1 Selección del corpus de evaluación

Para la comprobación de nuestra aproximación a través de los experimentos se decidió usar el repositorio de artículos científicos de la ACL Anthology [2]. Ofrece de manera abierta más de 20.000 artículos científicos pertenecientes a los diferentes congresos organizados por la *Association for Computational Linguistics* (ACL) durante los últimos 40 años.

A través del proyecto ACL Anthology Network (Radev et al., 2013), se ofrece una versión en texto plano de todos los artículos de la ACL Anthology desde sus inicios hasta 2012. Este proyecto ofrece además información adicional como son la redes de citas del artículo y la de citas recibidas por el autor, etc.

4.2 Creación del corpus de referencia (*gold standard*)

Para poder evaluar el nivel de generalidad o especificidad de un artículo hemos creado un corpus de referencia (*gold standard*). Este corpus fue creado a partir de los 100 primeros artículos devueltos por Lucene para la consulta *question answering*. Estos 100 artículos fueron anotados manualmente por un experto en el tema. La anotación se realizó en una escala de 0 a 5 para determinar su grado de generalidad partiendo del siguiente criterio:

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

- 5: El artículo aporta una visión general sobre el área (survey), un artículo que se centra en mostrar el estado de la cuestión para *question answering*.
- 4: El artículo es una visión general sobre un subtema dentro del tema principal. Por ejemplo, un artículo que muestra todo el estado de la cuestión sobre la clasificación de preguntas aplicado a sistemas de *question answering*.
- 3: El artículo presenta una aproximación general al área. Por ejemplo, un artículo que describe una aproximación general a resolver el problema de *question answering*.
- 2: El artículo presenta una aproximación en un dominio o área concreta. Por ejemplo, un artículo que describe una aproximación al problema de la clasificación de preguntas en sistemas de *question answering*, o una aproximación general a *question answering* en un idioma concreto.
- 1: El artículo presenta un proyecto o recurso concreto. Por ejemplo, un artículo que presenta un corpus de preguntas para entrenar sistemas de *question answering*.
- 0: El artículo no tiene que ver con el área, es decir, Lucene nos ha devuelto algo que no tiene que ver con *question answering*.

Con esto se generó el orden de lectura de los artículos devueltos por Lucene según un experto.

4.3 Experimentos

Se realizaron dos experimentos con el fin de comprobar las ideas expuestas. Estos experimentos tuvieron como objetivos determinar la cantidad óptima de tópicos y evaluar las aproximaciones descritas anteriormente.

Para la evaluación usamos *Normalized Discounted Cumulative Gain* (nDCG) (ver fórmula 3). Esta medida combina la puntuación del documento (entre 0 y 5) con la posición en la que ha sido devuelto dentro de la lista de artículos recomendados, dando como única medida la ganancia acumulada sin importar el tamaño de la lista de documentos recuperados (Järvelin y Kekäläinen, 2002).

Como *baseline* para comparar el rendimiento de nuestra aproximación utilizamos la salida original proporcionada por el motor de recuperación de información Apache Lucene (McCandless, Hatcher, y Gospodnetic, 2010).

En los experimentos propuestos se ha optado por mostrar en primer lugar aquellos de carácter más general.

A continuación se describen en detalles estos experimentos y sus resultados.

4.3.1 Experimento 1

El objetivo de este experimento fue determinar la cantidad óptima k de tópicos que se deben extraer del corpus de artículos para realizar su respectivo análisis de divergencia (σ) y similitud (Sim).

Para cumplir con este objetivo la secuencia de pasos que seguimos fue:

- Consulta, se realizó la consulta *question answering* sobre Lucene en el corpus de la ACL y se recuperaron 100 artículos, denotado por $P = (a_1, a_2, a_3, \ldots, a_{100})$.
- Ordenación del corpus de referencia (CR), se ordenaron los artículos de mayor a menor puntuación, es decir, 5, 5, 5, 4, 4, 4,..., 0, 0, 0 con el fin de obtener el mejor valor posible del sistema.
- Creación del corpus, se creó un corpus a partir de P (C_P).
- Extracción de tópicos, sobre el corpus C_P se experimentó con la obtención de distinto número de tópicos (5, 10, 15, 25, 50, 100, 150, 200, 250, y 300) para determinar la configuración óptima para nuestro sistema.
- Cálculo de medidas, para cada grupo de artículos creados por sus tópicos se aplicaron las dos aproximaciones definidas (desviación estándar de los tópicos de un documento y la similitud del coseno entre los tópicos de distintos documentos).
- Análisis de tópicos: se calculó $nDCG$ para cada uno de los conjuntos de tópicos propuestos (entre 5 y 300), a través de la fórmula 3:

$$nDCG_P = CR_{P_{[1]}} + \sum_{i=2}^{100} \frac{CR_{P_{[i]}}}{\log_2(i+1)} \quad (3)$$

  Los valores de $CR_{P_{[1]}}$ se obtienen viendo la posición que tiene a_i en el corpus de referencia.

La tabla 1 muestra los resultados de $nDCG$ obtenidos para 5, 10, 15, 25, 50, 100,

nDCG	Número de tópicos									
	5	10	15	25	50	100	150	200	250	300
Sim	0,8328	**0,8399**	0,7995	0,7956	0,7970	0,8000	0,8056	0,7754	0,8343	0,8382
σ	0,7537	0,7654	0,8042	0,8384	0,8320	0,8207	0,8274	0,8333	0,8452	**0,8476**
M	0,7933	0,8027	0,8019	0,8170	0,8145	0,8104	0,8165	0,8044	0,8398	**0,8424**

Tabla 1: Resultados de $nDCG$ para cada grupo de tópicos

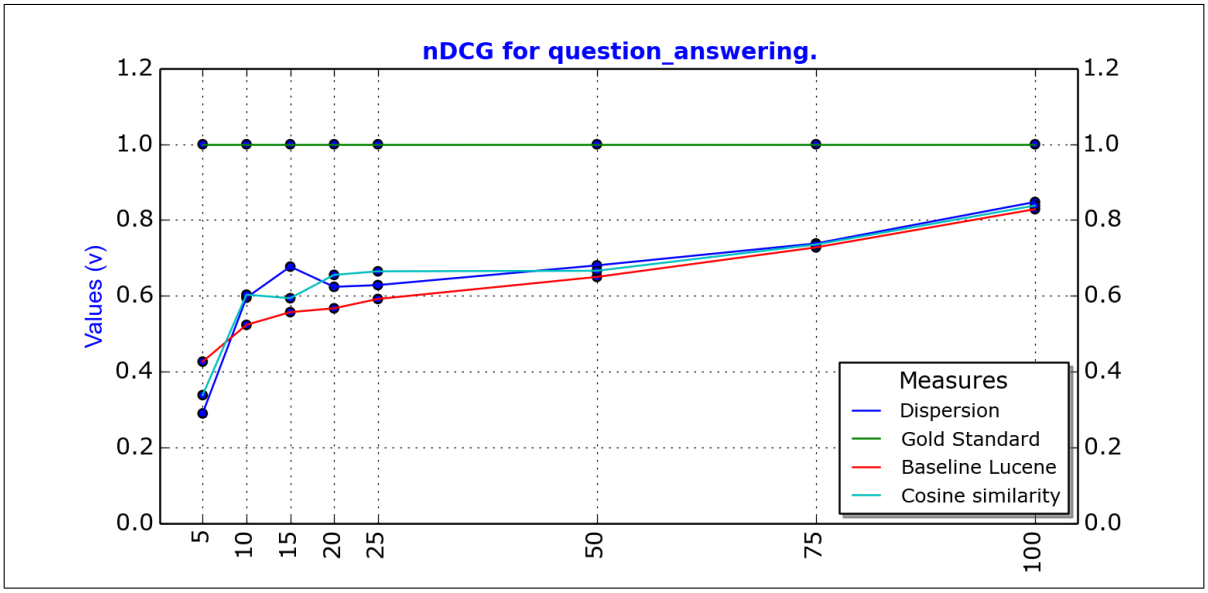

Figura 1: Gráfico de $nDCG$ por cantidad de artículos.

150, 200, 250 y 300 tópicos estos valores fueron normalizados al dividir el resultado de DCG por el valor máximo posible obtenido al ordenar CR. Podemos observar que la mejor similitud entre artículos se logra con 10 tópicos y la divergencia entre tópicos en un artículo con 300.

En el caso de la similitud existe una diferencia de 0,0017 entre 10 y 300 tópicos. Al ser esta diferencia muy pequeña y el mayor valor de la media (M=0,8429) en los 300 tópicos. Se decidió fijar a 300 tópicos el valor óptimo para el segundo experimento.

Con este experimento concluimos qeu con el análisis de 300 tópicos por artículos tenemos en las primeras posiciones los artículos mejores anotados con respecto a CR_P.

4.3.2 Experimento 2

Este experimento tuvo como objetivo lograr una lista de artículos donde los primeros que se recuperen sean los más generales que aborden sobre el tema realizado en la consulta.

Partimos realizando una consulta con Lucene sobre *question answering*, recuperando los 100 artículos más relevantes y aplicando Mallet para extraer 300 tópicos para cada artículo (la mejor configuración según el experimento realizado en el apartado anterior).

Con estos 300 tópicos procedimos a realizar dos aproximaciones: la primera basada en el cálculo de la desviación estándar para analizar la divergencia de los tópicos en cada artículo, y la segunda en la similitud del coseno para ver la co-ocurrencia de tópicos entre artículos.

En este caso utilizamos la medida $nDCG@k$, que es la versión *cut-off* de $nDCG$ que presta más atención a los primeros resultados de la lista de artículos recuperados (Wang et al., 2013). El propósito de usar esta medida es valorar de manera más positiva aquellos sistemas que devuelven mejores artículos en las primeras posiciones de la lista. Pensando en la usabilidad de este tipo de sistemas, la satisfacción del usuario pasa por devolver los artículos que le resulten relevantes en las primeras posiciones.

Como se puede observar en la figura 1 a partir de 25 artículos se produce un aumento del $nDCG$, pero este aumento viene dado por

nDGC@k	**Lucene**	*Sim*	σ
5	0,4263	0,3380 (-20,7 %)	0,2896 (32,1 %)
10	0,5233	0,6030 (15,2 %)	0,5950 (13,7 %)
15	0,5569	0,5937 (6,6 %)	**0,6771 (21,6 %)**
20	0,5670	**0,6552 (15,6 %)**	0,6237 (10,0 %)
25	0,5917	0,6646 (12,3 %)	0,6283 (6,1 %)

Tabla 2: Resultados de nDGC@k

el propio funcionamiento de la medida: el corpus de referencia tiene valores más bajos en las últimas posiciones, incluso ceros, mientras que nuestra aproximación y el experimento de referencia *baseline* contienen artículos con una valoración más alta en esas posiciones.

En cualquier caso, como se comentó más arriba, lo que nos interesa es mirar los valores que se obtienen en las primeras posiciones, ya que a la hora de recomendar artículos a un usuario no es conveniente proporcionar un número elevado de ellos, ya que el usuario lo que espera son resultados válidos en las primeras posiciones.

Por esta razón, vamos a centrar el estudio en los 25 primeros artículos, aplicando $nDCG@k$, donde $k = \{5, 10, 15, 20 \text{ y } 25\}$.

En la tabla 2 se reflejan los resultados de aplicar $nDCG@k$ al *baseline* (Lucene), notando que nuestras aproximaciones en 10, 15, 20 y 25 artículos respectivamente supera siempre al *baseline*.

5 Conclusiones y trabajo futuro

En este artículo se han presentado dos aproximaciones basadas en modelado de tópicos para recomendar artículos científicos según su grado de especificidad. Primero analizamos la divergencia de los temas que se abordan en cada artículo, y seguidamente la similitud de los temas entre artículos.

Para poder evaluar nuestras aproximaciones tuvimos que crear nuestro propio corpus (100 artículos) partiendo de los más de 20.000 artículos de investigación que contiene la ACL. Nuestro corpus consistió en recuperar los primeros 100 artículos en el área de la "búsqueda de respuestas" usando el motor de recuperación de información Apache Lucene.

Este listado de artículos en el orden que los recuperó Lucene fue nuestro *baseline* y para evaluar el nivel de generalidad o especificidad se etiquetaron de más general (5) a más específico (1), siendo nuestro corpus de referencia (gold standard).

Aplicando el modelado de tópicos hemos sido capaces de reordenar los documentos devueltos por un sistema de RI para que se muestren de manera más general a más específica.

El primer experimento consistió en determinar la mejor configuración posible en cuanto al número de tópicos a usar en el algoritmo LDA, obteniendo con 300 tópicos el resultado más óptimo.

El segundo experimento consistió en emplear nuestras dos aproximaciones (basada en la divergencia de tópicos en un documento y basada en la similitud del coseno entre tópicos de distintos documentos) para determinar la mejor reordenación posible de artículos, de más general a más específico.

Nuestras aproximaciones superaron al *baseline* de manera clara utilizando la medida $nDGC$, obteniendo una mejora del 21,6 % para $nDGC@15$ utilizando la medida de la divergencia de tópicos en un documento, y de un 15,6 % para $nDGC@20$ usando la similitud del coseno entre tópicos de distintos documentos.

Como trabajo futuro se plantea la forma de combinar la divergencia y similitud de los tópicos de los artículos para obtener mejores resultados a la hora de detectar la generalidad o especificidad de los artículos.

Bibliografía

Baeza-Yates, R. y B. Ribeiro-Neto. 1999. *Modern information retrieval*, volumen 463. ACM press New York.

Blei, D., L. Carin, y D. Dunson. 2010. Probabilistic topic models. *IEEE Signal Processing Magazine*, 27(6):55–65.

Blei, D., A. YNg, y M. Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Candan, K., E. Mehmet, T. Hedgpeth, J. Wook, Q. Li, y M. Sapino. 2009. Sea: Segment-enrich-annotate paradigm for adapting dialog-based content for improved accessibility. *ACM Transactions on Information Systems*, 27(3):1–45.

Ekstrand, M., P. Kannan, J. Stemper, J. Butler, J. Konstan, y J. Riedl. 2010. Automatically building research reading lists. En *Proceedings of the fourth ACM conference on Recommender systems*, páginas 159–166. ACM.

Giugni, M. y L. León. 2011. Clusterdoc un sistema de recuperación y recomendación de documentos basado en algoritmos de agrupamiento. *Telematique*, 9(2):13–28.

Järvelin, K. y J. Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

McCallum, A. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu (Consultado: 02 12 2014).

McCandless, M., E. Hatcher, y O. Gospodnetic. 2010. *Lucene in Action: Covers Apache Lucene 3.0*. Manning Publications Co.

McNee, S., I. Albert, D. Cosley, P. Gopalkrishnan, S. Lam, A. Rashid, J. Konstan, y J. Riedl. 2002. On the recommending of citations for research papers. En *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, páginas 116–125. ACM.

Ponte, J. y B. Croft. 1998. A language modeling approach to information retrieval. En *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, páginas 275–281. ACM.

Radev, D., P. Muthukrishnan, V. Qazvinian, y A. Abu-Jbara. 2013. The acl anthology network corpus. *Language Resources and Evaluation*, 47(4):919–944.

Robertson, S., S. Walker, S. Jones, M. Hancock-Beaulieu, y M. Gatford. 1994. Okapi at trec-3. En *Proceedings of TREC*, volumen 3, páginas 109–126.

Salton, G. 1991. Developments in automatic text retrieval. *Science*, 253(5023):974–980.

Salton, G. y C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Wang, C. y D. Blei. 2011. Collaborative topic modeling for recommending scientific articles. En *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, páginas 448–456. ACM.

Wang, J., Q. Li, Y. Chen, J. Liu, C. Zhang, y Z. Lin. 2010. News recommendation in forum-based social media. *The Journal of Information Science*, 180(24):4929–4939.

Wang, Y., L. Wang, Y. Li, D. He, T. Liu, y W. Chen. 2013. A theoretical analysis of ndcg type ranking measures. *CoRR*, abs/1304.6480. http://dblp.uni-trier.de/db/journals/corr/corr1304.html (Consultado: 15 02 2015).

Traducción automática

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 101-108 recibido 26-03-2015 revisado 28-04-2015 aceptado 15-05-2015

An Empirical Analysis of Data Selection Techniques in Statistical Machine Translation *

Análisis empírico de técnicas de selección de datos en traducción automática estadística

Mara Chinea-Rios
Universitat Politècnica de València
Camino de Vera s/n, Valencia, Spain
machirio@prhlt.upv.es

Germán Sanchis-Triches
Universitat Politècnica de València
Camino de Vera s/n, Valencia, Spain
gsanchis@dsic.upv.es

Francisco Casacuberta
Universitat Politècnica de València
Camino de Vera s/n, Valencia, Spain
fcn@prhlt.upv.es

Resumen: La adaptación de dominios genera mucho interés dentro de la traducción automática estadística. Una de las técnicas de adaptación esta basada en la selección de datos que tiene como objetivo seleccionar el mejor subconjunto de oraciones bilingües de un gran conjunto de oraciones. En este artículo estudiamos como afectan los corpus bilingües empleados por los métodos de selección de frases en la calidad de las traducciones.
Palabras clave: traducción automática estadistica, adaptación dominios, selección de frases bilingües, n-gramas infrecuentes, entropía cruzada

Abstract: Domain adaptation has recently gained interest in statistical machine translation. One of the adaptation techniques is based in the selection data. Data selection aims to select the best subset of the bilingual sentences from an available pool of sentences, with which to train a SMT system. In this paper, we study how affect the bilingual corpora used for the data selection methods in the translation quality.
Keywords: statistical machine translation, domain adaptation, bilingual sentence selection, infrequent n-gram, cross-entropy

1 Introduction

Statistical machine translation (SMT) system quality depends on the available parallel training data. Two factors are important: the size of the parallel training data and the domain. A small set of training data leads to poorly estimated translation models and consequently poor translation quality. Unfortunately, we do not have parallel data in all domains. For this reason, the translation quality gets worse when we do not have enough training data for the specific domain we need to tackle in our test set. Intuitively, domain adaptation methods try to make a better use of the part of the training data that is more similar, and therefore more relevant, to the text that is being translated (Sennrich, 2013).

There are many domain adaptation methods that can be split into two broad categories. Domain adaptation can be done at the corpus level, for example, by weighting, selecting or joining the training corpora. In contrast, domain adaptation can also be done at the model level by adapting directly the translation or language models.

Bilingual sentence selection (BSS) aims to select the best subset of bilingual sentences from an available pool of sentence pairs. Here, we focus on studying the performance of two different BSS strategies. We will refer to the pool of sentences as *out-of-domain* (OoD) corpus because we assume that it belongs to a different domain than the one to be translated. We will refer to the corpus of the domain to be translated as *in-domain* (ID) corpus.

Since we will be analysing the BSS techniques as applied to the specific case of SMT, we review briefly the SMT (Papineni, Roukos, and Ward, 1998; Och and Ney, 2002) frame

* The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 287576 (CasMaCat). Also funded by the Generalitat Valenciana under grant Prometeo/2009/014.

ISSN 1135-5948

work: given an input sentence $\mathbf{x}$ from a certain source language, the purpose is to find an output sentence $\mathbf{y}$ in a certain target language such that:

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} \sum_{m=1}^{M} \lambda_m h_m(\mathbf{x}, \mathbf{y}) \qquad (1)$$

where λ_m is the weight assigned to $h_m(\mathbf{x}, \mathbf{y})$ and $h_m(\mathbf{x}, \mathbf{y})$ is a score function representing an important feature for the translation of $\mathbf{x}$ into $\mathbf{y}$, as for example the language model of the target language, a reordering model, or several translation models. The weights λ_m are normally optimised with the use of a development set. The most popular approach for adjusting λ_m is the one proposed in (Och, 2003), commonly referred to as *minimum error rate training* (MERT).

The main contribution of this paper is:

- An empirical analysis of two BSS techniques with different corpora.

This paper is structured as follows. Section 2 summarises the related work. Section 3 presents the two BSS techniques selected, namely, infrequent n-grams recovery and cross entropy selection. In Section 4, experimental results are reported. Conclusions and future work are presented in Section 5.

2 Related work

State-of-the-art BSS approaches rely on the idea of choosing those sentence pairs in the OoD training corpus that are in some way similar to an ID training corpus in terms of some different metrics.

The simplest instance of this problem can be found in language modelling, where perplexity-based selection methods have been used (Gao et al., 2002). Here, OoD sentences are ranked by their perplex-ity score. Another perplexity-based approach is presented in (Moore and Lewis, 2010), where cross-entropy difference is used as a ranking function rather than just perplexity, in order to account for normalization. We apply this criterion for the task of selecting training data for SMT systems

Different works use perplexity-related BSS strategies (Axelrod, He, and Gao, 2011; Rousseau, 2013). In Axelrod, He and Gao (2011), the authors used three methods based in cross-entropy for extracting a pseudo ID corpus. This pseudo ID corpus is used to train small domain-adapted SMT systems. In (Rousseau, 2013) the authors describe the *XenC* open source toolkit for data selection. *XenC* uses the two strategies described in (Gao et al., 2002) and (Moore and Lewis, 2010).

Two different approaches are presented in (Gascó et al., 2012): one based on approximating the probability of an ID corpus and another one based on infrequent n-gram occurrence. The technique approximating the probability relies on preserving the probability distribution of the task domain by wisely selecting the bilingual pairs to be used. Hence, it is mandatory to exclude sentences from the pool that distort the actual probability. The technique based in infrequent n-gram occurrence will be explained in detail in the next section.

Other works have applied information retrieval methods for BSS (Lü, Huang, and Liu, 2007), in order to produce different submodels which are then weighted. In that work, authors define the baseline as the result obtained by training only with the corpus that shares the same domain with the test. Afterwards, they claim that they are able to improve the baseline translation quality by adding new sentences retrieved with their method. However, they do not compare their technique with a model trained with all the corpora available.

3 Data selection methods

In this section we present the two techniques that we have selected for our work. The first strategy we used in this work is infrequent n-grams recovery. This strategy was presented in Gascó et al., (2012).

The second strategy, proposed in Moore and Lewis, (2010), is based in cross-entropy. This strategy is used in many different works (Axelrod, He, and Gao, 2011; Rousseau, 201); Schwenk, Rousseau, and Attik, 2012; Sennrich, 2012). In these works, the authors report good results when using this strategy, and has become a defacto standard in the SMT research community.

3.1 Infrequent n-grams recovery

The main idea underlying the infrequent n-grams recovery strategy consists in increasing the information of the ID corpus by adding evidence for those n-grams that have been

seldom observed in the ID corpus. The n-grams that have never been seen or have been seen just a few times are called *infrequent n-grams*. An n-gram is considered infrequent when it appears less times than a given infrequency threshold t. Therefore, the strategy consists on selecting from the OoD corpus the sentences which contain the most infrequent n-grams in the source sentences to be translated.

Let X be the set of n-grams that appears in the sentences to be translated and $\mathbf{w}$ one of them; let $N_{\mathbf{x}}(\mathbf{w})$ be the counts of $\mathbf{w}$ in a given source sentence $\mathbf{x}$ of the OoD corpus, and $C(\mathbf{w})$ the counts of $\mathbf{w}$ in the source language ID corpus. Then, the infrequency score $i(\mathbf{x})$ is defined as:

$$i(\mathbf{x}) = \sum_{\mathbf{w} \in X} \min(1, N_{\mathbf{x}}(\mathbf{w})) \max(0, t - C(\mathbf{w})) \quad (2)$$

Then, the sentences in the OoD corpus are scored using Equation 2. The sentence $\mathbf{x}^*$ with the highest score $i(\mathbf{x}^*)$ is added to the ID corpus and removed from the OoD sentences. The counts of the n-grams $C(\mathbf{w})$ are updated with the counts $N_{\mathbf{x}^*}(\mathbf{w})$ within $\mathbf{x}^*$ and therefore the scores of the OoD corpus are updated. Note that t will determine the maximum amount of sentences that can be selected, since when all the n-grams within X reach the t frequency no more sentences will be extracted from the OoD corpus.

3.2 Cross-entropy selection

As mentioned in Section 2, one established method consists in scoring the sentences in the OoD corpus by their perplexity score. We follow the procedure described in Moore and Lewis (2010), which uses the cross-entropy rather than perplexity. Perplexity and cross-entropy are monotonically related. The perplexity of a given sentence $\mathbf{x}$ with empirical n-gram distribution p given a language model q is:

$$2^{-\sum_x p(x) \log q(x)} = 2^{H(p,q)} \quad (3)$$

where $H(p, q)$ is the cross-entropy between p and q. The formulation proposed by Moore and Lewis (2010) is: Let I be an ID corpus and G be an OoD corpus. Let $H_I(\mathbf{x})$ be the cross-entropy, according to a language model trained on I, of a sentence $\mathbf{x}$ drawn from G. Let $H_G(\mathbf{x})$ be the cross-entropy of $\mathbf{x}$ according to a language model trained on G. The cross-entropy score of $\mathbf{x}$ is then defined as

$$c(\mathbf{x}) = H_I(\mathbf{x}) - H_G(\mathbf{x}) \quad (4)$$

In this work we will also analyse the effect of varying the order of the n-grams considered, since this will also imply that the final sentence selection will be different. Specifically, we will consider $n = \{2, 3, 4, 5\}$ grams.

4 Experiments

4.1 Experimental set-up

We evaluated empirically the methods described in the previous section. As ID data, we used two different corpora. The EMEA[1] corpus (Tiedemann, 2009) is available in 22 languages and contains documents from the European Medicines Agency. The other ID corpus is the News Commentary[2] (NC) corpus. The NC corpus is composed of translations of news articles. The main statistics of the ID corpora used are shown in Table 1. For the OoD corpora, we used two corpora belonging to different domains readily available in the literature. Table 2 shows the main features of the two OoD corpora. The Europarl[3] corpus is composed of translations of the proceedings of the European parliament (Koehn, 2005). The PatTR corpus[4] (Wäschle and Riezler, 2012) is a parallel corpus extracted from the MAREC patent collection.

All experiments were carried out using the open-source SMT toolkit Moses version phrase-based (Koehn et al., 2007). The language model used was a 5-gram, standard in SMT research, with modified Kneser-Ney smoothing (Kneser and Ney, 1995), built with the SRILM toolkit (Stolcke, 2002). The phrase table was generated by means of symmetrised word alignments obtained with GIZA++ (Och and Ney, 2003). The de-coder features a statistical log-linear model including a phrase-based translation model, a language model, a distortion model and word and phrase penalties. The log-lineal combi-nation weights in Equation 1 were optimized using MERT (minimum error rate training)(Och, 2003).

For each domain, we trained two differ-ent baselines with which to compare the systems obtained by data selection. The

[1] www.statmt.org/wmt14/medical-task/
[2] www.statmt.org/wmt13
[3] www.statmt.org/europarl/
[4] www.cl.uni-heidelberg.de/statnlpgroup/pattr/

Corpus		$\|S\|$	$\|W\|$	$\|V\|$
EMEA-Domain	EN	1.0M	12.1M	98.1k
	FR		14.1M	112k
Medical-Test	EN	1000	21.4k	1.8k
	FR		26.9k	1.9k
Medical-Mert	EN	501	9850	979
	FR		11.6k	1.0k
NC-Domain	EN	157k	3.5M	65k
	FR		4.0M	76k
NC-Test	EN	3000	56.0k	4.8k
	FR		61.5k	5.0k
NC-Mert	EN	2050	43.4k	3.9k
	FR		47.1k	4.1k

Table 1: ID corpora main figures. (EMEA-Domain and NC-Domain) are the ID corpora, (Medical-Test and NC-Test) are the evaluation data and (Medical-Mert and NC-Mert) are development set. M denotes millions of elements and k thousands of elements, $|S|$ stands for number of sentences, $|W|$ for number of words (tokens) and $|V|$ for vocabulary size (types).

Corpus		$\|S\|$	$\|W\|$	$\|V\|$
Europarl	EN	2.0M	50.2M	157k
	FR		52.5M	215k
PatTR	EN	3.4M	78.6M	190k
	FR		81.8M	212k

Table 2: OoD corpora main figures (See Table 1 for an explanation of the abbreviations).

first baseline was obtained by training the SMT system only with ID training data: EMEA-Domain and NC-Domain, obtaining the `baseline-emea` and `baseline-nc` baselines, respectively. The second baseline was obtained by training the SMT system with a concatenation of either of the OoD corpora (Europarl or PatTR) and the ID training data (EMEA-Domain or NC-Domain):

- `bsln-emea-euro`: EMEA ∪ Europarl
- `bsln-nc-euro`: NC ∪ Europarl
- `bsln-emea-pattr`: EMEA ∪ PatTR
- `bsln-nc-pattr`: NC ∪ PatTR.

Results are shown in terms of BLEU (Papineni et al., 2002), measures the precision of uni-grams, bigrams, trigrams, and four-grams with respect to a set of reference translations, with a penalty for too short sentences.

4.2 Results for the infrequent n-grams technique

In this section, we present the experimental results obtained by infrequent n-grams recovery for each set-up presented in Section 4.1.

Figures 1 and 2 show the effect of adding sentences using infrequent n-grams selection, up to the point where the specific value of t does not allow to select further sentences. In addition, the result obtained with the two baseline systems is also displayed. We only show results for threshold values $t = \{10, 20\}$ for clarity, although experiments were also carried out for $t = \{10, 15, 20, 25, 30\}$ and such results presented similar curves.

Figure 1 shows the principal result obtained using the Europarl OoD corpus. Several conclusion can be drawn:

- The translation quality provided by the infrequent n-grams technique is large better in term of BLEU than the results achieved with the system `baseline-nc` and `baseline-emea`.
- Selecting sentences with the infrequent n-grams technique provides better results than including the OoD corpus in the SMT system with Medical domain (`bsln-emea-euro`). Specifically, the improvements obtained are in the range 0.70 BLEU points using less than 4% of the Europarl OoD corpus. Different result are obtained with News domain. In this scenario, the infrequent n-grams technique does not provide significantly better results than including the OoD corpus in the SMT system with News domain (`bsln-nc-euro`). But the results are very similar using less than 8% of the OoD corpus.
- As expected, $t = 20$ allows to select more sentences than $t = 10$, which also leads to higher BLEU scores. The results with $t = 10$ are slightly worse than with $t = 20$, for the same amount of sentences. We understand that this is because $t = 20$ entails a better estimation of the n-grams considered infrequent.

Figure 2 shows the principal results obtained using PatTR OoD corpus.

- In this scenario, the results achieved by the baseline systems do not show a significant difference when including the

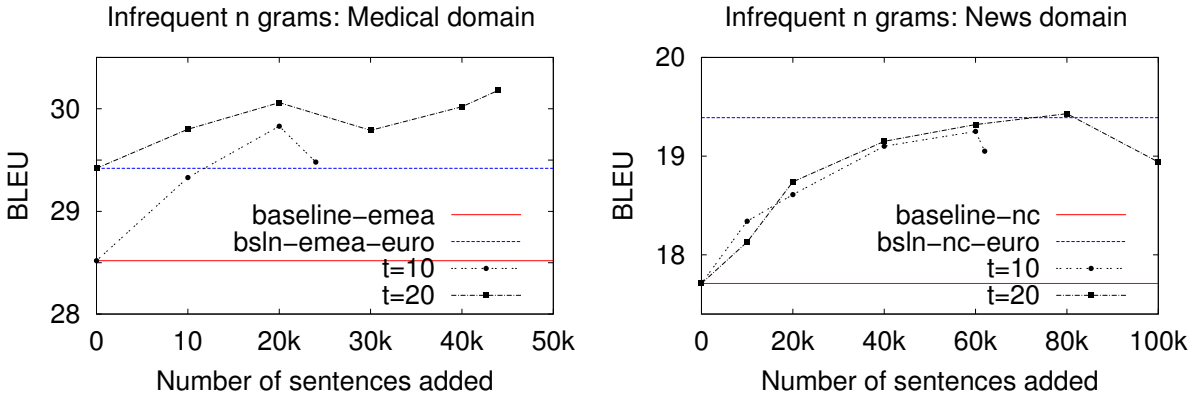

Figure 1: Effect over the BLEU score using infrequent n-grams recovery for two ID corpora EMEA and News and the Europarl OoD corpus. Horizontal lines represent the score when using the ID corpus and all the data available.

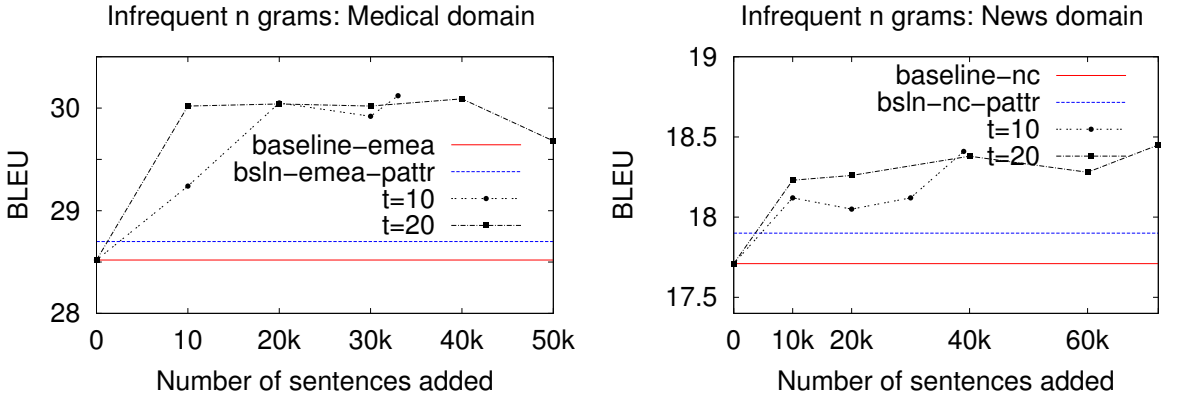

Figure 2: Effect over the BLEU score using infrequent n-grams recovery for two ID corpora EMEA and News and the PatTR OoD corpus. Horizontal lines represent the score when using the ID corpus and all the data available.

OoD PatTR data. We conclude that this corpus does not provide relevant information for the SMT system.

- The translation quality provided by the infrequent n-grams technique is large better in term of BLEU than the results achieved with all baseline systems, which evidence that the selection strategy is able to make a good use of the OoD data, even if such data as a whole does not seem to be useful. We understand that this is important, since it proves the utility of the BSS strategy.

4.3 Results for cross-entropy strategy

In this section, we present the experimental results obtained by the cross-entropy strategy for each set-up presented in Section 4.1.

Figures 3 and 4 show the effect of adding sentences by means of the cross-entropy strategy. We only show results using both 2-grams and 5-grams for clarity, although experiments were also carried out for $n = \{2, 3, 4, 5\}$.

- Adding sentences selected by means of cross-entropy improves over `baseline-emea` and `baseline-nc` from the very beginning, except the results obtained when testing in the medical domain and training with the PatTR OoD corpus.
- Cross-entropy data selection is not able to achieve improvements over training with all the data available when the Europarl corpus is considered as OoD corpus. When considering the PatTR, slight improvements are achieved, although such improvements are not very significant.
- In most cases, the order of the n-grams

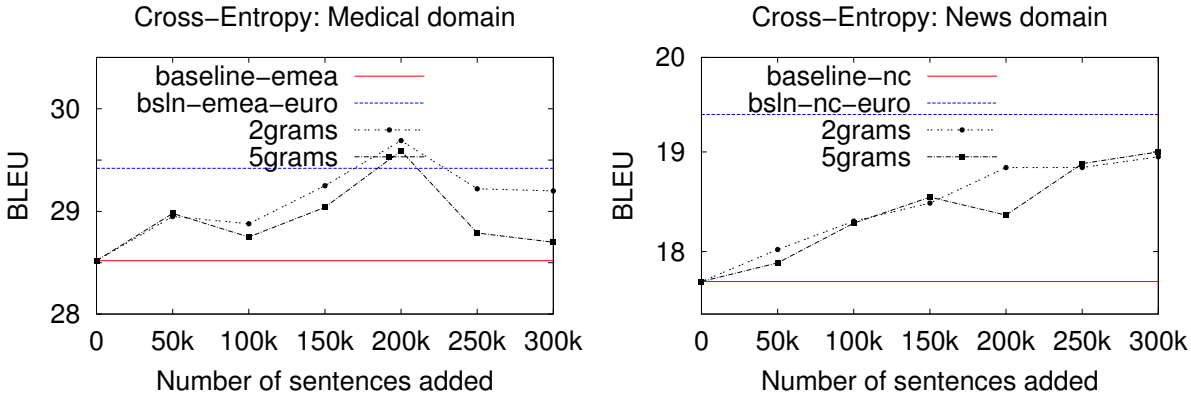

Figure 3: Effect to adding sentences over the BLEU score using cross-entropy strategy (with different n-gram value) for two ID corpora EMEA and News and the Europarl OoD corpus. Horizontal lines represent the score when using the ID corpus and all the data available.

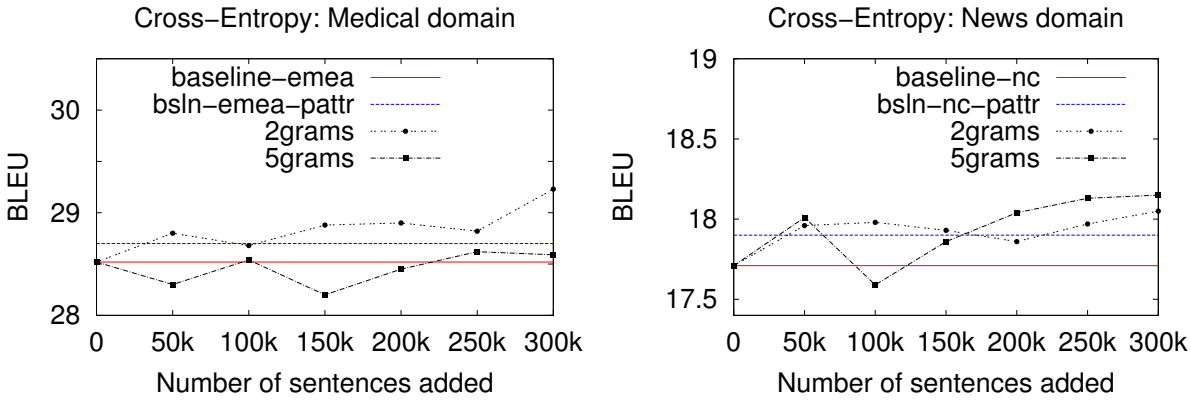

Figure 4: Effect to adding sentences over the BLEU score using cross-entropy strategy (with different n-gram value) for two ID corpora EMEA and News and the PatTR OoD corpus. Horizontal lines represent the score when using the ID corpus and all the data available.

used does not seem to affect significantly the translation quality, although using 2-grams provides slightly better results.

- Lastly, it is also worth noting that the results obtained with the cross-entropy strategy are slightly worse than the ones obtained with infrequent n-gram recovery in all the set-ups analysed, even though more sentences are considered when using cross-entropy.

4.4 Example Translations

Translation examples are shown in Table 3. In the first example, both the infrequent n-gram selection and baseline systems are able to obtain the sing % as appears in the reference. This in not only casual, since, by ensuring coverage for the infrequent n-grams only up to a certain t, we avoid distorting the specificities of the ID data. All the systems present the same lexical choice error with word (*développer*). However, this is so because this is the most likely translation in our data, both ID and OoD. The second example presents a sentence belonging to the NC-test set. None of the systems analysed achieved to produce the correct translation of "republican strategy". However, the "all" system did manage to produce the right reordering, even though the genre in *républicain* was not matched, and then word *á* was introduced instead of "pour". Note that, even if the cross-entropy translation of "counter" is different from the reference, it is semantically equivalent (even though BLEU would penalise it). This is, again, a lexical choice error.

4.5 Summary of the results

Table 4 shows the best results obtained with both strategies for each combination of the ID and OoD corpora. We can see the difference in number of selected sentences be-

Src	about 5 percent of people with ulcerative colitis develop colon cancer .
Bsl	environ 5 % des personnes avec colite ulcéreuse *de développer* un cancer du colon .
All	environ 5 *pour cent* des personnes avec colite ulcéreuse *développer* un cancer du colon .
Infr	environ 5 % des personnes avec colite ulcéreuse de *développer* un cancer du colon .
Entr	environ 5 *pour cent* des personnes avec colite ulcéreuse de *développer* un cancer du colon .
Ref	environ 5 % des personnes souffrant de colite ulcéreuse sont atteintes de cancer du côlon.
Src	a republican strategy to counter the re-election of obama
Bsl	une républicain stratégie pour contrer la réélection d'obama
All	une stratégie républicain á contrer la réélection d'obama
Infr	une républicain stratégie pour contrer la réélection d'obama
Entr	une républicain stratégie pour contrecarrer la réélection d'obama
Ref	une stratégie républicaine pour contrer la réélection d'obama

Table 3: Example of two translations for each of the SMT systems built: Src (source sentence), Bsl (baseline), All (all the data available), Infr (Infrequent n-grams), Entr (Cross-entropy) and Ref (reference).

Data	Strategy	BLEU	$\|S\|$
EMEA-Euro	ID	28.5	1.0M
	all data	29.4	1.0M+1.4M
	cross-entropy	29.7	1.0M+200k
	infreq. $t = 20$	30.2	1.0M+44k
EMEA-PatTR	ID	28.5	1.0M
	all data	28.7	1.0M+3.3M
	cross-entropy	29.2	1.0M+300k
	infreq. $t = 20$	30.2	1.0M+62k
NC-Euro	ID	17.7	117k
	all data	19.4	117k+1.4M
	cross-entropy	19.0	117k+300k
	infreq. $t = 20$	19.4	117k+80k
NC-PatTR	ID	17.7	117k
	all data	17.9	117k+3.3M
	cross-entropy	18.2	117k+300k
	infreq. $t = 20$	18.5	117k+72k

Table 4: Summary of the best results obtained with each setup. `Euro` stands for Europarl and $|S|$ for number of sentences, which are given in terms of the ID corpus size, and (+) the number of sentence selected.

tween infrequent n-grams and cross entropy. The cross-entropy strategy selects more sentences and the results achieved are worse than when using infrequent n-grams. We understand that this is because the infrequent n-grams technique selects more relevant sentences from the OoD corpus.

We observe performance differences between both ID corpora (EMEA and NC). Results obtained with the NC corpus seem to indicate that it is not an adequate corpus for testing adaptation techniques, as observed in our results and also in related work (Haddow and Koehn, 2012; Irvine et al., 2013). Hence, we disrecommend using the NC corpus for adaptation experiments, as it might lead to misleading results.

5 Conclusions and future work

Data selection has been receiving an increasing amount of interest within the SMT research community. In this work, we study the effect of using different data sets with two popular BSS strategies. The results obtained are similar in term of BLEU, although the best results were obtained by infrequent n-grams. Such conclusion is coherent across all combinations of corpora studied, i.e., ID and OoD. Finally, the BSS techniques obtain positive results using only a small fraction of the training data. These results show the importance of the data sets used: it is important to use a general-domain OoD corpus, such Europarl. Moreover, the NC corpus is not appropriate corpus to be used for the evaluation of domain adaptation methods.

In future work, we intend to combine the two strategies proposed and will develop new experiments with bigger and more diverse data sets. In addition, we will also study different data selection methods using a vectorial representation of sentences.

References

Axelrod, A., X. He, and J. Gao. (2011). Domain adaptation via pseudo in-domain data selection. In *Proc. of the EMNLP*, pages 355–362.

Gao, J., J. Goodman, M. Li, and K. Lee.

(2002). Toward a unified approach to statistical language modeling for chinese. *ACM TALIP*, 1:3–33.

Gascó, G., M.A. Rocha, G. Sanchis-Trilles, J. Andrés-Ferrer, and F. Casacuberta. (2012). Does more data always yield better translations? In *Proc. of the EACL*, pages 152–161.

Haddow, B. and P. Koehn. (2012). Analysing the effect of out-of-domain data on smt systems. In *Proc. of the Seventh Workshop on Statistical Machine Translation*, pages 422–432.

Irvine, A., J. Morgan, M. Carpuat, H. Daumé III, and D. Munteanu.(2013). Measuring machine translation errors in new domains. *Transactions of the Association for Computational Linguistics*, 1:429–440.

Kneser, R. and H. Ney. (1995). Improved backing-off for m-gram language modeling. In *Proc. of the International Conference on Acoustics Speech and Signal Processing*, pages 181–184.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, pages 79–86.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. (2007). Moses: open source toolkit for statistical machine translation. In *Proc. of the ACL*, pages 177–180.

Lü, Y., J. Huang, and Q. Liu. (2007). Improving statistical machine translation performance by training data selection and optimization. In *Proc. of the EMNLP-CoNLL*, pages 343–350.

Moore, R. C. and W. Lewis. (2010). Intelligent selection of language model training data. In *Proc. of the ACL*, pages 220–224.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proc. of the ACL*, pages 160–167.

Och, F. J. and H. Ney. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of the ACL*, pages 295–302.

Och, F. J. and H. Ney. (2003). A systematic comparison of various sta-tistical alignment models. *Computational linguistics*, 29:19–51.

Papineni, K., S. Roukos, T. Ward, and W. J. Zhu. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proc. of the ACL*, pages 311–318.

Papineni, K. A, S. Roukos, and R. T. Ward. (1998). Maximum likelihood and discriminative training of direct translation models. In *Proc. of the International Conference on Acoustics Speech and Signal Processing*, pages 189–192.

Rousseau, A. (2013). Xenc: An open-source tool for data selection in natural language processing. *The Prague Bulletin of Mathematical Linguistics*, 100:73–82.

Schwenk, H., A. Rousseau, and M. Attik. (2012). Large, pruned or continuous space language models on a gpu for statistical machine translation. In *Proc. of the NAACL*, pages 11–19.

Sennrich, R. (2012). Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proc. of the EACL*, pages 539–549.

Sennrich, R. (2013). *Domain adaptation for translation models in statistical machine translation.* Ph.D. thesis, University of Zurich.

Stolcke, A. (2002). Srilm-an extensible language modeling toolkit. In *Proc. of the Seventh International Conference on Spoken Language Processing.*

Tiedemann, J.(2009). News from opus- a collection of multilingual parallel corpora with tools and interfaces. In *Proc. of the Recent advances in natural language*, pages 237–248.

Wäschle, K. and S. Riezler. (2012). Analyzing Parallelism and Do-main Similarities in the MAREC Patent Corpus. *Multidisciplinary Information Retrieval*, pages 12–27.

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 109-116 recibido 26-03-2015 revisado 27-04-2015 aceptado 14-05-2015

A Bidirectional Recurrent Neural Language Model for Machine Translation*

Un modelo de lenguaje neuronal recurrente bidireccional para la traducción automática

Álvaro Peris, Francisco Casacuberta
PRHLT – Universitat Politècnica de València
Camino de Vera s/n, 46022, Valencia (Spain)
lvapeab@fiv.upv.es, fcn@prhlt.upv.es

Resumen: Se presenta un modelo de lenguaje basado en representaciones continuas de las palabras, el cual se ha aplicado a una tarea de traducción automática estadística. Este modelo está implementado por una red neuronal recurrente bidireccional, la cual es capaz de tener en cuenta el contexto pasado y futuro de una palabra para realizar predicciones. Debido su alto coste temporal de entrenamiento, para obtener datos de entrenamiento relevantes se emplea un algoritmo de selección de oraciones, el cual busca capturar información útil para traducir un determinado conjunto de test. Los resultados obtenidos muestran que el modelo neuronal entrenado con los datos seleccionados es capaz de mejorar los resultados obtenidos por un modelo de lenguaje de n-gramas.
Palabras clave: Modelado de lenguaje, redes neuronales recurrentes bidireccionales, selección de datos, traducción automática estadística.

Abstract: A language model based in continuous representations of words is presented, which has been applied to a statistical machine translation task. This model is implemented by means of a bidirectional recurrent neural network, which is able to take into account both the past and the future context of a word in order to perform predictions. Due to its high temporal cost at training time, for obtaining relevant training data an instance selection algorithm is used, which aims to capture useful information for translating a test set. Obtained results show that the neural model trained with the selected data outperforms the results obtained by an n-gram language model.
Keywords: Language modelling, bidirectional recurrent neural networks, instance selection, statistical machine translation.

1 Introduction

Many natural language processing applications, such as automatic speech recognition (ASR), handwritten text recognition (HTR) or statistical machine translation (SMT), require the use of a language model, which determines how well a word sequence is formed. The classical approach, the n-gram language model, is a count-based technique in a discrete representation space. Thanks to the smoothing techniques (e.g. Chen and Goodman (1998)), the n-gram models tackle the data sparseness problem, and are capable of obtaining predictions for non-seen events. This leads to simple, robust and fast models, which are may be trained over huge amounts of data. On the other hand, since words are treated as indices in a vector, there are no concepts such as similarity nor semantic relationships between words. In addition, the n-gram model, only considers few context words: Typically, the order of the n-gram models ranges from 2 to 5, therefore the model takes from 1 to 4 context words, and long-term relationships are lost (Bengio et al., 2003). In the last years, more complex language models have been successfully developed. One of these approaches rely in a distributed representation of words: A real-valued, dense and low-dimensional represen-

* The research leading to these results has received funding from the the Generalitat Valenciana under grant Prometeo/2009/014.

ISSN 1135-5948

tation in a continuous space. In these models, probability estimation is carried out in this continuous space, typically by means of a neural network. Furthermore, given the nature of continuous models, the learned function is inherently smoothed.

The proposed model belongs to this latter family and aims to overcome the aforementioned drawbacks of n-gram language models: First, by projecting the words into a continuous space, the model profits from a richer representations of words. Second, by using a bidirectional recurrent neural network (BRNN), the context of a word is determined not only by its preceding words, but also by its following words. This allows the model to produce more informed predictions.

Since the computational cost of such model is high, only a subset of the available training corpora is used to train the model. This can be seen as an instance of a domain adaptation problem, where the goal is to choose the most adequate sentences for translating a given test set from a sentence pool – in this case, the full training corpus. The selection of the training events is performed using a method belonging to the so-called *feature-decay* selection algorithms. The model is then applied to a SMT task, reranking N-best lists of translation hypotheses.

In this paper, we develop an extension of recurrent neural network language models, using a bidirectional neural network for carrying out the probability estimation. We show that this model can be appropriately trained for a given test set using a subset of all available data. This subset of data is chosen using an instance selection algorithm. Results show that the neural model combined with an n-gram language model enhances the performance of a SMT system.

The paper is structured as follows: In Section 2, related approaches are reviewed. In Section 3 the proposed language model is described. Section 4 states the motivation for selecting training instances. Performed experiments and results are shown in Section 5. Finally, conclusions about the work are obtained in Section 6.

2 Related work

The use of continuous spaces is nowadays a hot topic in the language modelling field. Since Bengio et al. (2003) proposed to perform a linear projection from the discrete to the continuous space and learn the probability function in this space, many other works followed these ideas. Bengio et al. and Schwenk (2013) performed the probability estimation through a feedforward neural network. Mikolov (2012) used a recurrent neural network (RNN) for that purpose. In his model, there was no projection layer, words were mapped directly to the hidden layer. Sundermeyer et al. (2012) combined both models, having a projection layer connected to a recurrent layer, with LSTM units. Pascanu et al. (2014) extended the RNN architecture, which led to *deep* RNN, and it was applied to language modelling. In the field of SMT, neural language models have also recent applications: Baltescu et al. (2014) coupled a feedforward neural language model into a SMT decoder. Wang et al. (2014) approximated a neural language model with an n-gram language model, according to bilingual information extracted from the phrase table.

Besides language modelling, continuous spaces have also been included as additional information sources in SMT systems. Sundermeyer et al. (2014) used a bidirectional LSTM network, architecturally similar to the proposed model, as translation model. Devlin et al. (2014) extended the original neural language model from Bengio et al. (2003), and developed a neural translation model. This model could be integrated into a hierarchical decoder and offered impressive results. Moreover, full-neural translation systems have been recently proposed (Bahdanau et al., 2014; Sutskever et al., 2014), offering encouraging results: In Luong et al. (2014), a full-neural system outperformed for the first time a state-of-the-art phrase-based system.

Instance selection techniques have been typically applied in the scope of domain adaptation or active learning. Gascó et al. (2012) showed that a SMT system trained with selected sentences outperformed a system trained with all available data. In this case, the selection criterion was to choose sentences which contained unseen (or seldom seen) n-grams. Other works used perplexity as selection criterion (Mandal et al., 2008). The selection method used in this paper is an instantiation of that proposed in Biçici and Yuret (2011), where the goal is to obtain a selection which maximizes the coverage of

the target part of the test set.

3 Bidirectional Recurrent Language Model

3.1 Recurrent neural networks

Recurrent architecture of neural networks are appropriate to model a temporal-discrete behaviour. Given an input sequence $\mathbf{x}_1^T = x_1, \ldots, x_T$, a RNN produces an output sequence $\mathbf{y}_1^T = y_1, \ldots, y_T$, computed as:

$$\mathbf{h}_t = f_h(x_t, \mathbf{h}_{t-1}) \quad (1)$$

$$\mathbf{y}_t = f_o(\mathbf{h}_t) \quad (2)$$

where $\mathbf{h}_t$ is the hidden state at timestep t, f_h is the hidden state function (e.g. sigmoid or hyperbolic tangent) and f_o is the output function (e.g. a multi-layer perceptron with an output layer performing the *softmax* function).

RNNs are typically trained via stochastic gradient descent, using the backpropagation through time algorithm (Werbos, 1990), to minimize a cost function under some optimality criterion, typically cross-entropy between the output of the system and the training data probability distribution.

3.2 Bidirectional recurrent neural networks

A drawback of regular RNNs is that the input sequence is only scanned in one direction, normally from past to future. In order to capture both past and future context, bidirectional RNNs were proposed by Schuster and Paliwal (1997). The main idea is to have two independent recurrent layers: One layer process the input sequence in forward time direction (from 1 to T), while the other layer process the input sequence reversed in time (from T to 1). Since hidden layers have no interaction between them, bidirectional RNNs can be trained using the same algorithms as those used for unidirectional RNNs. Following prior notation, bidirectional RNN is defined as:

$$\mathbf{h}_t^{\mathrm{f}} = f_h(x_t, \mathbf{h}_{t-1}^{\mathrm{f}}) \quad (3)$$

$$\mathbf{h}_t^{\mathrm{b}} = f_h(x_t, \mathbf{h}_{t+1}^{\mathrm{b}}) \quad (4)$$

$$\mathbf{y}_t = f_o(\mathbf{h}_t^{\mathrm{f}}, \mathbf{h}_t^{\mathrm{b}}) \quad (5)$$

where $\mathbf{h}_t^{\mathrm{f}}$ is the forward layer and $\mathbf{h}_t^{\mathrm{b}}$ is the backward layer. The output is a combination produced by the output function f_o of both backward and forward layers.

3.3 Bidirectional recurrent language model

The task of statistical language modelling consists in estimating the probability distribution over a sequence of words $\mathbf{x}_1^T = x_1, \ldots, x_T$. Applying the chain rule, the sequence probability $p(\mathbf{x}_1^T) = p(x_1, \ldots, x_T)$ can be decomposed as:

$$p(x_1, \ldots, x_T) = \prod_{t=1}^{T} p(x_t | x_1, \ldots, x_{t-1}) \quad (6)$$

In the RNN framework, information about the history $(x_1, \ldots, x_{t-1})$ is represented in the hidden recurrent layer. Thus, sequence probability is rewritten as:

$$p(x_1, \ldots, x_T) = \prod_{t=1}^{T} p(x_t | \mathbf{h}_t) \quad (7)$$

As we move to a BRNN, probability is conditioned by both forward and backward states:

$$p(x_1, \ldots, x_T) = \prod_{t=1}^{T} p(x_t | \mathbf{h}_t^{\mathrm{f}}, \mathbf{h}_t^{\mathrm{b}}) \quad (8)$$

In our language model architecture, input words are one-hot vectors, binary vectors with all elements set to 0 except the index that represents the input word, which is set to 1. Those vectors are projected into the continuous space through a projection layer and then fed to the BRNN. As architectural choices of the network, the hidden function (f_h) is the sigmoid function. The output function (f_o) is modelled with a 2-layer perceptron, which its first layer makes use of the sigmoid activation function and it is fully connected to the output layer, which makes use of the softmax cost function in order to obtain correct output probabilities:

$$\sigma(z_k) = \frac{\exp(z_k)}{\sum_{k'=1}^{K} \exp(z_{k'})} \quad (9)$$

where z_k is the k-th output unit. Each output unit represents a word in the vocabulary, hence, the output layer is vocabulary-sized. At test time, the probability of a sentence is normalized with respect to the length of the

sentence, in order to prevent benefits to short sentences (Graves, 2013). Since using the full vocabulary is computationally unaffordable, a shortlist is used: Only the most K frequent words are taken into account. The rest are mapped to a special token `<unk>`. Figure 1 shows a scheme of the model.

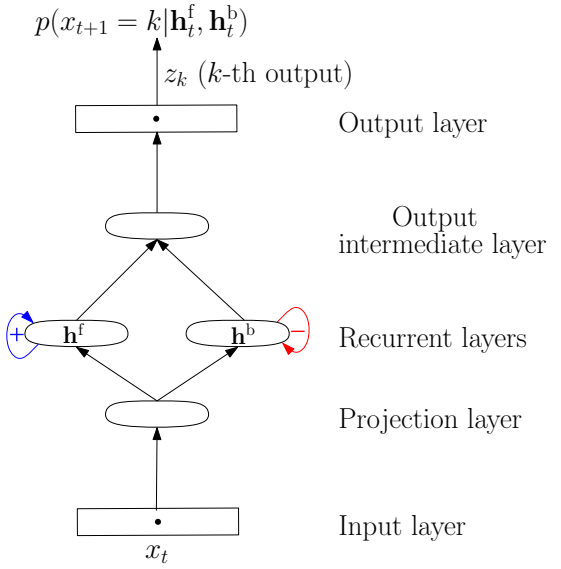

Figure 1: Architecture of the bidirectional language model, similarly depicted as Sundermeyer et al. (2014). Input and output layers have the size of the vocabulary.

4 Instance selection

Typically, corpora used for training SMT systems are much larger than the test set. Therefore, the training set contains noise and sentences that are irrelevant for a specific task. Moreover, neural models have a high time complexity and such large corpora present computational challenges at training time. Techniques for selecting the most appropriate set of training sentences for a given test set are suitable in this scenario. Within these approaches, feature-decay algorithms perform the selection of sentences from the training set aiming to maximize the coverage of the target language n-grams (features) of the test set (Biçici and Yuret, 2011).

The target side of the test set is unknown, only the corresponding source sentences are available. Since a source sentence potentially has many translations, performing a selection which objective is maximizing the coverage of the source part of the test does not guarantee an adequate coverage of the target part of the test. For treating this issue, feature-decay algorithms try to maximize the diversity of the selected instances. The method provides initial scores to the features, according to an initialization function. Iteratively, the features with highest scores are selected and scores of these features included in the selection are reduced. Therefore, it is expected that, in following iterations, different features will be included in the selection. Particular choices in the initialization, scoring and decaying functions provide different selection methods, such as n-gram coverage or TF-IDF (Eck et al., 2005).

FDA5, a parametrization of feature-decay algorithms has been recently proposed (Bicici and Yuret, 2015). In this algorithm, the selection of the data is performed according to 5 parameters: The feature initialization function considers frequency and inverse frequency of the tokens in a feature. Scores of the features decay following a polynomial and an exponential factor. Finally, the sentence scoring function is the sum of all feature values of a sentence, scaled by a sentence-length factor.

5 Experiments and results

The model was tested in the Spanish–English *EU* translation task – a selection from the *Bulletin of the European Union* (Khadivi and Goutte, 2003). The Thot toolkit (Ortiz-Martínez and Casacuberta, 2014) was used for building the translation models. The neural language model was used to rescore N-best lists, which were obtained executing a weight adjustment process, by means of the downhill simplex optimization method (Melder and Nead, 1965), using BLEU as function to maximize. At each iteration of the optimization process, a 200-best list was generated and merged with the list of the previous iteration. The process continued until no new elements were included in the N-best list. As result, the majority of the probability mass of translation hypotheses was included in the N-best list. The average size of the lists was $N = 4300$.

5.1 Data selection

For selecting an appropriate number of sentences, different selection sizes were tested. An n-gram language model was trained over the target side of the selected data and its perplexity was computed. We chose a selection which provide a good balance be-

tween complexity and quality. Figure 2 shows the relation between the number of source words selected and the number of training sentences selected (corpus coverage). If a word is included in the selection, all the sentences which contain such word belong to the selection. Figure 3 shows bigram coverage for the test set, according the number of source words selected. Finally, Figure 4 shows the number of out-of-vocabulary (OOV) words in the test test set with respect to the number of words selected. It was observed that performing a selection from one million words ahead, produced small variations both in test OOV words and in bigram coverage values.

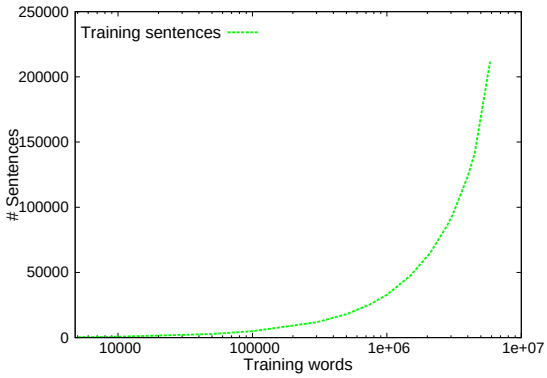

Figure 2: Training corpus coverage according the number of source words selected.

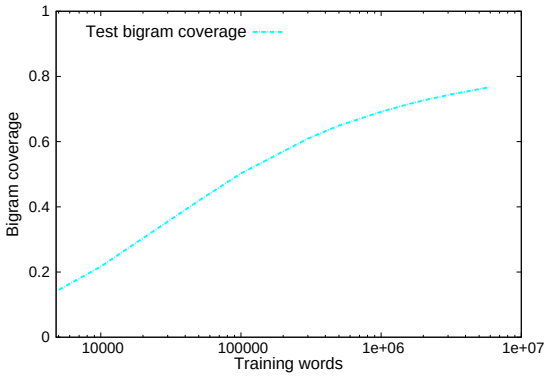

Figure 3: Relation between the test set bigram coverage of the selected corpus and the number of source words selected.

Figure 5 reports the perplexity obtained by different n-gram language models trained over the selected data. In order to obtain a fair comparison between language models, for computing perplexity, all neural and n-gram language models were trained using the same vocabulary. 4-gram performed slightly better than 5-gram when the full corpus was not selected. Thus, in following experiments, the

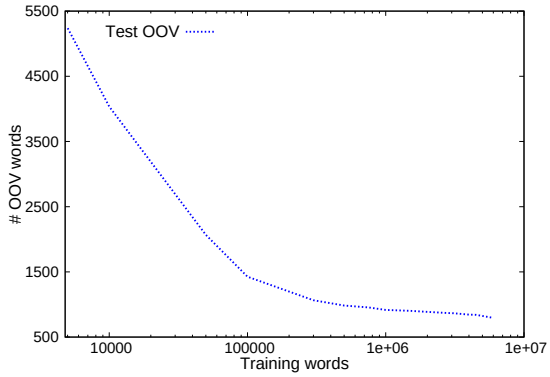

Figure 4: Number of OOV words in the test set according the number of words selected.

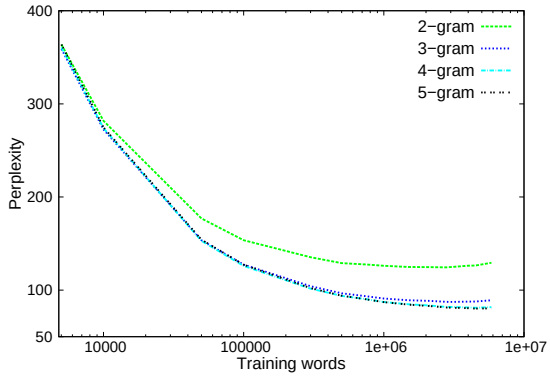

Figure 5: Perplexity obtained by different n-gram language models according the number of selected words.

order of the n-gram language model was set to 4. These results corroborated those given by the OOV words and bigram coverage: For selections larger than a million of words, obtaining a humble perplexity enhancement was expensive, as the number of instances that should be included rapidly grew.

Therefore for obtaining the reduced corpus, FDA5 considered 1 million words from the source corpus. Statistics of the selected instances are shown in Table 1. The obtained corpus contained only a 15% sentences of the original corpus, but the perplexity of the models trained over it, for the test set, was increased just a 6.4% with respect to the full corpus.

5.2 Language models

The hyperparameters of the neural language model (size of the projection layer, size of the recurrent layers and size of the output intermediate layer), were chosen following a perplexity minimization criterion. The learning rate was initially set to 1 and was halved at

		En	Sp
Full training set	Sentences	213k	
	Running words	5.2M	5.9M
	Vocabulary	50k	64k
Reduced training set	Sentences	33k	
	Running words	891k	1M
	Vocabulary	21k	28k
Test set	Sentences	800	
	Running words	20k	23k

Table 1: Statistics for full EU corpus, selected instances and test set (k and M stand for thousands and millions, respectively). The reduced corpus was obtained selecting one million words from the source part (Spanish) of the training set.

the start of each training epoch if the validation entropy did not decrease a 0.3% with respect the previous one (Mikolov, 2012). Following Pascanu et al. (2014), the network was initialized using the standard deviations of a Gaussian white noise distribution. The size of the shortlist was set to $K = 10,000$. An analogous unidirectional model was also trained.

Table 2 shows the perplexities obtained by the different language models over the test set. The bidirectional RNN language model offered a performance similar to that of the n-gram language model trained with the same data, while the perplexity of the unidirectional RNN was slightly higher.

Language model	**Perplexity**
Full n-gram	81.7
Reduced n-gram	87.3
Unidirectional RNN	95.6
Bidirectional RNN	87.9

Table 2: Test set perplexity for different language models. Full n-gram row refers to an n-gram trained over the complete corpus. Reduced n-gram refers to an n-gram trained only over the selected instances. Both neural models are trained over selected instances. All models were trained using the same vocabulary (10,000 words).

Table 3 shows the BLEU scores obtained by the different language models for the test set. The bidirectional model offered a small improvement with respect the unidirectional one, but both were worse than the n-gram language model. The neural language model was also linearly interpolated with an n-gram language model. The interpolation coefficient (λ) was determined by sampling in a development set. The sampling interval was $[0.1, 0.9]$, with a step of 0.1. The optimal value was found at $\lambda = 0.6$. This interpolation provided an enhancement of the system performance. That means that both approaches were complementary: Because of their nature, n-gram language models are robust modelling local dependencies. The neural network introduced additional information, which was useful in order to enhance the performance of the system. Although differences in the results obtained were statistically non-significant, we observed a trend in them.

Language model	**BLEU**
n-gram	30.8
Unidirectional RNN	30.2
Bidirectional RNN	30.3
Bidirectional RNN + n-gram	**31.3**

Table 3: Test set BLEU score for the different language models.

6 Conclusions

In this paper, a neural language model implemented by means of a bidirectional neural network has been presented. Since the computational training cost of the model was high, a subset of the training corpus was selected, using domain adaptation techniques. The network was successfully trained with the reduced corpus, obtaining a perplexity similar to that of an n-gram language model trained with the full corpus. It was shown that both neural and n-gram language models are complementary: The latter model was focused in local dependencies, while the first one was able to incorporate additional dependencies. That was supported by the results: The combination of both models provided enhancements with respect the use of the models solely.

Finally, the bidirectional architecture of the neural network language model exhibited a better behaviour than the unidirectional architecture, in terms of perplexity and translation quality.

References

Bahdanau, D., K. Cho, and Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. Technical report, arXiv preprint arXiv:1409.0473.

Baltescu, P., P. Blunsom, and H. Hoang. 2014. OxLM: A neural language modelling framework for machine translation. *The Prague Bulletin of Mathematical Linguistics*, 102(1):81–92.

Bengio, Y., R. Ducharme, P. Vincent, and C. Jauvin. 2003. A neural probabilistic language model. *Machine Learning Research*.

Biçici, E. and D. Yuret. 2011. Instance selection for machine translation using feature decay algorithms. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 272–283. Association for Computational Linguistics.

Bicici, E. and D. Yuret. 2015. Optimizing instance selection for statistical machine translation with feature decay algorithms. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 23(2):339–350.

Chen, F. and J. Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard U., Cambridge, MA.

Devlin, J., R. Zbib, Z. Huang, T. Lamar, R. Schwartz, and J. Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1370–1380. Association for Computational Linguistics.

Eck, M., S. Vogel, and A. Waibel. 2005. Low cost portability for statistical machine translation based on n-gram frequency and TF-IDF. In *International Workshop on Spoken Language Translation (IWSLT)*, pages 61–67.

Gascó, G., M. A. Rocha, G. Sanchis-Trilles, J. Andrés-Ferrer, and F. Casacuberta. 2012. Does more data always yield better translations? In *Proceedings of the 13th European Chapter of the Association for Computational Linguistics*, pages 152–161.

Graves, A. 2013. Generating sequences with recurrent neural networks. *arXiv:*1308.0850 [cs.NE].

Khadivi, S. and C. Goutte. 2003. Tools for corpus alignment and evaluation of the alignments (deliverable d4.9). Technical report, Technical report, TransType2 (IST-2001-32091).

Luong, T., I. Sutskever, Q. V. Le, O. Vinyals, and W. Zaremba. 2014. Addressing the rare word problem in neural machine translation. arXiv preprint arXiv:1410.8206.

Mandal, A., D. Vergyri, W. Wang, J. Zheng, A. Stolcke, G. Tur, D. Hakkani-Tur, and N. F. Ayan. 2008. Efficient data selection for machine translation. In *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*, pages 261–264. IEEE.

Melder, J. A. and R. Nead. 1965. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313.

Mikolov, T. 2012. *Statistical Language Models based on Neural Networks*. Ph.D. thesis, Brno University of Technology.

Ortiz-Martínez, D. and F. Casacuberta. 2014. The new Thot toolkit for fully-automatic and interactive statistical machine translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 45–48.

Pascanu, R., Ç. Gülçehre, K. Cho, and Y. Bengio. 2014. How to construct deep recurrent neural networks.

Schuster, M. and K. K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Schwenk, H. 2013. CSLM - a modular open-source continuous space language modeling toolkit. In *INTERSPEECH*, pages 1198–1202. ISCA.

Sundermeyer, M., T. Alkhouli, J. Wuebker, and H. Ney. 2014. Translation modeling with bidirectional recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural*

Language Processing (EMNLP), pages 14–25. Association for Computational Linguistics.

Sundermeyer, M., R. Schlüter, and H. Ney. 2012. LSTM neural networks for language modeling. In *Interspeech*, pages 194–197.

Sutskever, I., O. Vinyals, and Q. V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS 2014)*.

Wang, R., H. Zhao, B-L. Lu, M. Utiyama, and E. Sumita. 2014. Neural network based bilingual language model growing for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 189–195. Association for Computational Linguistics.

Werbos, P. J. 1990. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560.

Análisis de sentimientos y opiniones

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 119-126 recibido 15-03-2015 revisado 29-04-2015 aceptado 08-05-2015

Enriching User Reviews Through An Opinion Extraction System *

Enriqueciendo revisiones de usuarios mediante un sistema de extracción de opiniones

F. Javier Ortega Universidad de Sevilla javierortega@us.es
José A. Troyano Universidad de Sevilla troyano@us.es
Fermín L. Cruz Universidad de Sevilla fcruz@us.es
Fernando Enríquez Universidad de Sevilla fenros@us.es

Resumen: Las webs basadas en el contenido generado por usuarios (UGC) tienen una aplicabilidad potencial en un gran número de campos. En este trabajo realizamos un estudio de la utilidad de estos sistemas para determinar la percepción de los usuarios expresada en sus opiniones sobre productos o servicios. Para ello, hemos compilado y analizado opiniones compartidas por usuarios en TripAdvisor, centrándonos en dos aspectos: el contenido estructurado y el no estructurado. Hemos realizado un análisis cuantitativo y cualitativo de la información extraída por un sistema de minería de opiniones, siendo este último especialmente interesante ya que ofrece información valiosa sobre los puntos fuertes y débiles de los hoteles según la percepción de los usuarios, yendo más allá de la información estructurada. Por último, hemos realizado un estudio de la complementariedad de la información estructurada y la no estructurada, observando un gran incremento de la cantidad de información disponible conjuntando ambas.
Palabras clave: minería de opiniones, contenidos generados por usuarios

Abstract: Web sites based on User-Generated Content (UGC) have a potentially valuable applicability in a number of fields. In this work we carry out a study of the usefulness of these systems from the point of view of detecting the perception expressed by users about services or items. We have compiled and analyzed opinions shared by users on TripAdvisor focusing on two aspects: the structured and the unstructured data. We perform a quantitative and a qualitative analysis of the information extracted by an opinion extraction system from our dataset, being the last one especially interesting since it provides valuable knowledge about the strong and weak points of hotels according to user perceptions, going beyond the structured data. Finally, we provide a study on the complementarity of the knowledge extracted from both, the textual opinions and the structured data, observing a noticeable increment of the amount of information available with the conjunction of both sources.
Keywords: opinion mining, user-generated content

1 Introduction

Review websites have become a useful Web 2.0 tool for on-line customers in their decision making process in order to gather information about a specific service or item before purchasing it. These websites usually integrate a recommender system intended to offer the adequate product to each user, based on the previous opinions of users about similar products encoded in numerical ratings provided by users, or based on the opinions of other similar users about the same products. With the emergence of Opinion Mining, the analysis of textual opinions can be also a useful tool in this field. In this work we try to answer the research question about the extent in which Opinion Mining analysis can contribute to improve the quality of information that this type of systems can provide to their users.

In this sense, we distinguish in this work between the structured and the unstructured

* This work has been partially funded by the research projects AORESCU (P11-TIC-7684, Consejería de Innovación, Ciencia y Empresas, Junta de Andalucía), DOCUS (TIN2011-14726-E, Ministerio de Ciencia e Innovación) and ACOGEUS (TIN2012-38536-C03-02, Ministerio de Economía y Competitividad).

ISSN 1135-5948

data provided by users. The first one is usually guided by the interfaces of these reviews sites and gathered through "Likes/Dislikes" schemata or "Stars" or any other form of asking the user for a numerical rating about the item being opinionated. The nature of this type of information makes it easier to process. On the other hand, the unstructured data consists mostly of textual opinions written by users in natural language, usually without any kind of predefined pattern. For this reason, we need to pre-process this information in order to extract useful knowledge from it.

Among the diversity of topics being covered by review websites, maybe one of the most relevant in terms of industry and economy of many countries is tourism. This is the main reason to focus on this domain for our study. Another reason is the huge impact of these kind of on-line systems on tourism industry recently, as stated in many works (Buhalis and Law, 2008; Ye, Law, and Gu, 2009; Nieves and Haller, 2014; Yas-vari, Ghassemi, and Rahrovy, 2012; Noti, 2013; Ayeh et al., 2012), to the point that a new line of research, e-Tourism, has come up to address the opportunities and challenges arose from it.

Regarding the research on e-Tourism, there is a wide variety of works, covering the analysis of the trustworthiness that these new channels offer to the users (Cox et al., 2009; Munar and Jacobsen, 2013), the study of the eWord-of-Mouth (eWOM) phenomenon (Yasvari, Ghassemi, and Rahrovy, 2012; Barbagallo et al., 2012), the analysis of the influence of on-line reviews on the number of hotel room bookings (Ye et al., 2011), recommendation systems on tourism (Kabassi, 2010; Goossen et al., 2013), or even the development of systems for tourism packaging (Agarwal et al., 2013).

In this work we perform a study of a dataset composed by user opinions written in Spanish about hotels in the Canary Islands (Spain) extracted from one of the most relevant websites on this topic: TripAdvisor[1]. The study includes the evaluation of both, structured and unstructured data provided by users through TripAdvisor, analyzing the correlation between both types of information and their complementarity, in such way that we can measure the extent in which the knowledge provided by an opinion extraction system can enrich the user reviews.

The rest of the paper is structured as follows. The compilation of the dataset object of our study is discussed in section 2. In section 3 we briefly introduce TOES (Cruz et al., 2013), the opinion extraction system used in our work. In section 4 we discuss the relations between the structured and the unstructured data from the dataset and the output provided by the opinion extraction system, from three points of view: a quantitative analysis, a qualitative analysis and a study of the complementarity between both types of information. Finally, we point out the conclusions and future work in section 5.

2 Dataset Compilation

The selection of sources for our dataset has been guided by the relevance of the websites in the area and the amount of information that could be retrieved from them. So, we have chosen TripAdvisor over others because it is one of the most widely used tourism-related website. We decided to work with user-generated reviews about hotels in a specific location, in this case the Canary Islands, due to their particularities as the unique subtropical area in Europe and the importance of the tourism industry in their economy, which assures a huge amount of hotels and user-generated reviews of them, with a high variety of tourists with different needs and perceptions. Other relevant characteristic of TripAdvisor is the fact that any user is allowed to write a review about any item in the system (hotels, restaurants, etc.) with the only requirement of indicating (by clicking on a checkbox) that they have been there. Such a relaxed policy guarantees the provision of a high amount of user-generated content, in spite of the possible detriment in the quality and veracity of the reviews which are out of the scope of this work. We performed a search-driven crawling from TripAdvisor, given that our aim is to gather as much information as possible about hotels in a specific location, as follows:

1. Perform a search against the website with the required location.
2. Retrieve the list of hotels registered in the website for the given location.
3. For each hotel, retrieve all the structured information and the opinions of users.

[1] www.tripadvisor.com

We repeat these steps for each island, so our crawler obtained all the hotels located in the Canary Islands together with all the opinions in the system about them. Since we are interested in the characterization of the hotels and not in the creation of complete user profiles, we just retrieve the information about the opinions of the users about those hotels, leaving out the opinions of those users about hotels in other places.

The crawler has been implemented in Java using two well-known libraries in this task: *HtmlUnit*[2] and *WebHarvest*[3].

Metrics	Total	Spanish
Hotels	403	381
Reviews	78,535	12,950
Revs./Hotel	194.87	33.98
Users	68,441	11,039
Revs./User	1.14	1.17
Sentences	308,998	90,234
Sents./Rev.	3.93	6.96
Words	7,122,747	2,406,330
Words/Rev.	90.69	185.81

Table 1: Size of our dataset in terms of number of reviews, hotels and users, in addition to the number of sentences and words in the documents. The third column contains the same metrics applied only to those reviews written in Spanish.

The resulting resource after the execution of the above mentioned crawler is a dataset formed by structured and unstructured data about hotels in the Canary Islands and user reviews written in 2012 about those hotels in TripAdvisor. The structured data about the hotels consists of: name of the hotel, category (in the range of 0-5 stars), location, and the average of the scores provided by the users. About the opinions, we have gathered the user who wrote the opinion, the origin of the user, the profile (whether the user has traveled "solo", i.e. alone, with friends or with family), the textual opinion, and a set of detailed scores given by the users to six specific features: location, service, comfort, cleanliness, rooms and quality of the hotel, in addition to the overall score of each hotel. Table 1 contains some metrics of the resulting dataset distinguishing the whole collection and the subset formed by reviews written by Spanish users. In the table we show the number of hotels, reviews and users, the amount of textual information retrieved in terms of the number of sentences and words within the reviews and their average per review.

As a simple way of validating the compiled dataset, in Table 2 we show a comparison among the origin of tourists in the Canary Islands according to the gathered reviews and an official study carried out by a government institution, ISTAC [4], in the same period of the reviews in our dataset (2012). This official study consists of a personal interview to tourists in the main airports of the islands.

Origin	ISTAC	TripAdvisor
Germany	25.98%	2.49%
Belgium	2.74%	1.21%
France	2.79%	2.53%
UK	22.04%	51.06%
Netherlands	3.50%	0.51%
Ireland	1.50%	2.95%
Italy	1.95%	3.24%
Spain	22.31%	16.49%
Others	17.18%	19.53%

Table 2: Percentage of opinions in each resource according to the origin of users. The first column corresponds to the official statistics compiled by ISTAC in 2012, the second one corresponds to data from TripAdvisor.

As shown in Table 2, we can see that most of the users that write opinions in TripAdvisor about hotels in the Canary Islands are from the United Kingdom and Spain. On the other hand, the official statistics from ISTAC show that German, Spanish and British tourists add up to about 70% of the total number of opinions. Although there are some differences, these can be caused by the different nature of the compilation methods used by both sources. Nevertheless, in general we can see that our dataset is qualitatively comparable to the one from ISTAC, meaning that it can be considered a good sample of the tourists in the Canary Island.

3 Domain-adaptable Opinion Extraction System: TOES

The aim of this research work is to study the extent in which an opinion extraction system can be useful in order to enrich the

[2] `http://htmlunit.sourceforge.net`

[3] `http://web-harvest.sourceforge.net`

[4] Instituto Canario de Estadística, the Canary Islands Government `http://www.gobiernodecanarias.org/istac/temas_estadisticos/sectorservicios/hosteleriayturismo/demanda/`

user opinions within a reviews website. To that end, we have compiled a dataset containing a huge amount of user reviews about hotels. Next we need an opinion extraction system focused on this domain. Thus, we take advantage of TOES (Cruz et al., 2013), a domain-adaptable opinion extraction system that can be easily applied to our study.

TOES is intended to detect and classify the opinions in a text. The underlying idea is to capture knowledge about a particular product class and the way people write their reviews on it. This process consists in two phases: first, it detects the pieces of text expressing individual opinions about specific features of the item being opinionated; in the second step, it computes the polarity of each individual opinion and the intensity of the polarity, and assigns a score in the range [-1,1], representing -1 the most negative polarity and 1 the most positive.

TOES needs a training phase where a set of resources adapted to the domain are built. Some resources are automatically induced from a corpus of annotated reviews, while others are manually generated by an expert with some computational assessment. The training corpus is tagged by an expert aided by TOES. A taxonomy is built from the *feature words*, defining the characteristics that users are expected to write about. Using the taxonomy and the annotations of the expert, TOES builds a set of domain-dependent resources which are used for the detection and classification of opinions.

In our case TOES has been trained using a set of user-generated hotel reviews in Spanish extracted from TripAdvisor. This training set is formed by randomly chosen hotels from touristic Spanish cities like Madrid, Mallorca, Seville, etc., explicitly excluding the Canary Islands, so none of the hotels in the training dataset appear in our original dataset. Some metrics of the training set are shown in Table 3, including the number of annotated opinions that users express in their reviews.

Number of Reviews	1,200
Number of Words	213,843
Words per review	178.20
Annotated opinions	7,720

Table 3: Statistics of the reviews in the training set. The annotated opinions are the features commented by users in their reviews.

Once the domain-dependent resources have been created, TOES can extract user opinions from other texts on the same domain and also classify the polarity for each opinionated feature, determining whether the user expresses a positive or a negative opinion. Specifically, TOES provides, for each textual opinion, the set of features within that text in addition to the opinion words referring to the feature and the polarity of the opinion. In Figure 1 we can see a pair of input text and its corresponding output as an example.

INPUT:

Excelente ubicación para olvidarte del mundo. El personal es encantador. La piscina excelente. El restaurante, a pesar de tener una buena cocina falta variedad, por ejemplo en el desayuno no hay ni croissant, no hay opción de bebidas calientes (café) sino no esta abierto el restaurante y está cerrado por la tarde hasta las 19:00.

TOES OUTPUT:

1, 1, 0.050, 0.950, Excelente ubicación para olvidarte del mundo
1, 1, 0.032, 0.968, El personal es encantador
1, 1, 0.050, 0.950, La piscina excelente
1, 0, 0.991, 0.009, El restaurante, a pesar de tener una buena cocina falta variedad, por ejemplo en el desayuno no hay ni croissant, no hay opción de bebidas calientes (café) sino no esta abierto el restaurante y está cerrado por la tarde hasta las 19:00.

Figure 1: Output provided by TOES given the input text. The columns correspond to the identification of the document, the polarity of the opinion and the negative and positive scores, respectively, computed by TOES.

For more details on the performance of TOES on other domains and a thoroughly explanation of its characteristics, the interested reader can review (Cruz et al., 2013).

4 Enrichment of user reviews

Once our dataset is processed by the opinion extraction system, let us proceed to the study of the application of these results. We evaluate the contribution of the opinion extraction system to the user reviews from two points of view: quantitative and qualitative. Finally we study the contribution of the opinion extraction system in terms of knowledge gain. In order to perform these evaluations properly, we have manually mapped the categories offered by TripAdvisor to the taxonomy used by TOES (see Table 4).

4.1 Quantitative evaluation

From a quantitative stance, we show in this section an evaluation based on the compar-

TripAdvisor	TOES
Quality	Building, Hotel, Price
Comfort	Bed
Rooms	Rooms, Television, Bathroom, Facilities
Cleanliness	Cleanliness
Location	Location, Views
Services	Services, Staff, Internet, Food/Drink

Table 4: Mapping between the feature taxonomies of TOES and TripAdvisor.

ison of the information extracted by TOES and the structured information provided by users in the review websites. The aim of this evaluation is to assess the correlation between both types of information

After applying TOES to the textual opinions from our dataset, we can highlight some conclusions. First, we plot in Figures 2 and 3 a comparison between the distributions of frequencies of the scores in TripAdvisor and the textual opinions extracted by TOES, respectively, showing the number of hotels (x-axis) with respect to the number of opinions (y-axis) about each feature.

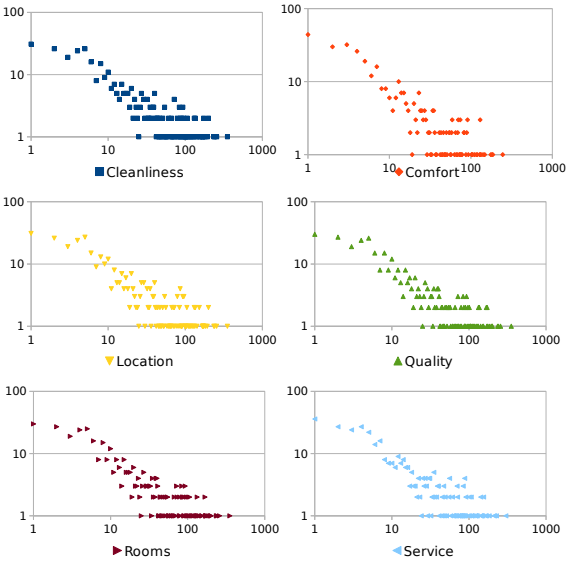

Figure 2: Number of numerical ratings (log-log) provided by users in TripAdvisor for each pre-established feature: Cleanliness, Comfort, Location, Quality, Rooms and Service.

In Table 5 we show a comparison between the scores given by users to each feature in the TripAdvisor taxonomy and the information extracted by TOES from the textual reviews of the users. This table has been computed by aggregating the count of opinions

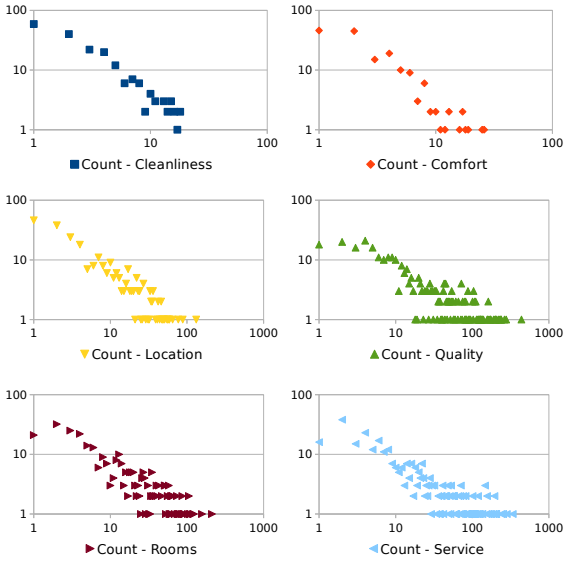

Figure 3: Number of opinions (log-log) provided by users in their textual reviews, according to TOES, for each pre-established feature: Cleanliness, Comfort, Location, Quality, Rooms and Service.

	TripAdvisor		TOES	
Features	**Pos.**	**Neg.**	**Pos.**	**Neg.**
Cleanliness	79.07%	20.93%	78.98%	21.02%
Comfort	63.95%	36.05%	66.41%	33.59%
Location	75.79%	24.21%	86.51%	13.49%
Quality	79.39%	20.61%	80.19%	19.81%
Rooms	78.72%	21.28%	81.93%	18.07%
Services	82.90%	17.10%	84.83%	15.17%
Average	76.64%	23.36%	79.81%	20.19%

Table 5: Percentage of positive and negative opinions per feature according to the scores in TripAdvisor(columns 2 and 3) and TOES (columns 4 and 5).

per feature for all the hotels in our dataset. Note that TripAdvisor allows its users to provide a score in the range [0-5] (stars) to each feature. We have considered as negative those scores < 3.

The columns corresponding to TripAdvisor scores have been obtained by computing the average of the scores given by users to each feature of the hotels in the dataset, and analogous for the scores in the columns corresponding to TOES. Regarding the data in Table 5, we observe that the overall results obtained by TOES from the textual opinions are fairly close to those expressed by users through the scores. In fact, the average of the differences between the scores from TripAdvisor and the polarity of opinions computed

by TOES is only $3,01\%$.

4.2 Qualitative evaluation

Through the quantitative evaluation we assess the reliability of TOES by comparing its results to the scores provided by the authors of the reviews to each feature in the TripAdvisor taxonomy. In this section, through the qualitative evaluation we show the capability of TOES of providing a finer-grained information by taking the analysis to the word-level of the specific terms that authors use to express their opinions.

With respect to the vocabulary used by users to express their opinions in the reviews, we can determine those words which are more frequently used in any, positive or negative opinions. One of these sets of words are represented through a word-cloud in Figure 4, corresponding to the most frequent feature words found in negative opinions.

Figure 4: Cloud of the most frequent feature words mentioned in negative user opinions, according to TOES.

From the figure we can infer that "comida" (food), "buffet" and "spa" receive most of the negative comments (obviously the word "hotel" is common in this domain for both, positive and negative opinions). On the other hand, in Table 6 we use another representation in order to highlight the top ten most used words in positive reviews.

These type of analyses can be performed for each hotel, being a useful tool for the providers of items being opinionated, in this case tourist services, in order to detect the pros and cons of the items provided in a finer grain than the one usually offered by the reviews website.

Words	Count
hotel	4,488
personal	2,656
habitacion	3,792
comida	2,006
piscina	1,584
servicio	979
trato	934
buffet	686
zona	624
limpieza	590

Table 6: Top 10 most used words in positive reviews in our dataset according to TOES.

4.3 Complementarity of informations

Given the nature of the structured information and the usability of the methods intended to gather it (usually the user must click on a number of stars or something analogous) in contrast to the more laborious activity of actually writing a text, we expected that most of the reviews contain a numerical value for the features proposed by TripAdvisor, while a smaller amount of them will include a proper written opinion. Table 7 shows the percentage of reviews without scores for each one of the features proposed by TripAdvisor, and also the percentage of reviews without textual opinions extracted by TOES for each feature.

Surprisingly, a higher percentage than expected of users do not provide numerical scores to all the features proposed by TripAdvisor. The case of *Comfort* is shocking: only 19.63% of reviews have a score, and less than 25% of them contain an opinion about it. The question now is: how the unstructured information can help to improve this lack of coverage of the structured one? In Figure 5 we plot a comparison of the percentage of reviews that contain structured information (scores) and written opinions about each feature in TripAdvisor, in addition to the union and intersection of both sets.

The most interesting observation in Figure 5 is provided by the last two columns of each feature: $TOES \cup TripAdvisor$ represents the percentage of user reviews with either, a score or a written opinion about the feature, while the column tagged as $TOES \cap TripAdvisor$ represents the percentage of user reviews that have both types of information. In other words, they correspond

Features	No Scores	No Text. Ops.
Cleanliness	58.80%	75.31%
Comfort	80.37%	76.42%
Location	58.99%	60.59%
Quality	58.53%	24.81%
Rooms	58.49%	47.54%
Service	68.60%	24.49%

Table 7: Percentage of reviews without scores or textual opinions in TripAdvisor, respectively, for the given features over the total of 12,950 reviews in Spanish in our dataset.

to the union and the intersection of those sets, respectively. These metrics highlight the improvement achieved by the inclusion of an automatic opinion mining tool like TOES in the system. Furthermore, the intersection of both sets is smaller than expected with only about 20% of user reviews, which means that, most of the times, users tend to provide only one type of information for each feature. Since there are a higher percentage of written opinions per feature than scores (except for *Cleanliness* and *Location*), we can state that a high percentage of users tend to score those features that they have not commented on.

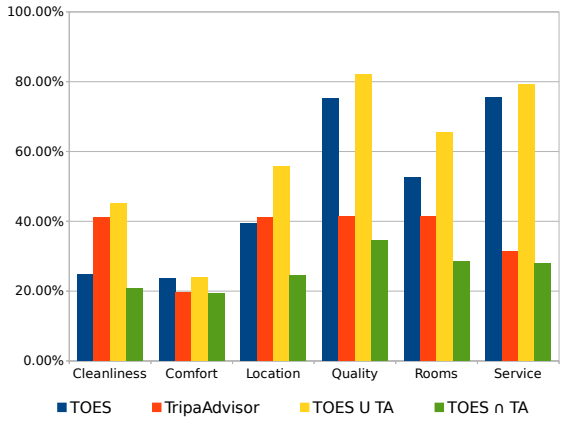

Figure 5: Percentage of reviews with numerical scores (TripAdvisor) and written opinions (according to TOES), together with the conjunction and the intersection of both sets, $TOES \cup TA$ and $TOES \cap TA$, respectively.

Following this idea, we can extract the *complementarity* of both, structured information in the form of scores in TripAdvisor and unstructured information in the form of textual opinions. We define the *complementarity* as the increment of information with respect to the total reviews of each feature. Table 8 contains the results of this metric for each feature in terms of the percentage of reviews without scores for the feature considered but with opinions extracted by TOES, with respect to the total number of reviews with relevant information for each feature.

Features	Str Inf	Str+Unstr	Compl.
Cleanliness	41.20%	45.21%	5.10%
Comfort	19.63%	23.92%	17.88%
Location	41.01%	55.84%	14.92%
Quality	41.47%	82.20%	26.89%
Rooms	41.51%	65.44%	19.22%
Service	31.40%	79.12%	47.13%

Table 8: Percentage of reviews with scores (Str Inf), reviews with scores or textual opinions (Str+Unst) and the complementarity of both (Compl.): increment of reviews with respect to the total reviews of each feature.

We can see in the table the noticeable complementarity of both sources for all the features proposed by TripAdvisor. According to this, in order to make a reliable recommender system based on user reviews, it should be mandatory to implement an opinion mining tool in order to make the most of the information provided by users, given that this unstructured information supports and even complements the structured data, providing a very useful source of additional knowledge about user opinions.

5 *Conclusions*

The e-Tourism research has an increasing interest on Opinion Mining and Recommendation Systems, since these fields can provide very valuable advances in the study of customers perceptions about the products or services enjoyed. In this work we have carried out a study intended to highlight in what extent an opinion extraction system can enrich or improve the information provided by users. In this sense, we have performed a quantitative study about the correlation between the numerical ratings and the knowledge extracted from textual opinions of users, in order to check whether these sources express similar perceptions. In our case, the correlation between the ratings in TripAdvisor and the opinions extracted from the textual reviews is very clear, with a 3% of difference in average between both types of information. On the other hand, we have performed a qualitative analysis of the output of an opinion extraction system in terms of the added-value obtained by the analysis of a finer grained and more detailed information

present in the textual opinions and not in the numerical ratings. Finally, we have studied the complementarity of both sources of information, obtaining a metric representing the contribution of the analysis of textual opinions to the structured information, showing that a resource built from both sources contains up to 47% more opinions on some features than using just the numerical ratings.

We plan to further our work by developing a method to automatically integrate structured and unstructured information in an aspect-based recommendation system, in addition to the study of the integration of multilingual opinion extraction systems to take advantage of the huge amount of textual opinions in diverse languages.

References

Agarwal, J., R H Goudar, N. Sharma, P. Kumar, V. Parshav, R. Sharma, and S. Rao. 2013. Cost effective dynamic packaging systems in e-tourism using semantic web. *Interna-tional Conference on Advances in Com-puting, Communications and Informatics*, pages 1196–1200.

Ayeh, J. K., D. Leung, N. Au, and R. Law. 2012. Perceptions and strategies of hospitality and tourism practitioners on social media: An exploratory study. In Matthias Fuchs, Francesco Ricci, and Lorenzo Cantoni, editors, *Information and Communication Technologies in Tourism*, Vienna. Springer Vienna.

Barbagallo, D., L. Bruni, C. Francalanci, and P. Giacomazzi. 2012. An empirical study on the relationship be-tween twitter sentiment and influence in the tourism domain. In *Information and Communication Technologies in Tourism*. pages 506–516.

Buhalis, D. and R. Law. 2008. Progress in information technology and tourism management: 20 years on and 10 years after the internet-the state of etourism research. *Tourism Management*, 29(4):609–623.

Cox, C., S. Burgess, C. Sellitto, and J. Buultjens. 2009. The role of user-generated content in tourists' travel planning behavior. *Journal of Hospitality Marketing & Management*, 18(8):743–764, oct.

Cruz, F. L., J. A. Troyano, F. Enríquez, F. J. Ortega, and C. G. Vallejo. 2013. 'long autonomy or long delay?' the importance of domain in opinion mining. *Expert Systems with Applications*, 40:3174–3184.

Goossen, M., H. Meeuwsen, J. Franke, ´A Maps Á Tourism, and Land. 2013. Á Destination inspiration using etourism tool. In Zheng Xiang and Iis Tussyadiah, editors, *Information and Communication Technologies in Tourism 2014*. Springer International Publishing, Cham.

Kabassi, K. 2010. Personalizing recommendations for tourists. *Telematics and Informatics*, 27(1):51–66, February.

Munar, A. M. and J. Kr. Steen Jacobsen. 2013. Trust and involvement in tourism social media and web-based travel information sources. *Scandinavian Journal of Hospitality and Tourism*, 13(1):1–19, April.

Nieves, J. and S. Haller. 2014. Building dynamic capabilities through knowl-edge resources. *Tourism Management*, 40:224–232, February.

Noti, E. 2013. Web 2.0 and the its influence in the tourism sector. *European Scientific Journal*, 9(20):115–123.

Yasvari, T. H., R. A. Ghassemi, and E. Rahrovy. 2012. Influential factors on word of mouth in service industries (the case of iran airline company). *International Journal of Learning and Development*, 2(5):227–242, October.

Ye, Q., R. Law, and B. Gu. 2009. The impact of online user reviews on ho-tel room sales. *International Journal of Hospitality Management*, 28(1):180–182, March.

Ye, Q., R. Law, B. Gu, and W. Chen. 2011. The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human Behavior*, 27(2):634–639, March.

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 127-134 recibido 27-03-2015 revisado 26-04-2015 aceptado 15-05-2015

Unsupervised Word Polarity Tagging by Exploiting Continuous Word Representations

Etiquetado no supervisado de la polaridad de las palabras utilizando representaciones continuas de palabras

Aitor García-Pablos, Montse Cuadros
Vicomtech-IK4 research centre
Mikeletegi 57, San Sebastian, Spain
{agarciap,mcuadros}@vicomtech.org

German Rigau
IXA Group
Euskal Herriko Unibertsitatea
San Sebastian, Spain
german.rigau@ehu.es

Resumen: El análisis de sentimiento es un campo del procesamiento del lenguaje natural que se encarga de determinar la polaridad (positiva, negativa, neutral) en los textos en los que se vierten opiniones. Un recurso habitual en los sistemas de análisis de sentimiento son los lexicones de polaridad. Un lexicón de polaridad es un diccionario que asigna un valor predeteminado de polaridad a una palabra. En este trabajo exploramos la posibilidad de generar de manera automática lexicones de polaridad adaptados a un dominio usando representaciones continuas de palabras, en concreto la popular herramienta Word2Vec. Primero mostramos una evaluación cualitativa de la polaridad sobre un pequeño conjunto de palabras, y después mostramos los resultados de nuestra competición en la tarea 12 del SemEval-2015 usando este método.
Palabras clave: word embeddings, polaridad de palabras, análisis de sentimiento

Abstract: Sentiment analysis is the area of Natural Language Processing that aims to determine the polarity (positive, negative, neutral) contained in an opinionated text. A usual resource employed in many of these approaches are the so-called polarity lexicons. A polarity lexicon acts as a dictionary that assigns a sentiment polarity value to words. In this work we explore the possibility of automatically generating domain adapted polarity lexicons employing continuous word representations, in particular the popular tool Word2Vec. First we show a qualitative evaluation of a small set of words, and then we show our results in the SemEval-2015 task 12 using the presented method.
Keywords: word embeddings, word polarity, sentiment analysis

1 Introduction

During the last decade the online consumer opinions have become a very valuable resource of information for companies. The huge amount of user generated content containing opinions about products, services and virtually about everything, requires automatic processing tools to handle all this data. Sentiment Analysis is the field of Natural Language Processing (NLP) that focus on determining the sentiment contained in opinion texts(Liu, 2012).

The determination of the sentiment in a text usually consists of finding subjective sentences or expressions and classifying them inside one of the possible sentiment values. Regardless if the sentiment is a continuous value or a categorical label (e.g. positive, very positive, negative, neutral, etc.), one of the key challenges in Sentiment Analysis is how to determine it for the words observed in the text under analysis.

There are many different approaches in the literature: some of them employ supervised machine learning methods to train a model that learns which words/expressions/sentences are positives and which are negatives. Other methods rely on sentiment lexicons, which are dictionaries manually developed or bootstrapped from texts using different techniques. A sentiment lexicon is an important tool for many Sentiment Analysis techniques. They can be used standalone as the only indicator for the polarity of a word, or as an additional feature for more sophisticated methods. Sentiment lexicons are domain dependent, meaning that some words or expressions may vary their po-

ISSN 1135-5948

larity from one domain to another(Choi and Cardie, 2009). If a process to create a Sentiment Lexicon is too complex or too time consuming it would be difficult to port to new domains and languages. In this paper we propose a simple yet promising approach employing continuous word representations to obtain domain-aware polarity for words.

The rest of the paper is structured as follows. Section 2 introduces previous work on deriving word polarities for Sentiment Analysis. Section 3 describes our proposed approach to obtain a polarity value for words using a continuous word embedding model, in particular by exploiting Word2Vec. Section 4 shows the results of our first experiments, in particular some qualitative analysis of adjectives in three different domains, and the results of our participation in the SemEval-2015 task 12 competition using the method exposed in this paper to calculate words polarity. Finally, section 5 contains our conclusions and future work.

2 Related Work

Sentiment analysis refers to the use of NLP techniques to identify and extract subjective information in digital texts like customer reviews about products or services. Due to the grown of the social media, and specialized websites that allow users posting comments and opinions, Sentiment Analysis has been a very prolific research area during the last decade (Pang and Lee, 2008; Zhang and Liu, 2014).

A key point in Sentiment Analysis is to determine the polarity of the sentiment implied by a certain word or expression (Taboada et al., 2011). Usually this polarity is also known as Semantic Orientation (SO). SO indicates whether a word or an expression states a positive or a negative sentiment, and can be a continuous value in a range from very positive to very negative, or a categorical value (like the common 5-star rating used to rate products).

A collection of words and their respective SO is known as sentiment lexicon. Sentiment lexicons can be constructed manually, by human experts that estimate the corresponding SO value to each word of interest. Obviously, this approach is usually too time consuming for obtaining a good coverage and difficult to maintain when the vocabulary evolves or a new language or domain must be analyzed. Therefore it is necessary to devise a method to automate the process as much as possible.

Some systems employ existing lexical resources like WordNet (Fellbaum, 1998) to bootstrap a list of positive and negative words via different methods. In Esuli, Se-bastiani and Moruzzi (2006) the authors em-ploy the glosses that accompany each Word-Net synset[1] to perform a semi-supervised synset classification. The result consists of three scores per synset: positivity, negativ-ity and objectivity. In Baccianella, Esuli and Sebastiani (2010) version 3.0 of SentiWordNet is introduced with improvements like a random walk approach in the WordNet graph to calculate the SO of the synsets. In Agerri and Garcia (2009) another system is introduced, Q-WordNet, which expands the polarities of the WordNet synsets using lexical relations like synonymy. In Guerini, Gatt, and Turchi (2013) the authors propose and compare different approaches based SentiWordNet to improve the polarity determination of the synsets.

Other authors try different bootstrapping approaches and evaluate them on WordNet of different languages (Maks et al., 2014; Vicente, Agerri, and Rigau, 2014). A problem with the approaches based on resources like WordNet is that they rely on the availability and quality of those resources for a new languages. Being a general resource, WordNet also fails to capture domain dependent semantic orientations. Likewise other approaches using common dictionaries do not take into account the shifts between domains (Ramos and Marques, 2005).

Other methods calculate the SO of the words directly from text. In Hatzivassiloglou et al., (1997) the authors model the corpus as a graph of adjectives joined by con-junctions. Then, they generate partitions on the graph based on some intuitions like that two adjectives joined by "and" will tend to share the same orientation while two adjec-tives joined by "but" will have opposite ori-entations.

On the other hand, in Turney (2002) the SO is obtained calculating the Pointwise Mutual Information (PMI) between each word and a very positive word (like "excellent") and a very negative word (like "poor") in a corpus. The result is a continuous numeric

[1] A WordNet synset in a set of synonym words that denote the same concept

value between -1 and +1.

These ideas of bootstrapping SO from a corpus have been further explored and sophisticated in more recent works (Popescu and Etzioni, 2005; Brody and Elhadad, 2010; Qiu et al., 2011)

2.1 Continuous word representations

Continuous word representations (also vector representations or word embeddings) represent each word by a n-dimensional vector. Usually, these vector encapsulates some semantic information derived from the corpus used and the process applied to derive the vector. One of the best known techniques for deriving vector representations of words and documents are Latent Semantic Indexing (Dumais et al., 1995) and Latent Semantic Analysis (Dumais, 2004).

Currently it is becoming very common in the literature to employ Neural Networks and the so-called Deep Learning to compute word embeddings (Bengio et al., 2003; Turian, Ratinov, and Bengio, 2010; Huang et al., 2012; Mikolov et al., 2013b). Word embeddings show interesting semantic properties to find related concepts, word analogies, or to use them as features to conventional machine learning algorithms (Socher et al., 2013; Tang et al., 2014; Pavlopoulos and Androutsopoulos, 2014). In Kim (2013) the word embeddings are explored to derive adjectival scales.

In this work we employ word embeddings, in particular the popular Word2Vec tool (Mikolov et al., 2013a; Mikolov, Yih, and Zweig, 2013) to obtain a polarity value for each word. As the Word2Vec model is trained on a corpus of the target domain it should be able to capture domain specific semantics, and in this case, domain specific polarities. The method is rather simple and its unsupervised nature makes it easy to apply to new languages or domains given a big enough text corpus.

3 Our approach

Our aim is to assess whether continuous word representations can be leveraged to automatically infer the polarity of the words for a given domain, just employing unlabeled text from the domain (for example, customer reviews) and a minimal set of seed words. The intuition behind this idea is based on the following assumptions:

- Continuous word representations, also called word embedding, can capture the semantic similarity between words.
- The polarity of a new given word with unknown polarity can be established by simply measuring its relative similarity with respect to two small seed sets of known positive and negative words from the domain (and its associated word embeddings).
- This fact can be exploited to arrange all the words in the vocabulary into a positive-negative axis.

Continuous word representations are mappings between entries in a vocabulary (i.e. the words) and numeric vectors of a certain size that represent the words. Depending on how these vectors are computed different linguistic or semantic facets would be captured. Albeit there are many different ways in the literature to obtain such vector representations, our experiments are based on Word2Vec[2]. Word2Vec is known to capture certain semantic patterns quite effectively, from semantically related words (e.g. obtaining "France", "Italy" and "Portugal" as similar words to the word "Spain") to more complex analogy patterns (e.g. "king" is to "man" which "queen" is to "woman").

Of course the performance and the kind of results that can be expected largely depend on the corpus used for training. We employ a domain specific dataset to obtain polarity values for a specific domain.

3.1 Datasets

To generate the Word2Vec word embeddings we have used datasets from different domains. The first dataset consists of customer reviews about restaurants. It is a 100k review subset obtained from the Yelp dataset[3]. From now on we will refer to it as Yelp-restaurants.

We also have used a second dataset of customer reviews about laptops. This dataset contains a subset of about 100k reviews from the Amazon electronic device review dataset from the Stanford Network Analysis Project

[2]We use the Word2Vec implementation contained in the Apache Spark Mllib library with its default parameters: vector size 100, skip-grams with context window 5, learning rate 0.025. `https://spark.apache.org/mllib/`

[3]`http://www.yelp.com/dataset_challenge`

(SNAP)[4], selecting reviews that contain the word "laptop".

In addition to these English datasets, we have also used the Spanish film reviews dataset from MuchoCine (Cruz et al., 2008) which contains about 4k reviews written in Spanish.

3.2 Generating the model

Word2Vec works processing plain text, taking every white-space separated token as a word. It builds a vocabulary with all the different word forms found in the training corpus. It is usual to set a minimum frequency threshold to discard those words that appear less than a certain number of times.

Before starting the process we perform a pre-processing step consisting of tokenizing, Part-of-Speech tagging and lemmatizing the datasets. For this pre-processing we use the IXA-pipes toolkit for both English and Spanish reviews [5].

Lemmatization of terms helps reducing word sparsity, because our datasets are not as big as the ones used in the literature. Part-of-Speech tagging serves to filter out non-content words (e.g. all determiners, pronouns, etc.). For other semantic tasks keeping every word might be necessary, but for our polarity-calculation task we only need content-words (nouns, verbs, adjectives and adverbs).

Once we have the dataset pre-processed, it takes only a few minutes on a commodity desktop computer to obtain the semantic word vectors using Word2Vec.

Using the Word2Vec vectors we can assign a polarity value to each word from the domain using the following simple equation:

$$\text{polarity}(w) = \text{sim}(w, POS) - \text{sim}(w, NEG) \quad (1)$$

Where POS is a set of known positive words for the domain of interest, and analogously NEG is a set of known negative words. In our experiments we have used domain independent words, like *excellent* and *horrible* respectively, or their equivalents in other languages (e.g. *excelente* and *horrible* in Spanish). We have used the cosine distance as a similarity function *sim* between the computed semantic vectors.

In that way we can obtain a continuous polarity value for every word in the domain. This word would be positive if the similarity between the target word and POS is greater than the similarity between that same word and NEG, and vice versa. The fact of obtaining a continuous value for the polarity could be an interesting property to measure the strength of the sentiment, but for now we simply convert the polarity value to a binary label: positive if the value is greater or equal to zero, and negative otherwise.

3.3 Dealing with multiword terms

Handling multiword terms is important in Sentiment Analysis systems (e.g. it is not the same to detect just "memory" than "flash memory" and/or "RAM memory", etc.). It is also important for tasks like opinion target detection or in this case, to better detect the sentiment bearing words. For example, in Spanish the word "pena" (sadness) would probably be taken as a negative word, but the expression "merecer la pena" (be worth) has the opposite polarity. Multiword terms can be also found as opinion expressions like "top notch" or "blazing fast". Finally, multiword terms arise from usual collocations of single terms, so they vary between domains.

Multiword terms are expressions that are formed by more than a single word, like idioms, typical expressions or usual word collocations. Multiword terms depend on the language and also on the topic or domain of the analysed texts. For example, in restaurant domain it is very common to find multiword terms related to recipes or names of dishes and ingredients (e.g. "black cod", "spring roll", "orange juice"). In the computer domain we have multiword terms for components like "RAM memory", "hard disk", "touch pad", "graphics card", "battery life", etc.

Handling multiword terms in advance is the only way to let Word2Vec indexing them as a vocabulary entry. Without multiword terms pre-processing it would not be possible to query the model for the polarity of expressions like "top notch", "high quality" or "high resolution", because their composing words would have been treated individually.

In order to bootstrap a list of candidate domain related multiword terms we have computed the Log-Likelihood Ratio (LLR)

[4] http://snap.stanford.edu/data/web-Amazon.html

[5] http://ixa2.si.ehu.es/ixa-pipes/

Restaurants	Laptops
happy hour	tech support
onion ring	power supply
ice cream	customer service
spring roll	operating system
live music	battery life
wine list	signal strength
filet mignon	sound quality
goat cheese	plug and play
bread sticks	numeric keypad

Table 1: Examples of multiword terms obtained for restaurants and laptop domains.

of word n-grams (with n<=3) to detect the more salient word collocations. Then we keep the top K candidates from the list ranked by the LLR measure. LLR is a common measure in the literature to estimate if two events (two words in this case) co-occur by chance or if they are truly correlated. In the case of word n-grams with n>2, the LLR is calculated taking the first word of the n-gram as the first event, and the rest of the n-1 words atomically as the second event.

With no other processing this leads to a very noisy list, in which many candidate collocations are formed by stopwords (determiners, pronouns, and other undesired words). To prevent this we first analyze the corpus to obtain the Part-of-Speech tags of the words. Then we run the calculation of the LLR for all the word n-grams in the text as before, but we keep the Part-of-Speech information of every word that composes a candidate multiword. Using the Part-of-Speech information of individual words that compose the candidate multiwords we filter out the ones that do not follow certain desired patterns (e.g. noun+noun, adj+noun, noun+prep+noun, etc.).

Table 1 shows some examples of the obtained multiword terms for the restaurant and laptop domains.

4 Experiments and results

In this section we present some preliminary results and propose ideas for further experimentation.

4.1 Qualitative tests

We have trained three Word2Vec models and we have generated some lexicons for a small set of highly frequent opinion words and expressions from each domain and corpus. More precisely we have manually annotated

Word	Pol. Score	Pol. label
delicious	0.4249	positive
tasty	0.4393	positive
inexpensive	0.3416	positive
top notch	0.2850	positive
lot of money	-0.3510	negative
slow	-0.1825	negative
arrogant	-0.2544	negative
mediocre	-0.0517	negative
fantastic	0.2408	positive
prompt	0.0766	positive
amazing	0.2659	positive
outstanding	0.1748	positive
fresh	0.3178	positive
terrible	-0.3065	negative
lousy	-0.0517	negative
poor	-0.2427	negative
yummy	0.2940	positive
pleasant	0.0112	positive
disappointing	-0.0591	negative
terrific	0.2641	positive
boring	-0.0438	negative
pathetic	0.1322	**positive***
nasty	-0.1636	negative

Table 2: Examples of polarity values obtained from the restaurants polarity lexicon.

200 adjectives taken from the dataset of each domain.

The first two models correspond to English restaurant reviews and laptop reviews. To train the Word2Vec models we have used Yelp-restaurants and Amazon-laptops datasets described in section 3.1 respectively.

Table 2 shows some results for adjectives in the restaurant domain. Table 3 shows some results for adjectives in the laptop domain. About 80% of the 200 manually annotated adjectives for restaurant domain are correctly annotated. For laptop domain the 70% of the 200 manually annotated adjectives are correctly annotated. Some examples of polarity values that seem incorrect or counter-intuitive are marked in bold with an asterisk.

In addition we have trained another model using Spanish movie reviews from MuchoCine corpus. The process to generate the Word2Vec model is the same and the only thing that must be adapted to calculate the polarity are the POS and NEG words. For example, instead of *excellent* and *horrible* we have translated them to their equivalents in Spanish, *excelente* and *horrible*.

Table 4 show some results for Spanish adjectives on the films domain. It seems that

Word	Pol. Score	Pol. label
slow	-0.0790	negative
fast	0.2007	positive
quick	0.1605	positive
crappy	-0.1753	negative
great	0.4287	positive
nice	0.2884	positive
old	-0.0387	negative
modern	0.0590	positive
glossy	0.0786	positive
top notch	0.3245	positive
incredible	0.2852	positive
funny	-0.3196	**negative***
pricey	0.0415	**positive***
bug	-0.2780	negative
break	-0.3196	negative
futuristic	0.0027	positive
trendy	0.1818	positive
high resolution	0.2122	positive
high quality	0.2069	positive
nothing but praise	-0.1318	negative
lot of problem	-0.3865	negative

Table 3: Examples of polarity values obtained from the laptops polarity lexicon.

Word	Pol. Score	Pol. label
bonito	0.1008	positive
bueno	0.6570	positive
fabuloso	0.4191	positive
increíble	0.3452	positive
inolvidable	0.3368	positive
fantástico	0.3929	positive
divertir	0.4111	positive
alucinante	-0.1228	**negative***
aburrir	-0.0501	negative
repetitivo	-0.1220	negative
absurdo	-0.0509	negative
estúpido	-0.0308	negative
brillante	0.7182	positive
genial	0.6745	positive
asombroso	-0.0625	**negative***
atractivo	0.3001	positive
enfermizo	-0.0708	negative
simple	-0.0708	negative
carismático	0.3336	positive
profundo	-0.3846	**negative***
deleznable	0.0712	**positive***

Table 4: Examples of polarity values obtained from the movies polarity lexicon (for Spanish).

the polarities are less accurate than in the English tests. Obviously, one possible reason could be the use of a much smaller corpora (4k film reviews vs. 100k for restaurants and laptops) which could be simply too small to obtain an accurate model. We leave this for future experiments.

4.2 Experiments at SemEval-2015

Finally, in order to perform a more systematic experiment to assess the validity of the polarities, we participated in the SemEval-2015 task 12 about Aspect Based Sentiment Analysis (ABSA) using this approach to calculate the polarity of words. Two training datasets were provided. The first dataset contains 254 annotated reviews about restaurants (a total of 1,315 sentences). The second dataset contains 277 annotated reviews about laptops (a total of 1,739 sentences). The annotation consists of quintuples of aspect-term, entity-attribute, polarity, and starting and ending position of the aspect-term. Since our method is unsupervised, we did not use the training datasets.

For the competition in the polarity annotation subtask similar datasets were provided, with the polarity slot empty. The subtask was about filling the polarity slot of each quintuple. For every sentence we performed the polarity annotation just counting positive and negative words (according to the Word2Vec polarity calculation) and assigning the most frequent polarity to the polarity slot of each quintuple. We took into account the negation words present in the sentence in order to reverse the polarity of the words within a certain window (one token before and two tokens after the current word). In particular, the negation words employed were: *not, neither, nothing, no, none, any, never, without, cannot.*

Table 5 shows the accuracy results for restaurant and laptop domain as they were reported in the competition. The table also shows the result of the best performing system in the competition for that subtask, the average score of all 14 participant systems and the baselines provided by the SemEval organizers. The SVM+BOW baseline is a supervised baseline that employs a Support Vector Machine based training and classification using a Bag-of-Words approach as features. The Majority baseline assigns the most frequent polarity in the training dataset. To our knowledge, best performing systems from other participants were supervised approaches trained on the provided training datasets.

Our system was performing a very basic

Polarity	Restaurants acc.	Laptops acc.
SVM+BOW	0.635	0.699
Majority	0.537	0.570
Our system	0.694	0.683
Best system	0.786	0.793
Average	0.713	0.713

Table 5: Polarity annotation accuracy results on the restaurant, laptops for slot 3.

and naive polarity annotation, relying only in the polarity values given by our trained Word2Vec model, but in our opinion the results are quite promising.

5 Conclusions and future work

In this paper we have described our approach with continuous word representations to calculate a polarity value for words for any domain and language. We use the popular Word2Vec tool to compute a vector model for the words coming from datasets of different domains. Having the appropriate corpora and seed words, this approach could provide a domain-specific lexicon with polarities for any language. As a first test, we perform a qualitative observation of the polarity values for three different domains and in two different languages. Then, we presented the results obtained in the SemEval-2015 task 12 competition, in the polarity annotation subtask, achieving quite good results despite the simple and unsupervised nature of the approach. The idea introduced in this work requires further research to assess if the method works for other domains and languages. Additional investigation is required on the effect of different parameters (dimensionality of Word2Vec vectors, number of training iterations, size of context window, size of corpora, etc.).

Acknowledgements

This work has been supported by Vicomtech-IK4.

References

Agerri, R. and A. Garcia. 2009. Q-WordNet: Extracting Polarity from WordNet Senses. *Seventh Conference on International Language Resources and Evaluation Malta Retrieved May*, 25:2010.

Baccianella, S., A. Esuli, and F. Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 0:2200–2204.

Bengio, Y., R. Ducharme, P. Vincent, and C. Janvin. 2003. A Neural Probabilistic Language Model. *The Journal of Machine Learning Research*, 3:1137–1155.

Brody, S. and N Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, (June):804–812.

Choi, Y. and C. Cardie. 2009. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 590–598. Association for Computational Linguistics.

Cruz, F. L, J. A. Troyano, F. Enriquez, and J. Ortega. 2008. Clasificación de documentos basada en la opinión: experimentos con un corpus de crıticas de cine en espanol. *Procesamiento de Lenguaje Natural*, 41.

Dumais, S, G Furnas, T Landauer, S Deerwester, S Deerwester, et al. 1995. Latent semantic indexing. In *Proceedings of the Text Retrieval Conference*.

Dumais, S. T. 2004. Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230.

Esuli, A., F. Sebastiani, and V. G. Moruzzi. 2006. SentiWord-Net : A Publicly Available Lexical Resource for Opinion Mining. *Proceedings of LREC 2006*, pages 417–422.

Fellbaum, C. 1998. *WordNet*. Wiley Online Library.

Guerini, M., L. Gatti, and M. Turchi. 2013. Sentiment Analysis : How to Derive Prior Polarities from SentiWordNet. *Emnlp*, pages 1259–1269.

Hatzivassiloglou, V., V. Hatzivassiloglou, K.R. McKeown, and K.R. McKeown. 1997. Predicting the semantic orientation

of adjectives. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages:181.

Huang, E. H, R. Socher, C. D. Manning, and A. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882.

Kim, J. 2013. Deriving adjectival scales from continuous space word representations. *Emnlp*, (October):1625–1630.

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

Maks, I., R. Izquierdo, F. Frontini, R. Agerri, A. Azpeitia, and P. Vossen. 2014. Generating Polarity Lexicons with WordNet propagation in five languages. pages 1155–1161.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*, pages 1–12, January.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. *arXiv preprint arXiv: ...*, pages 1–9, October.

Mikolov, T., W. Yih, and G. Zweig. 2013. Linguistic regularities in continuous space word representations. *Proceedings of NAACL-HLT*, pages 746–751.

Pang, B. and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Pavlopoulos, J. and I. Androutsopoulos. 2014 Aspect term extraction for sentiment analysis: New datasets, new evaluation measures and an improved unsupervised method. *Proceedings of LASMEACL*, pages 44–52.

Popescu, AM and O. Etzioni. 2005. Extracting product features and opinions from reviews. *Natural language processing and text mining*, (October):339–346.

Qiu, G., B. Liu, J. Bu, and C. Chen. 2011. Opinion word expansion and target extraction through double propa-gation. *Computational linguistics*, (July 2010).

Ramos, C. and N. C. Marques. 2005. Determining the Polarity of Words through a Common Online Dictionary.

Socher, R, A. Perelygin, J.Y. Wu, and J. Chuang. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. *newdesign.aclweb.org*.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(September 2010):267–307.

Tang, D., F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1555–1565.

Turian, J., L. Ratinov, and Y. Bengio. 2010. Word Representations: A Simple and General Method for Semi-supervised Learning. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394.

Turney, P. D. 2002. Thumbs Up or Thumbs Down? Semantic Orienta-tion Applied to Unsupervised Classifica-tion of Reviews. *Computational Linguis-tics*, (July):8.

Vicente, S., R. Agerri, and G. Rigau. 2014. Simple, Robust and (almost) Unsupervised Generation of Polarity Lexicons for Multiple Languages. *Eacl2014*.

Zhang, L. and B. Liu. 2014. Aspect and entity extraction for opinion mining. In *Data mining and knowledge discovery for big data*. Springer, pages 1–40.

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 135-142 recibido 27-03-2015 revisado 23-04-2015 aceptado 20-05-2015

Is This Tweet Satirical? A Computational Approach for Satire Detection in Spanish*

¿Es satírico este tweet? Un método automático para la identificación del lenguaje satírico en español

Francesco Barbieri
Universitat Pompeu Fabra
francesco.barbieri@upf.edu

Francesco Ronzano
Universitat Pompeu Fabra
francesco.ronzano@upf.edu

Horacio Saggion
Universitat Pompeu Fabra
horacio.saggion@upf.edu

Resumen: La lingüística computacional está cada vez mas interesada en el procesamiento del lenguaje figurado. En este artículo estudiamos la detección de noticias satíricas en español y más específicamente la detección de sátira en mensajes de Twitter. Nuestro modelo computacional se basa en la representación de cada mensaje con un conjunto de rasgos diseñados para detectar el estilo satírico y no el contenido. Nuestros experimentos muestran que nuestro modelo siempre funciona mejor que un modelo de bolsa de palabras. También mostramos que el sistema es capaz de detectar este tipo de lenguaje independientemente de la cuenta de Twitter que lo origina.
Palabras clave: Detección Automática Sátira, Lenguaje Figurado, Análisis de Sentimientos

Abstract: Computational approaches to analyze figurative language are attracting a growing interest in Computational Linguistics. In this paper, we study the characterization of Twitter messages in Spanish that advertise satirical news. We present and evaluate a system able to classify tweets as satirical or not. To this purpose, we concentrate on the tweets published by several satirical and non-satirical Twitter accounts. We model the text of each tweet by a set of linguistically motivated features that aim at capturing the style more than the content of the message. Our experiments demonstrate that our model outperforms a word-based baseline. We also demonstrate that our system models global features of satirical language by showing that it is able to detect if a tweet contains or not satirical contents independently from the account that generated the tweet.
Keywords: Satire Detection, Figurative Language, Sentiment Analysis

1 Introduction

Computational approaches to analyze figurative language are attracting a growing interest in Computational Linguistics. Characterizing the figurative meaning of a sentence or text excerpt is extremely difficult to achieve by automated approaches. Properly dealing with figurative language constitutes a core issue in several research fields, including Human-Computer Interaction and Sentiment Analysis (Turney, 2002; Pang and Lee, 2008; Pak and Paroubek, 2010). Both of them would benefit of systems able to recognize figurative language. In the case of Sentiment Analysis for example, the literal sense of a text can be different and is often the opposite of its figurative meaning.

In this research we consider the case of satire, an important form of figurative language. Satire is a phenomena where humor and irony are employed to criticize and ridicule someone or something. Even if often misunderstood, "in itself, satire is not a comic device —it is a critique — but it uses comedic devices such as parody, exaggeration, slapstick, etc. to get its laughs." (Colletta, 2009). We focus on the study of satirical news in Spanish presenting a system able to separate satirical from non-satirical news. More specifically, we concentrate on Twitter

* The research described in this paper is partially funded by the SKATER-UPF-TALN project (TIN2012-38584-C06-03).

ISSN 1135-5948

messages published by several satirical and non-satirical Twitter accounts. As satirical Twitter accounts we consider "El Mundo Today" and "El Jueves", and as non-satirical ones the popular newspapers "El Mundo" and "El País".
Two examples respectively of satirical and non-satirical tweets are:

- **Satire** (from @elmundotoday)
  Ferran Adrià confiesa que su secreto es echarle a todo vinagre de Módena caramelizado.
  (Ferran Adrià confesses that his secret is adding to each dish caramelised Modena vinegar)
- **Non-Satire** (from @El_Pais)
  La enciclopedia de Ferran Adrià se pone en marcha. Más de 80 personas trabajarán en el nuevo proyecto del chef
  (The Ferran Adrià's Encyclopedia project begins. More than 80 people are going to work on the new chef's project).

As we read in the non-satirical tweet from the newspaper "El Pais", the popular Spanish chef Ferran Adrià is going to compile an Encyclopaedia of all the Spanish traditional dishes. The satirical news makes fun of this, saying that the only secret to make a good dish is adding Modena vinegar.

In this paper, we model each tweet by linguistically motivated features, which aim at capturing not the content but the style of the message. Our experiments demonstrate that our model outperforms a word-based baseline, in detecting if a tweet is satirical or not. We also show that our system detects satire independently from the Twitter account generating the messages.

The paper is organized as follows. The second Section is an overview of the state of the art on the characterization of satire. In Section 3 we describe the tools we used to process Spanish tweets. In Section 4 we introduce the features of our model. In Section 5 we describe the experiments we carried out to evaluate our model and present their results. In Section 6 we discuss the performance of our model. In the last section we present our conclusions and our plans for future work.

2 Related Work

Satire is a form of communication where humor and irony are used to criticize someone's behavior and ridicule it. Satirical authors may be aggressive and offensive, but they "always have a deeper meaning and a social signification beyond that of the humor"(Colletta, 2009). Satire loses its significance when the audience do not understand the real intents hidden in the ironic dimension. Indeed, the key message of a satirical utterance lays in the figurative interpretation of the ironic sentence. Satire has been often studied in literature (Peter, 1956; Mann, 1973; Knight, 2004; LaMarre, Landreville, and Beam, 2009), but rarely with a computational approach. The work of Burfoot and Baldwin (2009) attempts to computationally model satire in English. They retrieved news-wires documents and satiric news articles from the web, and build a model able to recognize satirical articles. Their approach included standard text classification (Binary feature weights and Bi-normal separation feature scaling), lexical features (including profanity and slang) and semantic validity. To characterize the semantic validity of an excerpt, they identify its named entities and query the web for the conjunction of those entities, expecting that satirical conjunctions were less frequent than the ones from non-satirical news.

As said above, irony plays a key role in satire. The standard definition of irony is "saying the opposite of what you mean" (Quintilien and B., 1953). Grice (1975) believes that irony is a rhetorical figure that violates the maxim of quality, while Giora (1995) says that irony can be any form of negation with no negation markers. Wilson (2002) defined irony as echoic utterance that shows a negative aspect of someone's else opinion. Utsumi (2000) and Veale (2010a) stated that irony is a form of pretence that is violated. Since 2010 researchers designed models to detect irony automatically. Veale (2010b) proposed an algorithm for separating ironic from non-ironic similes in English, detecting common terms used in this ironic comparison. Reyes et. al (2012; 2013) proposed a model to detect irony and humour in English tweets, pointing out that Ambiguity (computed using perplexity on Google n-gram) and skip-grams which capture word sequences that contain (or skip over) arbitrary gaps, are the most informative features. Barbieri and Saggion (2014) designed an irony detection system that avoided the use of the

word-based features. However, irony has not been studied intensively in languages other than English. Few studies addressed irony detection in non-English languages like Portuguese (Carvalho et al., 2009; De Freitas et al., 2014), Dutch (Liebrecht, Kunneman, and van den Bosch, 2013) and Italian (Barbieri, Ronzano, and Saggion, 2014).

3 Data and Text Processing

We parse the textual contents of tweets in order to extract relevant linguistic and semantic features as described in this Section. We use the tool Freeling (Carreras et al., 2004) to perform sentence splitting, tokenization, POS tagging, and Word Sense Disambiguation (WSD) of tweets. WSD in Freeling relies on the Spanish Wordnet distributed by the TALP Research Centre. The Spanish Wordnet is mapped by means of the Inter-Lingual-Index to the English Wordnet 3.0 whose synset IDs are in turn characterized by sentiment scores by means of SentiWordnet[1]. In order to define the usage frequency of the words of a tweet, we use a corpus we built from a dump of the Spanish Wikipedia[2] as of May 2014.

In order to train and test our system we retrieved tweets from four twitter accounts (two satirical and two non-satirical) from June 2014 to January 2014. We gathered tweets from the satirical accounts "El Mundo Today" and "El Jueves". The non-satirical tweets were retrieved from the Twitter accounts of real newspapers: "El Mundo" and "El Pais". For each account we gathered 2,766 tweets, hence the final corpus includes 11,064 tweets. After downloading the tweets we filtered them by removing tweets that were not relevant to our study (for instance: "Buy our new issue" or "Watch the video"). We left only tweets that advertize actual news (satirical or non-satirical). We share this dataset[3] as a list of tweet IDs since per Twitter policy it is not possible to share tweets contents.

4 Our Method

This Section describes the two systems we compare with respect to their ability to classify the tweets of our dataset as satirical or not. The first system (Section 4.1) is the actual satire-detection system we present in this paper; it relies on lexical and semantic features to characterize each word of a tweet. The second system (Section 4.2) constitutes our baseline to evaluate our real approach and model tweets by relying on lemma occurrences (BOW approach). Both systems exploit Support Vector Machine[4] (Platt and others, 1999) to classify tweets as satirical or not.

4.1 Satire Detection Model

We implement a similar model to (Barbieri and Saggion, 2014) for irony detection. We characterize each tweet by seven classes of features:Frequency, Ambiguity, Part Of Speech, Synonyms, Sentiments, Characters, and Slang Words These features aim to describe intrinsic aspects of the words included in satiric tweets. The interesting propriety of the intrinsic word features is that they do not rely on words-patterns hence detect more abstract (and Twitter account-independent) traits of satire.

4.1.1 Frequency

We access the frequency corpus (see Section 3) to retrieve the frequency of each word of a tweet. Thus, we derive three types of Frequency features: *rarest word frequency* (frequency of the most rare word included in the tweet), *frequency mean* (the arithmetic average of all the frequency of the words in the tweet) and *frequency gap* (the difference between the two previous features). These features are computed including all the words of each tweet. We also calculate these features by considering only Nouns, Verbs, Adjectives, and Adverbs.

4.1.2 Ambiguity

To model the ambiguity of the words in the tweets we use the WordNet Spanish synsets associated to each word. Our hypothesis is that if a word has many meanings (synsets associated) it is more likely to be used in an ambiguous way. For each tweet we calculate the *maximum number of synsets* associated to a single word, the *mean synset number* of all the words, and the *synset gap* that is the difference between the two previous features. We determine the value of these features by including all the words of a tweet as well as

[1] http://sentiwordnet.isti.cnr.it/

[2] We thank Daniel Ferrés for his help.

[3] http://sempub.taln.upf.edu/tw/sepln2015/

[4] We relied on the LibLINEAR implementation, http://www.csie.ntu.edu.tw/ cjlin/liblinear

by considering only Nouns, Verbs, Adjectives or Adverbs.

4.1.3 Part Of Speech

The features included in the Part Of Speech (POS) group are designed to capture the syntactic structure of the tweets. The features of this group are eight and each one of them counts the number of occurrences of words characterized by a certain POS. The eight POSs considered are Verbs, Nouns, Adjectives, Adverbs, Interjections, Determiners, Pronouns, and Apposition.

4.1.4 Synonyms

We consider the frequencies (for each language its own frequency corpora, see Section 3) of the synonyms of each word in the tweet, as retrieved from WordNet. Then we compute, across all the words of the tweet: the *greatest* and the *lowest number of synonyms* with frequency higher than the one present in the tweet, the *mean number of synonyms* with frequency greater / lower than the frequency of the related word present in the tweet. We determine also the greatest / lowest number of synonyms and the mean number of synonyms of the words with frequency greater / lower than the one present in the the tweet. We compute the set of Synonyms features by considering both all words of the tweet together and only the words belonging to each one of the four POSs listed before.

4.1.5 Sentiments

The sentiment of the words in tweets is important for two reasons: to detect the *sentiment* (e.g. if tweets contain mainly positive or negative terms) and to capture unexpectedness created by a negative word in a positive context or vice versa. Relying on Sentiment lexicons (see Section 3) we computed the *number of positive / negative words*, the *sum of the intensities of the positive / negative scores of words*, the *mean of positive / negative score of words*, the *greatest positive / negative score*, the *gap between the greatest positive / negative score and the positive / negative mean*. Moreover we simply count (and measure the ratio of) the *words with polarity* not equal to zero, to detect subjectivity in the tweet. As previously done, we compute these features by considering both all the words of a tweet and only Nouns, Verbs, Adjectives, and Adverbs.

4.1.6 Characters

Even if Satirical news try to mimic the same punctuation style than non-satirical newspapers, we also wanted to capture the punctuation style of the authors, and the type of characters employed in a tweet. This is because punctuation is very important in social networks: ellipses can be sign of satire for instance, or a full stop of negative emotion. Each feature that is part of this set is the number of occurrences of a specific punctuation mark, including: ".", "!", "?", "$", "%", "&", "+", "-", "=". We also compute the numbers of Uppercase and Lowercase characters, and the length of the tweet.

4.1.7 Bad Words

Since Twitter messages often include *slang words*, we count them as they may be used often in satirical and informal messages (we compiled a list of 443 "slang words" in Spanish).

4.2 Word-Based Baseline

All the features belonging to this group are useful to model common word-patterns. These features are used to train our baseline system that classifies tweets as satirical or not, in order to carry out a comparative evaluation with our actual system that relies on the other groups of features described in the previous section. We compute the five word-based features: *lemma* (lemmas of the tweet), *bigrams* (combination of two lemmas in a sequence) and *skip 1/2/3 gram*. For each of these feature we keep the 1,000 most frequent occurrences in each training set considered (we carry out several experiments considering distinct training sets, thus considering distinct feature occurrence in each experiment, see Section 5).

5 Experiments and Results

In order to test the performances of our system we run two kind of balanced binary classificati on experiments where the two classes are "satire" and "non-satire". Our dataset includes two newspaper accounts, N1 and N2, and two satirical news accounts, S1 and S2. In the **first binary balanced classification experiment**, we train the system on a dataset composed of 80% of tweets from one of the newspaper accounts and 80% of tweets from one of the satirical accounts (5,444 tweets in total). Then we test the system on a dataset that includes 20% of the

Train	Test	Word Based	Intrinsic Features	All Features
N1 vs S1	N2 vs S2	0.622	**0.754**	0.727
N1 vs S2	N2 vs S1	0.563	0.712	**0.723**
N2 vs S1	N1 vs S2	0.592	**0.805**	0.709
N2 vs S2	N1 vs S1	0.570	**0.778**	0.737
N1-N2 vs S1-S2	N1-N2 vs S1-S2	0.735	0.814	**0.852**

Table 1: F1 of each newspaper/satirical account combination, where N1="El Pais", N2="El Mundo", S1="El Mundo Today", and S2="El Jueves". In **bold** the best results (not by chance confirmed by two-matched-samples t-test with unknown variances) between word-based and Intrinsic features.

tweets of a newspaper account that is different from the one used in the training and 20% of the tweets of a satirical account that has not been used for training. The final size of our testing set is 1,089 tweets. We run the following configurations:

- Train: N1 (80%) and S1 (80%)
  Test: N2 (20%) and S2 (20%)
- Train: N1 (80%) and S2 (80%)
  Test: N2 (20%) and S1 (20%)
- Train: N2 (80%) and S1 (80%)
  Test: N1 (20%) and S2 (20%)
- Train: N2 (80%) and S2 (80%)
  Test: N1 (20%) and S1 (20%)

With these configurations we never use tweets from the same account in both the training and testing datasets, thus we can evaluate the ability of our system to detect satire independently from the features of a specific Twitter account. As a consequence we avoid the *account modeling / recognition* effect, as the system is never trained on the same accounts where it is tested. Moreover, in order to study the learning progression in relation to the number of tweets, we divide each training set in ten folds and test the systems using 1 to 10 folds to train it. In other words, we start using a tenth of the training tweets, and progressively add a tenth of tweets more until reaching the size of the whole training set.

In the **second binary balanced classification experiment**, the training set is composed of 80% of the tweets of each account. The test includes the remaining 20% of the tweets of each account. Hence the training set includes 8,710 tweets and the test set includes 2,177 tweets.

For the two classification experiments just introduced, we test three models: the baseline (BoW, see Section 4.2), our model (Section 4.1), and the union of them. The results are reported in Table 1. The reader can note that in each experiment our system outperforms the baseline. In the first configuration the baseline achieves F1 between 0.563 to 0.622 with a mean of 0.586 in the four combinations. In the same experiment our system obtains better F1 in every configuration, with values in the range 0.712-0.805 with a mean of 0.762. In Table 2 we show the results of the three systems when only a portion of the training set is used (refer to the first column, Fold). For each fold and each system we report the mean of the four account combinations. We can see that even if the BoW slightly improves its performance when adding more tweets to the training set, our system always performs better. Additionally, our system achieves high accuracy even when using a tenth of the training tweets: with only 544 tweets the F1 of our system (the mean of the four combinations) is 0.743.

Fold	BoW	Intr.	All
1	0.526	**0.743**	0.729
2	0.530	**0.754**	0.709
3	0.556	**0.755**	0.713
4	0.559	**0.762**	0.725
5	0.565	**0.759**	0.729
6	0.579	**0.755**	0.726
7	0.571	**0.756**	0.728
8	0.576	**0.760**	0.722
9	0.576	**0.757**	0.721
10	0.586	**0.762**	0.724

Table 2: First binary classification experiments with progressive training set size (from 544 (fold 1) to 5440 (fold 10) tweets. For each experiment is reported the mean of the F1 of the four account combinations.

In the second experiment (the union of all the accounts) the baseline model improves its

performance, but our model is still better. The F1 are respectively 0.735 for the baseline model and 0.814 for our model.

In order to understand the contribution of each feature in the two models we computed the information gain on the training set of the second binary classification experiment (where tweets from all the accounts are included). The best 20 features of each model are shown in Table 3 and Table 4. The most relevant words to detect satire are slang expressions, articles and specific nouns; in the list verbs are not present and "serio" ("serious") is the only adjective. The best features of our model are the ones of the Character group, followed by Part of Speech, Frequency, Slang Words and Ambiguity. To notice, Synonyms and Sentiment groups do not contribute as much as the other groups.

IG	Lemma Feat.	Translation
0.023	manda_güevos	(slang)
0.011	en	in
0.011	de	of
0.011	que	that
0.009	saludos	greetings
0.007	por	to
0.007	sobre	on
0.007	le	him/her
0.006	el_minuto	the_minute
0.006	partido	match
0.006	uno	one
0.006	el_por	the_to
0.006	serio	serious
0.006	porque	because
0.005	méxico	mexico
0.005	mucho_gracia	thanks_a lot
0.005	te	you
0.005	tú	you
0.005	el_sobre	the_envelope

Table 3: Best lemma-based features ranked computing the information gain on the N1-N2 vs S1-S2 training set.

6 Discussion

Our system outperforms the baseline in each experiment settings. In Table 1 we can see that in the cross-account experiments (training on two accounts and testing in the other two accounts) the baseline is not able to recognize satire. Indeed lemma-based features are useful to model the vocabulary of a specific Twitter account instead of abstracting less domain/account dependent features. This is also proven by the high results obtained by the baseline in the second experiment (last row of Table 1), where both the training and test sets include tweets from the same accounts (all the accounts considered). The difference between the mean score on the first configuration (first four rows of Table 1) and the second one (last row of Table 1) is 0.15. On the other hand our model is more stable and obtains good results in each configuration. The difference in performance is lower (0.52) suggesting that our model does not depend on specific accounts. However in the first experiment our model does not obtain the same results in all the combinations: it is hard to explain exactly why, more experiments are needed. With a closer look, we can see that the best configuration is obtained when training on "El Mundo vs El Mundo Today" and testing on "El Pais vs El Jueves" (inverting these train and test datasets we obtain the worse configuration). This suggests that the combination "El Mundo vs El Mundo Today" includes diverse examples than "El Pais vs El Jueves". Indeed it is possible to detect satire with high accuracy when training on the first dataset but not in the second one. Vice versa, a system trained on "El Pais vs El Jueves" recognizes fewer ironic tweets of "El Mundo vs El Mundo Today".

Another important point is how fast our system learns to detect satirical tweets. Our system achieves very good performances if trained on a dataset made of only 544 tweets, and with less than half of the tweets available for training (Table 2, fold 4, 2176 tweets) our system obtains the best F1.

The information gain experiments give interesting information on how the systems work. The baseline model that relies on lemma-based features depends on accounts and topics. Indeed, the lemma-based feature with most information gain is a Spanish slang ("manda_güevos") as in the account "El Jueves" is often used while never present in the newspaper accounts. The other relevant features of the baseline model are nouns that depend on topics like "el partido" ("the match") and "méxico". In our model (Table 4) the most important features are the ones relative on the style of the message (Character and Frequency features). Indeed, features like the length of the message, the case and the length of the words, and number of exclamation points have high information gain. The structure of the message (Part of

IG	Feature Group	Feature Name
0.231	Characters	Length
0.087	Characters	Uppercase
0.080	Characters	First uppercase word
0.078	Characters	Lowercase
0.072	Part of Speech	Number Nouns
0.040	Characters	Longest word
0.038	Characters	Mean word length
0.037	Characters	Number of !
0.034	Part of Speech	Number Apposition
0.026	Frequency	Frequency gap (Nouns)
0.025	Frequency	Rarest Adjective
0.024	Ambiguity	Mean number Synsets
0.023	Part of Speech	Number Numbers
0.022	Frequency	Frequency Mean (Nouns)
0.021	Part of Speech	Number Pronouns
0.020	Frequency	Rarest Noun
0.017	Badwords	Badwords Ratio
0.017	Characters	Number of -
0.017	Frequency	Frequency mean (Adjectives)
0.015	Frequency	Frequency gap
0.013	Ambiguity	Max Number Synsets

Table 4: Best 20 features considering the Information Gain calculated on the N1-N2 vs S1-S2 training set (second experiment configuration where all the accounts are included).

Speech group) is also important as features like number of nouns and apposition are relevant on satire detection. The ambiguity feature plays an important role too, and the satirical tweets present words with greater polisemy (number of synsets associated) than newspaper tweets. Finally, a simple but relevant feature is the presence of "slang words", than obviously are more used in the satirical news.

We were not able to compare our approach with other satire detection systems (Burfoot and Baldwin, 2009) since approaches and dataset are very different. An important incompatibility is we only used lexical information, while Burfoot and Baldwin also included meta-information by searching the web. The other relevant difference was the dataset: they considered whole satirical and non-satirical articles, while we only use messages at most 140 characters long (tweets). Moreover, their research was on English articles.

7 Conclusion and Future Work

In this paper we present a system for the automatic detection of Spanish satirical news. We retrieve text from Twitter accounts of newspapers and satirical Twitter accounts. Our system classifies tweets by relying on linguistically motivated features that aim at capturing not the content but the style of the message. We show with cross-account experiments (experiments that never share tweets of the same Twitter accounts among training and test sets) that our system detects satire with good accuracy considerably improving performance with respect to a Bag of Words baseline. Bag of Words baselines are able to model the dictionary of specific accounts more than to detect satire.

In the future we aim to improve our model adding new features (e.g. distributional semantic) and increase our dataset by incorporating new Twitter accounts so as to perform a more extensive evaluation of our results.

References

Barbieri, F., F. Ronzano, and H. Saggion. 2014. Italian Irony Detection in Twitter: a First Approach. In *The First Italian Conference on Computational Linguistics CLiC-it 2014*, page 28.

Barbieri, F. and H. Saggion. 2014. Modelling Irony in Twitter. In *Student Research Workshop at EACL*, pages 56–64, Gothenburg, Sweden, April. ACL.

Burfoot, C. and T. Baldwin. 2009. Automatic satire detection: Are you having a laugh? In *Proceedings of the*

ACL-IJCNLP 2009 conference short papers, pages 161–164. ACL.

Carreras, Xavier, Isaac Chao, Lluis Padró, and Muntsa Padró. 2004. Freeling: An open-source suite of language analyzers. In *Language Resources and Evaluation Conference*.

Carvalho, P., L. Sarmento, M. Silva, and E. de Oliveira. 2009. Clues for detecting irony in user-generated contents: oh...!! it's so easy;-). In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 53–56. ACM.

Colletta, L. 2009. Political satire and postmodern irony in the age of stephen colbert and jon stewart. *The Journal of Popular Culture*, 42(5):856–874.

De Freitas, L. A, A. Vanin, D. Hogetop, M. Bochernitsan, and R. Vieira. 2014. Pathways for irony detection in tweets. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, pages 628–633. ACM.

Giora, R. 1995. On irony and negation. *Discourse processes*, 19(2):239–264.

Grice, H. 1975. Logic and conversation. *1975*, pages 41–58.

Knight, C. 2004. *The literature of satire*. Cambridge University Press.

LaMarre, H., K. Landreville, and M. Beam. 2009. The irony of satire political ideology and the motivation to see what you want to see in the colbert report. *The International Journal of Press/Politics*, 14(2):212–231.

Liebrecht, Christine, Florian Kunneman, and Antal van den Bosch. 2013. The perfect solution for detecting sarcasm in tweets# not. *WASSA 2013*, page 29.

Mann, J. 1973. *Chaucer and Medieval Estates Satire: The Literature of Social Classes and the General Prologue to the Canterbury Tales*. Cambridge University Press Cambridge.

Pak, A. and P. Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Language Resources and Evaluation Conference*.

Pang, Bo and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January.

Peter, J. 1956. Complaint and satire in early english literature.

Platt, John et al. 1999. Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methodssupport vector learning*, 3.

Quintilien and Harold E. B. 1953. *The Institutio Oratoria of Quintilian. With an English Translation by HE Butler*. W. Heinemann.

Reyes, A., P. Rosso, and T. Veale. 2013. A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation*, pages 1–30.

Reyes, Antonio, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12.

Turney, P. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th ACL*, pages 417–424.

Utsumi, Akira. 2000. Verbal irony as implicit display of ironic environment: Distinguishing ironic utterances from nonirony. *Journal of Pragmatics*, 32(12):1777–1806.

Veale, T. and Y. Hao. 2010a. An ironic fist in a velvet glove: Creative misrepresentation in the construction of ironic similes. *Minds and Machines*, 20(4):635–650.

Veale, T. and Y. Hao. 2010b. Detecting Ironic Intent in Creative Comparisons. In *ECAI*, volume 215, pages 765–770.

Wilson, Deirdre and Dan Sperber. 2002. Relevance theory. *Handbook of pragmatics*.

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 143-150 recibido 27-03-2015 revisado 27-04-2015 aceptado 20-05-2015

CRiSOL: Base de conocimiento de opiniones para el español*

CRiSOL: Opinion Knowledge-base for Spanish

M. Dolores Molina González, Eugenio Martínez Cámara, M. Teresa Martín Valdivia
Departamento de Informática
Universidad de Jaén, E-23071 - Jaén, España
{mdmolina, emcamara, maite}@ujaen.es

Resumen: El presente trabajo se centra en la clasificación de polaridad de comentarios de hoteles en español (COAH) y presenta un nuevo recurso léxico, CRiSOL. Este nuevo recurso toma como base la lista de palabras de opinión iSOL, a la cual incluye los valores de polaridad de los *synsets* de SentiWordNet. Debido a que SentiWordNet no es un recurso para español, se ha tenido que usar como pivote la versión española de WordNet incluida en el Repositorio Central Multilingüe (MCR). Se ha desarrollado un clasificador de la polaridad no supervisada para evaluar la validez de CRiSOL. Los resultados obtenidos con CRiSOL superan los obtenidos por los lexicones base iSOL y SentiWordNet por separado, lo cual nos anima a seguir trabajando en esta línea.
Palabras clave: Análisis de Opiniones, combinación de recursos de opinión, clasificación de la polaridad

Abstract: In this paper we focus on Spanish polarity classification in a corpus of hotel reviews (COAH) and we introduce a new lexical resource called CRiSOL. This new resource is built on the list of Spanish opinion words iSOL. CRiSOL appends to each word of iSOL the polarity value of the related synset of SentiWordNet. Due to the fact that SentiWordNet is not a Spanish linguistic resource, a Spanish version of WordNet had to be used. The Spanish version of WordNet chosen was Multilingual Central Repository (MCR). An unsupervised classifier has been developed with the aim of assessing the validity of CRiSOL. The results reached by CRiSOL are higher than the ones reached by iSOL and SentiWordNet, so that encourage us to continue this research line.
Keywords: Sentiment Analysis, opinión resources combination, polarity classification

1 Introducción

Con el paso de los años, el estudio computacional de la opinión se ha ido convirtiendo en una aplicación del Procesamiento del Lenguaje Natural (PLN) que no cesa de atraer el interés de nuevos investigadores. El persistente interés está motivado principalmente por la continuada progresión de la necesidad de conocer la orientación de las opiniones que se publican en Internet.

El Análisis de Opiniones (AO) es la tarea encargada del estudio de la opinión en el ámbito del PLN. Según Cambria y Hussain (2012), el AO se define como el conjunto de técnicas computacionales para la extracción, clasificación, comprensión y evaluación de opiniones expresadas en fuentes publicadas en Internet, comentarios en portales web y en otros contenidos generados por usuarios. En efecto, esta actualización de la definición clásica de AO de Pang y Lee (2008) sintetiza las distintas operaciones que implica el procesamiento de la opinión.

Una de las operaciones que se mencionan en la definición de Cambria y Hussain (2012) es la de clasificación. La clasificación de la polaridad tiene como fin la determinación de la categoría de opinión que se le puede asignar a un mensaje. La categoría puede ser binaria, positiva y negativa, o estar conformada por diversos niveles de intensidad de opinión. En la experimentación que aquí se presenta, se evalúa un sistema de clasificación de la polaridad binaria.

* 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.

ISSN 1135-5948

La clasificación de la polaridad se puede realizar mediante un sistema basado en aprendizaje automático (Pang, Lee, y Vaithyanathan, 2002), no supervisado (Turney, 2002) o híbrido (Prabowo y Thelwall, 2009). Independientemente de la estrategia que se emprenda, en muchos casos los sistemas requieren de la información que aportan los recursos lingüísticos de opinión. La disponibilidad de recursos de opinión no es muy abundante, y menos aún en español, de manera que la generación de nuevos recursos, o la mejora de los existentes, son aportaciones valiosas a la tarea de AO.

Las listas de palabras de opinión son recursos léxicos constituidos por palabras categorizadas normalmente en dos clases de opinión, positiva y negativa. Por otro lado, las bases de conocimiento de opinión son recursos que toman como sustento bases de conocimiento léxicas, como puede ser WordNet, y asignan a los conceptos que las conforman unos valores de polaridad.

La orientación semántica que asigna una lista de opinión a una palabra es totalmente rígida, ya que, o la palabra pertenece a una clase u a otra. Por lo tanto, cabe preguntarse ¿se mejoraría la capacidad de clasificación de una lista de palabras de opinión si se incluyen a sus palabras los valores de polaridad de una base de conocimiento de opinión?

Por otra parte, en ocasiones las bases de conocimiento de opinión insertan ruido en los sistemas de clasificación de la polaridad. Sin embargo, si se filtra la información de una base conocimiento con una lista de palabras de opinión ¿sería posible una mejora de la clasificación?

Estas dos preguntas son las que tratamos de responder en el presente artículo. Para ello, se estudiará la clasificación de la polaridad mediante la construcción de un nuevo recurso que combina la base de conocimiento de opinión SentiWordNet (SWN) (Baccianella, Esuli, y Sebastiani, 2010) y la lista de opinión iSOL (Molina-González et al., 2013). El nuevo recurso CRiSOL (*Combined Resources in iSOL*) toma como base el lexicón iSOL y para cada uno de los términos de dicho lexicón, se intenta asociar su *synset* en SWN extrayendo e integrando la tripleta de valores que mantiene dicho recurso (positivo, neutro, y negativo). El sistema resultante es evaluado sobre un corpus de opiniones en el dominio de hoteles, COAH (Molina-González et al., 2014).

El artículo se estructura de la siguiente manera: la siguiente sección resumirá algunos trabajos relacionados. La Sección 3 se circunscribirá a la descripción de los recursos lingüísticos que se han empleado. Posteriormente se detallará el sistema construido. En la Sección 5 se encontrará el análisis de los resultados obtenidos. La última sección detallará las conclusiones alcanzadas, así como las actuales líneas de trabajo.

2 Trabajos relacionados

El presente trabajo se encuadra en la investigación relacionada con la clasificación de la polaridad en un idioma distinto al inglés, basado en el desarrollo de un sistema fundamentado en el uso de una lista de palabras y en la combinación de recursos lingüísticos de opinión, con el fin de emplear la información oportuna que propicie la mejor clasificación posible.

Se remarca el hecho de que la investigación que se expone no es sobre textos en inglés, porque la mayor parte de la investigación en AO se centra en dicha lengua, como se puede comprobar en (Pang y Lee, 2008; Liu, 2012; Tsytsarau y Palpanas, 2012).

Un ejemplo de la relevancia de emplear recursos lingüísticos en AO, tanto para inglés como para español, se puede encontrar en (Brooke, Tofiloski, y Taboada, 2009). En dicho trabajo, los autores concluyen que la inclusión de la información que aportan los recursos de opinión es beneficiosa para un sistema de clasificación de la polaridad, ya sea éste supervisado o no supervisado.

En el contexto de la generación de recursos lingüísticos para AO, destacan aquellos en los que se presentan nuevos corpus de opiniones, como por ejemplo el corpus *Spanish Movie Review* (Cruz et al., 2008), la versión española de SFU corpus (Brooke, Tofiloski, y Taboada, 2009), el corpus EmotiBlog (Boldrini et al., 2009) o el corpus que se va a emplear en esta evaluación, el corpus COAH (Molina-González et al., 2014). En el ámbito de la generación de listas de palabras de opinión deben ser resaltados algunos estudios. Rangel, Sidorov, y Suárez-Guerra (2014) presentan un léxico de emociones en español compuesto por 2.036 vocablos acompañados de un valor, que representa la probabilidad de uso afectivo (PFA) del término con respecto a una de las siguientes emociones: alegría,

enfado, miedo, tristeza, sorpresa y repulsión. ML-SentiCon (Cruz et al., 2014) es un recurso que integra listas de palabras de opinión en español, inglés, gallego, vasco y catalán.

En cuanto a la combinación de métodos de clasificación y de recursos también se pueden encontrar ejemplos en la literatura relacionada con AO. Aunque no se trata de una experimentación sobre textos en español, en (Kennedy y Inkpen, 2006) se muestra como la combinación de un método supervisado con un clasificador de polaridad basado en una lista de palabras de opinión mejora los resultados de los dos clasificadores base por separado. Centrándonos exclusivamente en español, Martínez-Cámara et al. (2014), siguiendo una metodología de *stacking* (Wolpert, 1992), combinan con éxito los dos recursos que se emplean en el presente trabajo, iSOL y SentiWordNet. En dicho trabajo, se evaluaron dos sistemas no supervisados para la clasificación de las opiniones recogidas en el corpus *Spanish Movie Review*, demostrándose que la combinación de dos clasificadores basados en dos recursos de opinión mejora los resultados que obtienen por separado.

3 Recursos

En esta sección se describe, en primer lugar, el corpus de opiniones sobre hoteles. Este corpus se llama COAH y está disponible libremente. En segundo lugar se comentarán los lexicones usados para la experimentación. Se parte del lexicón independiente del dominio iSOL, al que le seguirá la descripción de SentiWordNet, recurso léxico ampliamente usado en documentos escritos en inglés, y el recurso lingüístico Repositorio Central Multilingüe (MCR). Para terminar se detallará la combinación usada de estos recursos léxicos para generar el nuevo recurso CRiSOL.

3.1 Corpus COAH

COAH[1] (*Corpus of Opinion about Andalusian Hotels*) es un corpus que contiene comentarios sobre 10 hoteles de cada una de las ocho provincias andaluzas, obteniendo un total de 1.816 opiniones escritas en español en los últimos años sobre los 80 hoteles elegidos en total. En (Molina-González et al., 2014) se detalla la generación del corpus.

Este corpus se compone de dos tipos de información. Una sobre el hotel (nombre, dirección) y otra sobre la opinión del huésped del hotel (valoración global, la identificación del usuario, la valoración de relación calidad/precio, la limpieza, etc.).

La valoración global del hotel está 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 representa una valoración positiva. Los hoteles con valor 3 se pueden catalogar como hoteles neutros, ni buenos ni malos, y por tanto, difíciles de clasificar. 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 1.

Clases	**Opiniones**
Positiva	1.020
Negativa	511
Total	**1.531**

Tabla 1: Clasificación binaria del corpus COAH

3.2 iSOL

Este lexicón fue generado a partir de una traducción automática del inglés al español del lexicón de Bing Liu generando el recurso SOL (*Spanish Opinion Lexicon*) (Martínez-Cámara et al., 2013).

La corrección manual de SOL dio lugar a iSOL. Por un lado, debido a la inflexión morfológica española, se tiene que mientras un adjetivo inglés, por lo general, no posee ni género ni número, y es representado por un solo término, al adjetivo español le corresponde hasta cuatro posibles palabras traducidas del inglés, dos para el género (masculino o femenino) y dos para el número (singular o plural). Por otra parte, siguiendo la filosofía de Bing Liu se introdujo en las listas algunas palabras mal escritas o inexistentes en el Diccionario de la Real Academia Española (DRAE) ya que aparecen con mucha frecuencia en el contenido de los medios de comunicación social, como por ejemplo "kaput", "pillín" o "coñacete". Finalmente iSOL se compone de 2.509 palabras positivas y 5.626 palabras negativas. Por ende, iSOL contiene 8.135 palabras de opinión.

[1] `http://sinai.ujaen.es/coah`

3.3 SentiWordNet

SentiWordNet (SWN) es uno de los lexicones más usados en AO y está construido sobre la base de datos léxica WordNet. Asigna a cada *synset* en WordNet tres propiedades (positivo, neutro y negativo), e indica la probabilidad de que el concepto sea positivo, neutro o negativo. Al tratarse de valores de polaridad, la suma de la tripleta debe ser 1. SWN cubre la totalidad de los *synsets* de WordNet, en concreto 117.000.

En SWN cada entrada contiene la categoría morfológica y un índice, que identifican unívocamente al *synset* en WordNet, junto a las tres propiedades que indican la probabilidad de que el *synset* sea positivo, neutro o negativo.

3.4 MCR

MCR (*Multilingual Central Repository*) (Atserias et al., 2004; Gonzalez-Agirre, Laparra, y Rigau, 2012) es un recurso lingüístico a gran escala que puede ser usado en procesos semánticos que necesitan gran cantidad de conocimiento lingüístico.

MCR integra en el mismo marco de trabajo de EuroWordNet, diversas versiones de WordNet para diferentes lenguas, inglés, español, vasco, catalán y gallego. Los *synsets* han sido construidos siguiendo el modelo propuesto por EuroWordNet, en los cuales los WordNet se enlazan mediante un índice entre lenguas (*InterLingual Index-ILI*). Por medio de este *ILI* los lenguajes están conectados, haciendo posible ir desde una palabra de un idioma a otras palabras similares traducidas a otros idiomas. Este hecho es el que nos permite enlazar SentiWordNet para el idioma español utilizando el *ILI* de MCR en el SWN en inglés.

La versión final de MCR contiene alrededor de 1,6 millones de relaciones semánticas entre los synsets, siendo la mayoría de ellos adquiridos mediante métodos automáticos. Este recurso está en continuo crecimiento siendo la última versión disponible MCR 3.0.

3.5 CRiSOL

Como se ha comentado en la Sección 3.2, iSOL es un lexicón de palabras de opinión en español compuesto por 2.509 palabras positivas y 5.626 palabras negativas.

En el presente artículo se pretende generar un nuevo recurso que combine la información de opinión de iSOL y de SentiWordNet. Para ello se intenta añadir a iSOL las puntuaciones de polaridad de los conceptos de SentiWordNet. iSOL es un recurso formado por palabras, o mejor dicho, por formas, ya que, tanto lemas como algunas de sus derivaciones constituyen iSOL. Por otro lado, SentiWordNet es un recurso conformado por conceptos en inglés, de manera que se hace obligatorio el uso de un recurso auxiliar para enlazar las formas de iSOL y SentiWordNet, el cual será el ya descrito MCR.

El proceder habitual en el uso de una base de conocimieto léxica basada en la estructura de WordNet, como es el caso de MCR y de SentiWordNet, se corresponde con el uso del identificador de los conceptos (*ILI*) para recuperar la información asociada al concepto. En este caso no se cuenta con *ILIs*, sino con formas lingüísticas de una lista de palabras de opinión. MCR asocia a cada lema un *ILI*, lo cual identifica inequivocamente uno de los posibles conceptos del lema. Tomando ese *ILI*, ya sí es posible acudir a SentiWordNet y obtener las puntuaciones de polaridad asociadas a dicho concepto. Por tanto, el proceso de generación de CRiSOL comenzó con la obtención de los lemas de las palabras de iSOL. Una vez obtenidos los lemas, el siguiente paso fue encontrar el *ILI* asociado al lema en MCR. Como es sabido, un lema puede tener asociados varios identificadores, dado que es común que un lema represente a varios conceptos. Para la primera versión de CRiSOL, se siguió como heurística el tomar como *ILI* el primero de los asociados al lema. El último paso fue el de recuperar de SentiWordNet los valores de polaridad asociados al *ILI*.

La Figura 1 representa el proceso de generación de CRiSOL, la cual se trata de una base de conocimiento compuesta por los mismos términos de iSOL, de los cuales 4.434, además de contar con la etiqueta de polaridad de iSOL, están complementados por la categoría morfológica y las puntuaciones de polaridad de SentiWordNet.

4 Experimentos y resultados

Antes de llevar a cabo los experimentos, las opiniones de hoteles del corpus COAH han sido preprocesadas con el fin de tener en cuenta los mismos criterios que se utilizaron en la generación del lexicón iSOL. Por ejemplo, las letras mayúsculas se han cambiado a minúsculas y se han eliminado las tildes.

Suponiendo que C es un comenta-

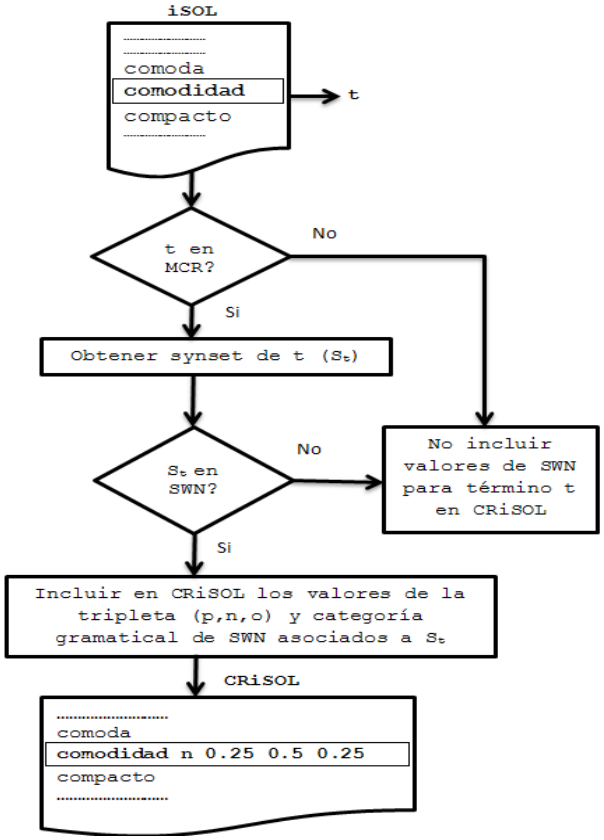

Figura 1: Proceso de generación de CRiSOL

rio de COAH, que t es un término de C, que $SWN[synset][positivo]$ y $SWN[synset][negativo]$ representan el valor de polaridad positiva y negativa de SWN, y que $iSOL^+$ e $iSOL^-$ es la lista de términos positivos y la lista de términos negativos de iSOL, se implementan cinco experimentos con el fin de responder a las preguntas hechas anteriormente.

El primer experimento (Algoritmo 1), sigue una metodología simple basada en la cuenta del número de palabras en cada opinión del corpus COAH incluidas en las listas iSOL. Así, nuestro método clasifica el comentario como positivo si el número de palabras positivas encontradas es igual o mayor que el número de palabras negativas encontradas, o como negativo en el resto de casos.

El segundo experimento (Algoritmo 2) hace uso de SentiWordNet. Como se ha comentado en la Sección 3.4, a través del *ILI* del MCR es posible ir desde una palabra de un idioma a otras palabras similares traducidas a otros idiomas. Este hecho es el que nos permite conseguir una versión de SentiWordNet para el español utilizando el *ILI* de MCR en el SWN en inglés. Usando este método,

```
Entrada: COAH, iSOL
inicio
    para cada C en COAH hacer
        Positivas ⟵ 0
        Negativas ⟵ 0
        para cada t en C hacer
            si t ∈ iSOL+ entonces
                Positivas ⟵ Positivas + 1
            fin
            si t ∈ iSOL− entonces
                Negativas ⟵ Negativas + 1
            fin
        fin
        si Positivas ≥ Negativas entonces
            C_Polaridad ⟵ positivo
        si no
            C_Polaridad ⟵ negativo
        fin
    fin
fin
```

Algoritmo 1: iSOL: Clasificación de la polaridad basada en el uso de la iSOL.

los valores de la positividad y negatividad de las palabras encontradas en SWN para cada comentario del corpus se sumarán respectivamente. Así para cada comentario se obtendrán dos resultados, uno de positividad y otro de negatividad, y si el primer valor es mayor o igual que el segundo se considerará el comentario como positivo, siendo catalogado como comentario negativo en caso contrario.

```
Entrada: COAH, MCR, SWN
inicio
    para cada C en COAH hacer
        Positivas ⟵ 0
        Negativas ⟵ 0
        para cada t en C hacer
            si t ∈ MCR entonces
                synset ⟵ MCR[t]
                si synset ∈ SWN entonces
                    Positivas ⟵ Positivas +
                    SWN[synset][positivo]
                    Negativas ⟵ Negativas +
                    SWN[synset][negativo]
                fin
            fin
        fin
        si Positivas ≥ Negativas entonces
            C_Polaridad ⟵ positivo
        si no
            C_Polaridad ⟵ negativo
        fin
    fin
fin
```

Algoritmo 2: SWN: Clasificación de la polaridad basada en el uso de SentiWordNet.

El tercer experimento (Algoritmo 3) utiliza los resultados del experimento 2 para aquellas palabras que existan en SWN. El resto de palabras se buscarán en iSOL y si están contenidas la polaridad positiva y negativa se hallará contando las palabras encontradas en cada lista iSOL. Halladas las polaridades positivas y negativas usando SWN e iSOL, se sumarán las positivas por un lado y las negativas por el otro. Se considera un comentario positivo si el valor de polaridad positiva es

mayor o igual que el valor de la polaridad negativa, siendo el comentario negativo para el resto de casos.

```
Entrada: COAH, MCR, SWN, iSOL
inicio
    para cada C en COAH hacer
        Positivas ⟵ 0
        Negativas ⟵ 0
        para cada t en C hacer
            si t ∈ MCR entonces
                synset ⟵ MCR[t]
                si synset ∈ SWN entonces
                    Positivas ⟵ Positivas +
                    SWN[synset][positivo]
                    Negativas ⟵ Negativas +
                    SWN[synset][negativo]
                si no
                    si t ∈ iSOL+ entonces
                        Positivas ⟵
                        Positivas + 1
                    fin
                    si t ∈ iSOL− entonces
                        Negativas ⟵
                        Negativas + 1
                    fin
                fin
            fin
        fin
        si Positivas ≥ Negativas entonces
            C_Polaridad ⟵ positivo
        si no
            C_Polaridad ⟵ negativo
        fin
    fin
fin
```

Algoritmo 3: SWN_iSOL: Clasificación basada en la ampliación de SWN con iSOL.

El cuarto y quinto experimento utilizan el lexicón enriquecido CRiSOL, sumando los atributos de polaridad (positividad y negatividad) de aquellas palabras que aparecen en cada comentario del corpus de hoteles.

El cuarto experimento (Algoritmo 4) clasifica el comentario como positivo si la polaridad positiva total hallada mediante la suma de las polaridades procedentes de CRiSOL en el comentario es mayor o igual que la polaridad negativa total hallada. La opinión será negativa en el resto de los casos.

El quinto experimento (Algoritmo 5) hace uso de los resultados obtenidos en el experimento cuarto. Así, a estos resultados se les añadirá el número de palabras positivas o negativas existentes en CRiSOL pero que no se han encontrado en SWN mediante la metodología explicada en la Sección 3.4 y por tanto no tienen valor en los atributos de positividad y negatividad. Las sumas resultantes se compararán y al igual que en casos anteriores, este experimento clasificará el comentario como positivo si el valor resultante positivo total es mayor o igual que el valor resultante negativo total, o como comentario negativo en el resto de casos.

En la Tabla 2 se muestran los resultados obtenidos por los cinco algoritmos expuestos anteriormente.

```
Entrada: COAH, CRiSOL
inicio
    para cada C en COAH hacer
        Positivas ⟵ 0
        Negativas ⟵ 0
        para cada t en C hacer
            si t ∈ CRiSOL[SWN] entonces
                Positivas ⟵ Positivas +
                CRiSOL[SWN][t][positivo]
                Negativas ⟵ Negativas +
                CRiSOL[SWN][t][negativo]
            fin
        fin
        si Positivas ≥ Negativas entonces
            C_Polaridad ⟵ positivo
        si no
            C_Polaridad ⟵ negativo
        fin
    fin
fin
```

Algoritmo 4: CRiSOL[SWN]: Clasificación de la polaridad basada en el uso de los valores de SWN que se encuentran en CRiSOL

```
Entrada: COAH, CRiSOL
inicio
    para cada C en COAH hacer
        Positivas ⟵ 0
        Negativas ⟵ 0
        para cada t en C hacer
            si t ∈ CRiSOL Y CRiSOL[fuente] =
            SWN entonces
                Positivas ⟵ Positivas +
                CRiSOL[SWN][t][positivo]
                Negativas ⟵ Negativas +
                CRiSOL[SWN][t][negativo]
            fin
            sinó, si t ∈ CRiSOL[fuente] = iSOL
            entonces
                si t ∈ CRiSOL[iSOL+]
                entonces
                    Positivas ⟵ Positivas + 1
                fin
                si no
                    Negativas ⟵
                    Negativas + 1
                fin
            fin
        fin
        si Positivas ≥ Negativas entonces
            C_Polaridad ⟵ positivo
        si no
            C_Polaridad ⟵ negativo
        fin
    fin
fin
```

Algoritmo 5: CRiSOL: Clasificación de la polaridad basada en el uso de CRiSOL.

5 Análisis de resultados

De los resultados recogidos en la Tabla 2 se deben destacar varios aspectos. Primeramente resaltar el buen comportamiento por separado de iSOL, y de SWN. Es remarcable también el hecho de que una lista de palabras de opinión se comporta mejor que una

Algoritmo	Macro-P	Macro-R	Macro-F1	Accuracy
iSOL	91,64 %	83,21 %	87,22 %	88,50 %
SWN_MCR	88,85 %	82,27 %	85,71 %	87,46 %
SWN_iSOL	90,52 %	85,17 %	87,76 %	89,22 %
CRiSOL[SWN]	88,19 %	83,36 %	85,70 %	87,52 %
CRiSOL	90,26 %	87,13 %	88,66 %	90,07 %

Tabla 2: Resultados obtenidos en la clasificación binaria de corpus COAH usando diferentes lexicones

base de conocimiento de opinión, aunque este comportamiento ya se intuía, porque como se indica en la Introducción, las bases de conocimiento de opinión en ocasiones insertan ruido al proceso de clasificación. Ambos recursos adolecen del mismo problema, un bajo *recall*, porque su capacidad de clasificación está limitada al vocabulario que cubren.

Para mejorar la cobertura de ambos recursos, la solución inmediata que se piensa es la de su uso conjunto (SWN_iSOL), y como se puede apreciar en la tabla de resultados, dicha combinación proporciona unos mejores resultados. La mejora se produce por aminorar el problema anterior, es decir, por mejorar la cobertura de la clasificación.

El problema de la inserción de ruido en el proceso de clasificación por parte de una base de conocimiento de opinión se puede solucionar con el filtrado de la misma con una lista de palabras de opinión, y es eso precisamente lo que se hace en CRiSOL[SWN]. Los resultados muestran que se mejora la capacidad de clasificación de la base de conocimiento, debido principalmente a un incremento de la cobertura del clasificador. Por último, si se emplea CRiSOL, es decir, el uso combinado de una lista de palabras de opinión, y una base de conocimiento de opinión filtrada por la lista, los resultados que se obtienen son globalmente mejores que el uso por separado de ambos recursos.

6 Conclusiones y Trabajo Futuro

La evaluación llevada a cabo permite contestar a las dos preguntas planteadas en la Introducción. Por un lado, filtrar una base de conocimiento de opinión, como SWN, con una lista de palabras de opinión, como iSOL (Algoritmo 4), es beneficioso para la posterior clasificación de la polaridad. Por otro lado, y lo más relevante, que la combinación de dicho filtro, con las polaridades propias de la lista de palabras (Algoritmo 5) mejora tanto la clasificación que proporcionan por separado tanto la lista de opinión, como la base de conocimiento. Por consiguiente, se puede afirmar que los resultados han demostrado la validez de CRiSOL.

La afirmación anterior es el punto de partida de una serie de nuevos trabajos. Sin considerar la revisión manual de CRiSOL, la investigación que se está llevando en este momento se centra en el estudio de la manera óptima de insertar el conocimiento de SWN en CRiSOL. Asimismo, se está evaluando la posibilidad de incluir bases de conocimiento adicionales.

En AO es de vital importancia la consideración del dominio en el que se circunscriben las opiniones. En (Molina-González et al., 2014) se expone un método de adaptación de listas de palabras de opinión a un dominio concreto, el cual será tenido en cuenta para añadir a CRiSOL información relativa a un dominio determinado.

Bibliografía

Atserias, J., L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, y P. Vossen. 2004. The meaning multilingual central repository. En *GWC 2012 6th International Global Wordnet Conference*. Brno: Masaryk University.

Baccianella, S., A. Esuli, y F. Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. En *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, páginas 2200–2204, Valletta, Malta.

Boldrini, E., A. Balahur, P. Martínez-Barco, y A. Montoyo. 2009. Emotiblog: a fine-grained model for emotion detection in non-traditional textual genres. En *WOMSA*, páginas 22–31.

Brooke, J., M. Tofiloski, y M. Taboada. 2009. Cross-linguistic sentiment analysis:

From english to spanish. En *Proceedings of the International Conference RANLP-2009*, páginas 50–54, Borovets, Bulgaria, September. ACL.

Cambria, E. y A. Hussain. 2012. *Sentic Computing*, volumen 2 de *SpringerBriefs in Cognitive Computation*. Springer Netherlands.

Cruz, F., 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 del Lenguaje Natural*, 41:73–80.

Cruz, Fermín L., Jose A. Troyano, Beatriz Pontes, y F. Javier Ortega. 2014. Ml-senticon: a multilingual, lemma-level sentiment lexicon. *Procesamiento del Lenguaje Natural*, 53:113–120.

Gonzalez-Agirre, A., E. Laparra, y G. Rigau. 2012. Multilingual central repository version 3.0. En *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Kennedy, A. y D. Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125.

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

Martínez-Cámara, E., M. T. Martín-Valdivia, M. D. Molina-González, y J. M. Perea-Ortega. 2014. Integrating spanish lexical resources by meta-classifiers for polarity classification. *Journal of Information Science*, 40(4):538–554.

Martínez-Cámara, E., M. T. Martín-Valdivia, M. D. Molina-González, y L. A. Ureña López. 2013. Bilingual experiments on an opinion comparable corpus. En *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, páginas 87–93, Atlanta, Georgia. ACL.

Molina-González, M. D., E. Martínez-Cámara, M. T. Martín-Valdivia, y L. A. Ureña-López. 2014. Cross-domain sentiment analysis using spanish opinionated words. En *Natural Language Processing and Information Systems*, volumen 8455. Springer International Publishing, páginas 214–219.

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

Pang, B. y L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Pang, Bo, Lillian Lee, y Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. En *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, volumen 10 de *EMNLP '02*, páginas 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

Prabowo, R. y M. Thelwall. 2009. Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2):143 – 157.

Rangel, I. D., G. Sidorov, y S. Suárez-Guerra. 2014. Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. *Onomázein*, 1(29):31–46.

Tsytsarau, M. y T. Palpanas. 2012. Survey on mining subjective data on the web. *Data Min. Knowl. Discov.*, 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*, ACL '02, páginas 417–424, Stroudsburg, PA, USA. ACL.

Wolpert, D. H. 1992. Stacked generalization. *Neural Networks*, 5(2):241 – 259.

Proyectos

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 153-156 recibido 26-02-2015 revisado 27-04-2015 aceptado 08-05-2015

AORESCU: análisis de opinión en redes sociales y contenidos generados por usuarios

AORESCU: Opinion Analysis in Social Networks and User-Generated Contents

José A. Troyano Jiménez
ITALICA - Universidad de Sevilla
Av. Reina Mercedes, s/n. 41012 - Sevilla
troyano@us.es

L. Alfonso Ureña López
SINAI - Universidad de Jaén
Campus Las Lagunillas s/n, 23071 - Jaén
laurena@ujaen.es

Manuel J. Maña López
LABERINTO - Universidad de Huelva
Carretera de Palos s/n, 21819 - Huelva
manuel.mana@dti.uhu.es

Fermín Cruz Mata
ITALICA - Universidad de Sevilla
Av. Reina Mercedes, s/n. 41012 - Sevilla
fcruz@us.es

Fernando Enríquez de Salamanca Ros
ITALICA - Universidad de Sevilla
Av. Reina Mercedes, s/n. 41012 - Sevilla
fenros@us.es

Resumen: El proyecto AORESCU tiene como objetivos la recopilación y el procesamiento de la información generada por los usuarios sobre una entidad con idea de obtener a partir de ella una serie de indicadores que permitan evaluar la imagen que los usuarios tienen de la misma. La información recuperada puede ser estructurada (p.e. valoraciones numéricas) y no estructurada (fundamentalmente en forma de textos en lenguaje natural). Las técnicas y herramientas utilizadas en el proyecto son adaptables a cualquier dominio. No obstante, se ha elegido el ámbito turístico como dominio de aplicación al tratarse de un sector con una importante actividad económica y para el que es fácil encontrar contenidos para analizar. El proyecto tiene cuatro partes fundamentales: la recuperación de información de distintas fuentes sobre las entidades que pertenecen al dominio de aplicación (hoteles, restaurantes, espacios naturales, monumentos,…), la definición de un modelo de datos para representar esta información, el desarrollo de herramientas de análisis de textos para procesar los comentarios de los usuarios y el desarrollo de una aplicación web que permita analizar los datos procesados.
Palabras clave: Análisis de Opiniones, Procesamiento de Lenguaje Natural, Recuperación de Información, Extracción de Opiniones

Abstract: AORESCU project main goals are focused on the retrieval and processing of information generated by users about an entity. The idea is to get insights from this information that help us to understand the perception of users about an entity. We can retrieve two types of information from web 2.0 sources: structured information (e.g. numerical rating) and unstructured (mainly in the form of texts in natural language). The techniques and tools used in the project are adaptable to any domain. We chose the tourism sector as application domain since it is a sector with an important economic activity and because it is easy to find user generated content about touristic resources. The project has four main phases: the retrieval of information from different sources about the entities (for the tourism sector, these entities are hotels, restaurants, natural spaces, monuments,...), the definition of a data model to represent this information, the development of text analysis tools to process user comments and the development of a web application to query and analyze the processed data.
Keywords: Opinion Analysis, Natural Language Processing, Information Retrieval, Opinion Extraction

ISSN 1135-5948

1 Introducción

La necesidad de conocer qué se dice de una persona, una empresa o cualquier organización no es algo nuevo. Ya en 1852 un agente de prensa polaco llamado Romeike fundó en Londres la primera empresa de *press clipping*. Este nuevo modelo de negocio, en aquellos tiempos, consistía en elaborar informes basados en recortes de prensa para personajes públicos, que estaban interesados en saber qué se decía de ellos. Desde entonces, el valor de la imagen no ha dejado de crecer, hasta el punto de que en algunos casos ya no está claro qué es más importante para una empresa: invertir en mejorar su producto o invertir en imagen.

En la actualidad ya no sólo se trata de analizar medios tradicionales como los periódicos. La irrupción del concepto web 2.0, que posibilita a cualquier usuario publicar contenidos mediante distintos canales (foros, blogs, microblogs, redes sociales…), multiplica el número de fuentes de información y plantea nuevos problemas de análisis de las mismas.

Los retos que plantea la extracción de información desde estas fuentes son complicados y desde hace unos años están siendo abordados por investigadores en varios campos. En concreto, las áreas de trabajo denominadas análisis de sentimientos y minería de opiniones están centradas en la resolución de este tipo de problemas. En ambos casos se aplican técnicas propias del procesamiento del lenguaje natural y de la minería de textos para extraer conocimiento desde textos subjetivos. El análisis de sentimientos se centra en determinar la actitud del autor de un texto con respecto a un determinado tema. La minería de opiniones, por su parte, analiza los textos a un nivel de granularidad más fino y se plantea identificar qué opina el autor del texto sobre aspectos concretos del tema sobre el que escribe (un producto, una institución, una persona, un partido político…).

2 Objetivos

El objetivo principal del proyecto es el desarrollo de un sistema para el análisis de la opinión expresada en contenidos generados por usuarios sobre una determinada entidad (empresa, producto, institución, personaje...). La idea es obtener de forma automática una serie de indicadores que resuman la imagen que los usuarios tienen sobre la entidad en función de la información generada por ellos mismos.

Los objetivos específicos del proyecto AORESCU son:

- Procesar tanto información no estructurada (por ejemplo comentarios escritos en lenguaje natural) como estructurada (como por ejemplo vínculos entre usuarios, etiquetas o información temporal).
- Aplicar técnicas de minería de textos y de procesamiento del lenguaje natural para analizar las opiniones expresadas en textos sin ningún tipo de formato.
- Desarrollar una herramienta de análisis de textos adaptable a diferentes dominios con facilidad. Para ello se separan los aspectos genéricos del análisis de contenidos, de los propios de cada dominio, quedando recogidos estos últimos en una serie de recursos lingüísticos específicos del dominio.
- Definir una taxonomía de características para cada dominio que recoja los aspectos sobre los cuales el sistema será capaz de extraer las opiniones desde los textos.
- Desarrollar una aplicación concreta en el contexto del sector turístico, que permita identificar la opinión de un colectivo de usuarios sobre productos y servicios turísticos.

3 Propuesta

Los objetivos planteados en el proyecto presentan una serie de retos relacionados con distintas líneas de investigación del área del Procesamiento del Lenguaje Natural. En concreto, el proyecto requiere fundamentalmente de la aplicación de técnicas de recuperación de información y extracción de información. Las técnicas de recuperación de información son necesarias para acceder a los contenidos publicados por usuarios sobre las entidades objeto de análisis. Las dificultades que presentan las particularidades de los textos escritos por usuarios (como la baja calidad o la concisión) se unen a las típicas dificultades de

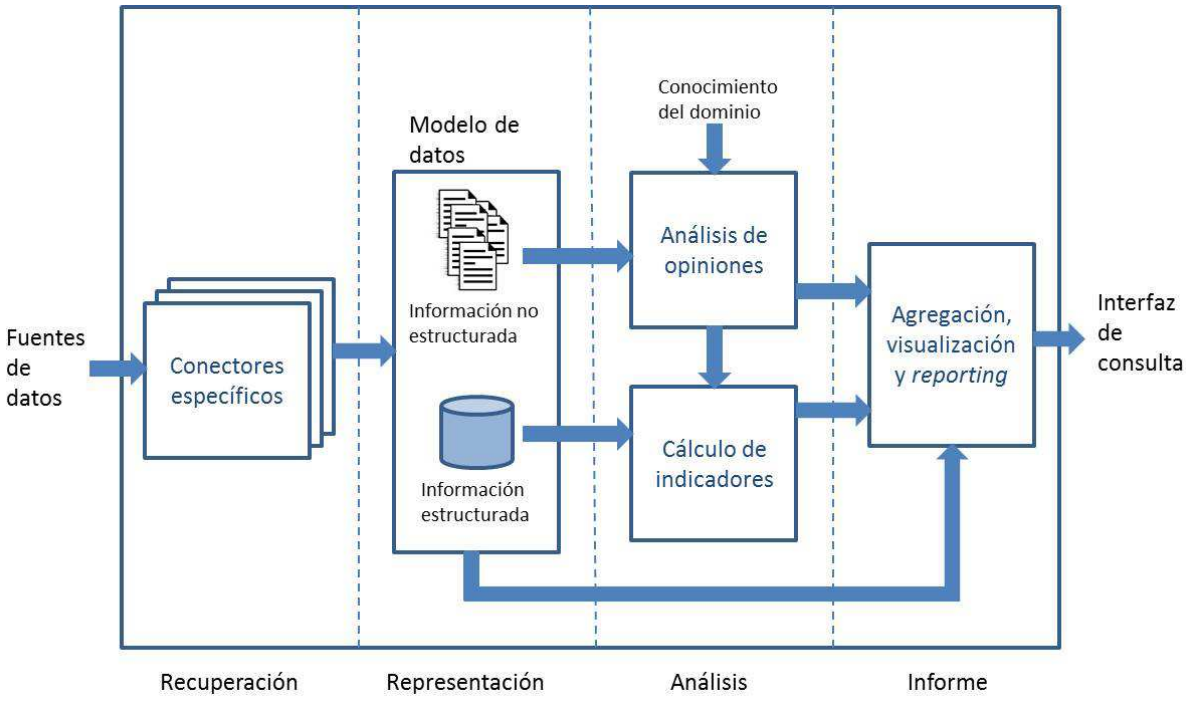

Figura 1. Arquitectura general del sistema

la recuperación de información propias de las distintas ambigüedades encontradas en cualquier texto en lenguaje natural. Las técnicas de extracción de información, por su parte, constituyen un elemento clave para el proyecto. Con ellas somos capaces de procesar los comentarios de los usuarios, que son las contribuciones más valiosas a la hora de obtener una imagen clara de la percepción que los usuarios tienen de una entidad. Si con las informaciones estructuradas es factible obtener una valoración cuantitativa que responda a la pregunta ¿cómo se valora? (por ejemplo agregando los *ratings* de distintos usuarios que han opinado), con la información no estructurada se pueden responder a preguntas más complejas, del tipo ¿por qué se valora? o ¿cuáles son los aspectos mejor y peor valorados? Evidentemente estas preguntas son mucho más interesantes pero también requieren de métodos más sofisticados para obtener la información que permita responderlas.

El proyecto AORESCU se articula en torno al desarrollo de un sistema que permita integrar las soluciones a los retos investigadores anteriormente mencionados en un entorno real. La Figura 1 muestra la arquitectura general del sistema, en ella se observan las cuatro fases principales en las que se ha dividido el proyecto: recuperación, representación, análisis e informe. La fase de recuperación contempla la implementación de conectores específicos para cada una de las fuentes analizadas (como por ejemplo redes sociales o webs de opinión). Dependiendo de las fuentes, esta recuperación se realiza con la ayuda de una API si ésta es ofrecida por la fuente o mediante *crawlers* específicos si no se dispone de dicha interfaz. La fase de representación está dedicada al desarrollo de modelos de datos que permitan almacenar la información extraída para su posterior procesamiento. La fase de análisis constituye el núcleo del proyecto y en ella se incluyen los algoritmos que permiten obtener indicadores de la percepción de los usuarios a partir del procesamiento de los contenidos publicados por ellos. Por último, la fase de informe tiene como objeto el desarrollo de una interfaz de consulta que permita explorar de forma dinámica el conocimiento generado a partir del análisis de las fuentes de información procesadas.

4 Resultados

AORESCU tiene un plazo de ejecución de 3 años de los cuales ya se han cubierto 18 meses. Nos encontramos, por tanto, en el ecuador del proyecto. En este período de tiempo se han

conseguido resultados tanto en el plano investigador como en el plano aplicado.

Aunque las técnicas aplicadas en el proyecto son aplicables a cualquier dominio, para poder llevar a la práctica estas ideas se hace necesario decidir un dominio de aplicación. Una vez que este dominio está concretado, se pueden validar experimentalmente las técnicas de adaptación a dominios específicos y también se puede concretar el desarrollo de un sistema siguiendo la arquitectura presentada en la figura 1.

Se ha elegido el sector turístico al tratarse de un sector con una importante actividad económica y para el que es fácil encontrar contenidos para analizar. A partir de ahí, se han elegido fuentes de datos para analizar distintas categorías relacionadas con el turismo (alojamiento, gastronomía, naturaleza, cultura, espectáculos y servicios). En este momento el sistema se encuentra desarrollado a un cincuenta por ciento. Se han desarrollado los conectores para las fuentes de datos, la capa de representación también está finalizada y se ha iniciado el desarrollo de los indicadores de análisis e interfaces de consulta para las categorías de alojamiento y gastronomía.

Se está siguiendo una metodología de desarrollo por fases en el desarrollo del sistema, aunque en paralelo se trabaja en el plano investigador experimentando con técnicas que permitan extraer información que pueda ser integrada en forma de indicadores en el sistema.

En el ámbito investigador, son varias las contribuciones que se han publicado en el período que lleva el proyecto. Los trabajos están relacionados fundamentalmente con las etapas de recuperación y análisis que son las que plantean los retos más interesantes desde el punto de vista investigador.

En el ámbito de la recuperación, el trabajo (Cotelo et al., 2014) presenta un método para obtener de forma automática consultas adaptativas a partir de un conjunto de *hashtags* semilla.

En el contexto del análisis de información no estructurada, se han publicado trabajos relacionados con la clasificación de documentos (Montejo-Ráez et al., 2013), con la extracción de información (Cruz et al., 2013) y con el análisis de estructuras y fenómenos lingüísticos en textos de opinión como son la negación y la especulación (Cruz Díaz et al., 2015).

También se han publicado trabajos que tienen como objeto la generación de recursos léxicos que sirvan de apoyo a tareas de análisis de opinión. En esta línea están (Molina-González et al., 2013) y (Cruz et al., 2014) en los que se presentan sendos métodos para la construcción de lexicones de palabras de opinión.

Agradecimientos

El proyecto AORESCU (P11-TIC-7684 MO) está financiado por la Consejería de Innovación, Ciencia y Empresas de la Junta de Andalucía.

Bibliografía

Cotelo, J.M., Cruz, F.L. Troyano, J.A. 2014. Dynamic topic-related tweet retrieval. *JASIST*. 65(3): 513-523

Cruz Díaz, N.P., Taboada, Mitkov, R. 2015. A Machine Learning Approach to Negation and Speculation Detection for Sentiment Analysis. *JASIST*. Pendiente de publicación.

Cruz, F.L., Troyano, J.A., Enríquez, F., Ortega, F.J., Vallejo, C.G. 2013. 'Long autonomy or long delay?' The importance of domain in opinion mining. *Expert Systems with Applications*. 40(8): 3174-3184.

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

Molina-González, M. Dolores, Martínez-Cámara, Eugenio, Martín-Valdivia, M. Teresa, Perea-Ortega, Jose M. 2013. Semantic Orientation for Polarity Classification in Spanish Reviews. *Expert Systems with Applications*. 40(18):7250-7257.

Montejo-Ráez, Arturo, Martínez-Cámara, Eugenio, Martín-Valdivia, M. Teresa, Ureña-López, L. Alfonso. 2014. A Knowledge-Based Approach for Polarity Classification in Twitter. *JASIST*. 65(2):414-425.

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 157-160 recibido 23-03-2015 revisado 27-04-2015 aceptado 14-05-2015

EXTracción de RElaciones entre Conceptos Médicos en fuentes de información heterogéneas (EXTRECM)*

EXTracción de RElaciones entre Conceptos Médicos

Arantza Díaz de Ilarraza†
Koldo Gojenola‡
UPV/EHU
†Paseo Manuel Lardizabal, 1, 20018 San Sebastián
‡Paseo Rafael Moreno Pitxitxi, 3, 48013 Bilbao
a.diazdeilarraza,koldo.gojenola@ehu.eus

Lourdes Araujo
Raquel Martínez
UNED
C/ Juan del Rosal, 16. 28040 Madrid
lurdes,raquel@lsi.uned.es

Resumen: En este proyecto se plantea la extracción de relaciones entre conceptos médicos en documentos científicos, historiales médicos e información de carácter general en Internet, en varias lenguas utilizando técnicas y herramientas de Procesamiento de Lenguaje Natural y Recuperación de Información. El proyecto se propone demostrar, mediante dos casos de uso, los beneficios de la aplicación de este tipo de tecnologías lingüísticas al dominio de la salud.
Palabras clave: identificación relaciones entre conceptos, dominio médico, minería de textos

Abstract: This project addresses extraction of medical concepts relationship in scientific documents, medical records and general information on the Internet, in several languages by using advanced Natural Language Processing and Information Retrieval techniques and tools. The project aims to show, through two use cases, the benefits of the application of language technology in the health sector.
Keywords: identification of concept relationship, medical domain, text mining

1 Descripción general

El proyecto EXTRECM (`http://ixa.si.ehu.es/extrecm`) tiene como objetivo principal proporcionar un acceso eficiente y fiable al gran volumen de información al que en este momento acceden los profesionales de la salud de manera manual o casi artesanal. Este volumen no solo corresponde a documentos científicos alojados en repositorios específicos, sino que consideramos, además, que la información contenida en la web sobre páginas especializadas y/o redes sociales puede aportar información de distinta naturaleza basada en la experiencia de los pacientes, que puede complementar a las otras fuentes. Es importante que estos profesionales puedan disponer de mecanismos que les faciliten el "acceso avanzado" a la información contenida en todos estos millones de documentos de naturaleza heterogénea. Por "acceso avanzado" entendemos un acceso que permita concentrarse en el concepto médico deseado y recuperar la información relacionada con dicho concepto médico presente en los diferentes documentos y fuentes de información heterogéneas. En este proyecto nos planteamos precisamente el reto de desarrollar y aplicar las tecnologías de tratamiento del lenguaje a diversos tipos de documentos que manejan los profesionales del área de la salud en múltiples idiomas y a escala web.

Los profesionales en el sector de la salud pública tienen que acceder a conocimiento preciso y completo para poder tomar decisiones con la mayor cantidad de información posible. Cada vez es más difícil tomar estas decisiones dado el gran volumen de datos que ha de considerarse. Este volumen dificulta encontrar manualmente relaciones que pueden ser utilizadas en la extración de conoci-

* TIN2013-46616-C2-1-R, TIN2013-46616-C2-2-R

miento. Este proyecto se centra en tres tipos de colecciones de documentos: publicaciones científicas, historiales clínicos e información de carácter general de la web, redes sociales, blogs de usuarios, etc. Las redes sociales, y particularmente Twitter, permiten introducir las valoraciones de los pacientes y allegados con respecto a una enfermedad, tratamiento, medicamento, etc. que normalmente quedan fuera de las fuentes de consulta habituales de los profesionales de la salud.

Los profesionales médicos del área de la salud están habituados a realizar consultas a algunos de estos tipos de documentación, aunque normalmente se limitan a búsquedas por palabras clave. En este proyecto se le da una nueva perspectiva a estos profesionales que están inmersos en una constante carrera para estar informados y responder adecuadamente a cualquier cambio, desarrollo o novedad. Por este motivo es importante disponer de tecnología que filtre, seleccione y organice dicha información.

Las técnicas de PLN (Procesamiento del Lenguaje Natural) y RI (Recuperación de Información) nos permitirán crear un sistema de vigilancia tecnológica, tanto de novedades científicas de interés para los expertos, como de preocupaciones e intereses sociales relacionados con la salud y reflejados en las redes. Esta vigilancia iría más allá de la información aportada por una búsqueda clásica, ya que incluiría relaciones indirectas entre los conceptos involucrados. En este momento no existe ningún sistema de consulta avanzada sobre términos/conceptos médicos en el cual el experto en medicina (doctores u otro personal sanitario) pueda formular su pregunta de forma "dirigida" por el sistema y además en inglés, español o euskera. Tampoco ningún recurso que sea capaz de realizar búsquedas en fuentes de información heterogéneas: historiales clínicos, publicaciones científicas, redes sociales e Internet en general. El proyecto EXTRECM supone una innovación que permitirá obtener las respuestas partiendo de repositorios médicos muy extensos usados por la comunidad médica internacional. Así, ayudará a eliminar posibles barreras idiomáticas y pondrá al alcance del personal sanitario toda la información existente en los mencionados repositorios.

2 Grupos involucrados

El proyecto tiene una naturaleza multidisciplinar y será abordado mediante la colaboración entre grupos de investigación expertos en tecnologías de la lengua y del área de la salud. Esta colaboración puede ayudar a la creación de sinergias entre las dos partes, con el objetivo principal de crear herramientas de procesamiento de textos médicos que mejoren la eficiencia y competitividad de los sistemas de salud y hospitalarios, posibilitando el acceso a ingentes cantidades de información.

Los grupos implicados en el proyecto son:

- Grupo IXA[1] la Universidad del País Vasco UPV/EHU. Tiene una amplia trayectoria en investigación en Procesamiento de Lenguaje Natural y lingüística computacional, y de participación en proyectos de investigación. Tiene líneas de investigación abiertas en el dominio médico.
- Grupo NLP&IR[2] de la UNED. Dispone de una amplia experiencia en Acceso Inteligente a la Información y Adquisición y Representación de Conocimiento Léxico, Gramatical y semántico. Tiene una amplia trayectoria en la realización de proyectos de investigación.
- Hospitales de Galdakao (HGA) y Basurto (HUB), integrados en el grupo de trabajo IXA pertenecientes al Servicio Público de Salud. Este grupo es pionero en el tratamiento e implantación de los historiales clínicos electrónicos, siendo un socio fundamental en este proyecto. Aportará su experiencia en el área de la detección de efectos adversos manifestados explícita o implícitamente en los historiales clínicos (caso de uso).
- Orphanet[3], entidad internacional dedicada al objetivo de contribuir a la mejora del diagnóstico, cuidado y tratamiento de los pacientes con enfermedades raras. En este proyecto esta entidad se integra en el grupo de trabajo de la UNED. Ellos nos aportan su conocimiento y necesidades en el escenario de recuperación de información para enfermedades raras (caso de uso).

[1] `http://ixa.si.ehu.es/Ixade`
[2] `http://nlp.uned.es/`
[3] `http://www.orpha.net/`

2.1 Casos de uso

Las técnicas propuestas se aplicarán a dos casos de uso específicos de interés para las instituciones médicas que colaboran en el proyecto: los hospitales de Galdakao y Basurto para el caso de identificar efectos adversos (EA) a medicamentos, y Orphanet para el caso de asociar discapacidades a enfermedades raras (EERR).

Los grupos IXA y UNED colaboran en la construcción de herramientas de PLN para el dominio de la salud en un entorno multilingüe. Esas herramientas se utilizan y se ponen a prueba en tipos de documentos heterogéneos.

IXA aplica esas herramientas de PLN en su colaboración con los hospitales de Galdakao y Basurto, para quienes identifica posibles reacciones adversas a medicamentos en informes médicos y después en los documentos recuperados de Internet. Por su parte, UNED aplica esas herramientas a distintos tipos de documentos recuperados de Internet, identificando posibles discapacidades asociadas a enfermedades raras, que interesan a Orphanet.

Además, los distintos tipos de documentos tratados en ambos grupos y las técnicas adaptadas a ellos se generalizan abordando los casos de uso de interés del proyecto de forma cruzada. A partir de una selección de medicamentos indicados por el grupo IXA, el grupo UNED aplicará las técnicas desarrolladas para la búsqueda de reacciones adversas a esa selección de medicamentos. Por su parte el grupo IXA aplicará las técnicas desarrolladas para identificar casos de EERR en los informes de que dispone, de manera que puedan relacionar reacciones adversas a medicamentos y discapacidades.

El trabajo conjunto de los dos grupos de investigadores, junto con los investigadores incluidos en los correspondientes equipos de trabajo, en su mayoría profesionales de la salud, se considera un motor generador de nuevas ideas, que podrán ser puestas en práctica en el mundo de la salud.

3 Objetivos

En la figura 1 se muestra la interrelación entre los principales objetivos del proyecto. Tomando como base el estado del arte en el área, se trata de definir los requerimientos del usuario sobre el tipo de consultas avanzadas a grandes volúmenes de información que se desean realizar en el dominio médico en los tres grandes bloques de documentación con los que nos planteamos trabajar: historiales clínicos, web sociales y artículos médicos.

Desde el punto de vista cuantitativo, es importante identificar los volúmenes de datos con los que tienen que trabajar nuestros expertos. Desde el punto de vista cualitativo, es importante estudiar la estructura de los documentos con los que se va a trabajar, formatos, tipos de información que se maneja, etc. Hemos de observar el modo de trabajo de los expertos incluidos en los grupos de trabajo de cada subproyecto ya que como resultado de esa observación conoceremos los requisitos que tienen que cumplir los sistemas que se desarrollen con el objetivo de servir de la manera más precisa a sus necesidades.

También es necesario seleccionar y preparar y, en su caso, etiquetar el conjunto de documentos de referencia para la evaluación de los resultados para los idiomas inglés, castellano y euskera. Las exigencias son diferentes para cada uno de los idiomas y el trabajo de etiquetado será más exigente para los documentos en castellano que para los escritos en euskera. Este paso va ligado al diseño e implementación de los módulos de acceso, recuperación, filtrado y organización de la información relacionados con los casos de uso.

Otra parte fundamental del proyecto es la preparación, diseño e implementación de los módulos de procesamiento del lenguaje. La idea general del proyecto es utilizar una arquitectura abierta. Son necesarios procesadores básicos para todas las lenguas del proyecto tales como tokenización, lematización y etiquetado morfosintáctico, análisis sintáctico, desambigüación semántica, y reconocimiento de entidades nombradas. Algunos de estos procesadores pueden ser módulos genéricos de procesamiento del lenguaje que han de adaptarse al dominio de la salud, pero hay otro grupo importante de herramientas que tienen que desarrollarse expresamente para cumplir los objetivos de este proyecto.

Utilizando las técnicas y herramientas mencionadas se abordará la construcción de prototipos para tratar los dos casos de uso considerados: detección de eventos adversos a medicamentos, y discapacidades asociadas a efectos adversos, en los distintos tipos de documentos considerados. Nos proponemos investigar nuevas técnicas para la detección de las conexiones más relevantes entre los con-

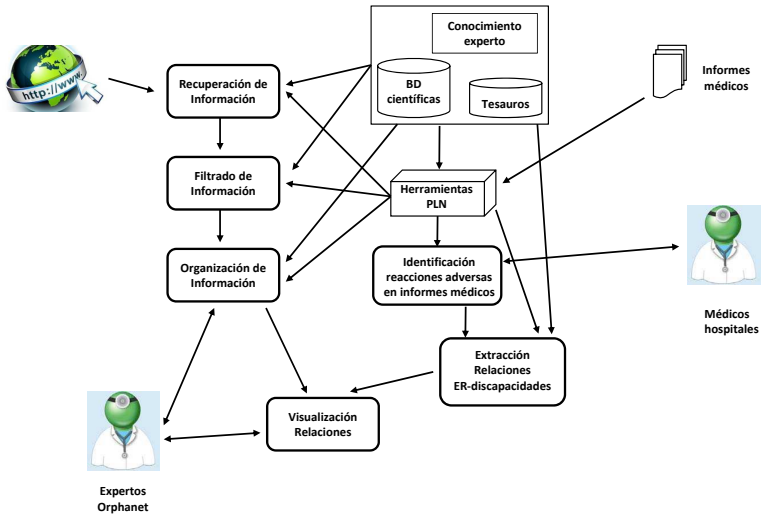

Figura 1: Relación entre los componentes del proyecto: los hospitales colaboradores y distintas fuentes de Internet proporcionan información que se procesa con técnicas de PLN desarrolladas entre ambos grupos, IXA y UNED. Orphanet y los hospitales contribuyen a la evaluación de los resultados obtenidos.

ceptos considerados. Estas conexiones pueden interpretarse como una asociación entre los conceptos implicados, que permitirá la generación de conocimiento y la confirmación de relaciones procedentes de otras fuentes. Para cumplir este objetivo se necesita por una parte identificar este tipo de conceptos mediante el uso de ontologías de dominio, reconocimiento de entidades nombradas, patrones, sinónimos, etc. y por otra parte aplicar métodos que permitan seleccionar las relaciones realmente significativas y presentarlas de forma accesible a los profesionales de la salud.

Otra fase fundamental del proyecto es la evaluación de los casos de uso con expertos. Y finalmente abordaremos la generalización de los resultados obtenidos en el proyecto. Se trata de analizar la generalidad de las técnicas desarrolladas y fomentar la interacción entre los grupos. Para comprobar esta hipótesis aplicaremos los sistemas desarrollados a los casos de uso cruzados. Así, el grupo IXA proporcionará al grupo UNED una serie de casos de reacciones adversas a medicamentos encontrados en los informes médicos. Por su parte, el grupo UNED aplicará las técnicas desarrolladas para buscar información en documentación científica y redes sociales que confirme esta hipótesis y pueda aportar detalles adicionales a esta información. Para ello las relaciones que se considerarán en este caso son medicamentos y reacciones adversas.

4 Situación Actual

El proyecto está en una fase inicial ya que lleva pocos meses activo. En relación a las colecciones que se van a utilizar en el proyecto, por una parte se está trabajando en la ampliación de los corpus anotados manualmente con efectos adversos a medicamentos, y en la mejora del asistente de anotación manual. Por otra parte se están compilando diferentes corpus con información sobre enfermedades raras y las posibles discapacidades asociadas. Esta información abarca la web y los artículos científicos.

De cara a la identificación de relaciones entre medicamentos y efectos adversos se están aplicando tanto técnicas supervisadas como no supervisadas. Además se está ampliando la cobertura del anotador morfosintáctico de conceptos médicos con nuevas abreviaturas y acrónimos. También se está trabajando en la anotación automática de discapacidades y en el filtrado de documentos relevantes.

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 161-164 recibido 23-03-2015 revisado 27-04-2015 aceptado 14-05-2015

IPHealth: Plataforma inteligente basada en *open*, *linked* y *big data* para la toma de decisiones y aprendizaje en el ámbito de la salud

IPHealth: Intelligent platform based on open, linked and big data for decision-making and learning in the field of health

Manuel de Buenaga
Diego Gachet
Dpto. de Sistemas Informáticos
Universidad Europea de Madrid
C/ Tajo s/n – 28670
Villaviciosa de Odón (Madrid)
buenaga@uem.es
diego.gachet@uem.es

Manuel J. Maña
Jacinto Mata
Dpto. de Tecnologías de la Información
Universidad de Huelva
Carretera Palos s/n – 21819
Palos de la Frontera (Huelva)
manuel.mana@dti.uhu.es
jacinto.mata@dti.uhu.es

L. Borrajo
E.L. Lorenzo
Dpto. de Informática
Universidad de Vigo
Edificio Politécnico
Campus Universitario
32004 Ourense
eva@uvigo.es
lborrajo@uvigo.es

Resumen: El proyecto IPHealth tiene como principal objetivo diseñar e implementar una plataforma con servicios que permitan un acceso integrado e inteligente a la información relacionada en el entorno biomédico. Se proponen tres escenarios de uso: (i) la asistencia a los profesionales sanitarios durante el proceso de toma de decisiones en el ámbito clínico, (ii) el acceso a información relevante sobre su estado de salud a pacientes crónicos y dependientes y (iii) el soporte a la formación basada en la evidencia de los nuevos estudiantes de medicina. Se propondrán técnicas más efectivas para diversas técnicas de PLN y extracción de información de grandes conjuntos de datos tanto provenientes de sensores como utilizando conjuntos de datos y texto de libre acceso. Se diseñará una arquitectura y un framework de aplicaciones Web que permita la integración de procesos y técnicas de minería de texto y datos e integración de información de una forma rápida, uniforme y reutilizable (mediante plugins).
Palabras clave: Minería de Texto, Minería de Datos, Integración de Información, Open Data, Big Data, Smart Sensors, Sistemas de Salud Personal

Abstract: The IPHealth project's main objective is to design and implement a platform with services that enable an integrated and intelligent access to related in the biomedical domain. We propose three usage scenarios: (i) assistance to healthcare professionals during the decision making process at clinical settings, (ii) access to relevant information about their health status and dependent chronic patients and (iii) to support evidence-based training of new medical students. Most effective techniques are proposed for reveral NLP tecniques and extraction of information from large data sets from sets of sensors and using open data. A Web application framework and an architecture that would enable integration of processes and techniques of text and data mining will be designed. Also, this architecture have to allow an integration of information in a fast, consistent and reusable (via plugins) way.
Keywords: Text Mining, Data Mining, Information Integration, Open Data, Big Data, Smart Sensors, Personal Health

1 Introducción

El proyecto IPHealth se centra en la aplicación de técnicas de integración de información para conseguir abarcar fuentes de tipo heterogéneo y de minería de textos y datos para facilitar la extracción de conocimiento útil en el contexto de la medicina personalizada (PM, Personalized Medicine). Es un proyecto financiado por el Ministerio de Economía y Competitividad, correspondientesa la convocatoria 2013 del Programa Estatal de Investigación, Desarrollo e

Innovación Orientada a los Retos de la Sociedad (TIN2013-47153-C03) y que se desarrolla desde enero de 2014 hasta diciembre de 2016.

Los numerosos avances y descubrimientos en el ámbito biomédico que han tenido lugar en la última década, tanto a nivel de tecnología como de investigación básica, han supuesto un importante revulsivo en el enfoque y la práctica clínica moderna de la medicina basada en la evidencia (EBM, Evidence-Based Medicine) (Kumar, 2011) y la medicina personalizada.

La medicina personalizada busca la identificación de terapias personalizadas que hagan seguro y efectivo el tratamiento individualizado de pacientes específicos. Una de las grandes dificultades para llevar a cabo esta práctica clínica de forma efectiva es que en la actualidad no existen sistemas flexibles de información capaces de proporcionar conocimiento preciso, actualizado e interrelacionado basado en el acceso estratificado a múltiples orígenes de datos de tipo heterogéneo (Fernald, 2011). Toda esta información, generada en estudios experimentales, ensayos clínicos y en la práctica clínica diaria, así como recientemente a través de sensores biomédicos y grandes conjuntos de datos y texto de libre acceso y entrelazado (Open y Linked Data) debería convertirse en una fuente extraordinaria de conocimiento para el avance de la medicina personalizada. Sin embargo, la medicina personalizada se enfrenta en la actualidad a grandes retos. Es necesario integrar información heterogénea dispersa en múltiples orígenes, de diferentes género, dominio, estructura y escala, donde además juega un papel muy importante la componente textual.

2 Objetivos y beneficios del proyecto

El objetivo general del proyecto es analizar, experimentar y desarrollar nuevas técnicas de minería de texto e integración de información sobre grandes cantidades de datos de fuentes de información y conocimiento heterogéneas, como elementos clave en sistemas inteligentes de acceso a la información biomédica, desarrollando una plataforma que implemente este tipo de servicios. El objetivo general se desglosa en los siguientes objetivos concretos:

O1. Desarrollo de técnicas de minería de texto bilingües (Inglés y Español) adaptadas al dominio clínico, como generación de resúmenes, recuperación de imágenes a partir de texto, recuperación de información, reconocimiento de entidades nombradas y extracción de información.

O2. Diseño y desarrollo de una arquitectura Big-Data para el almacenamiento, análisis y consulta en tiempo real de datos de trabajo. Los datos de trabajo incluyen tanto los procedentes de sensores biomédicos inalámbricos como la obtenida de fuentes de datos públicas.

O3. Diseño de una plataforma Web para soporte e integración de técnicas inteligentes que permita a los usuarios un acceso remoto y amigable a las herramientas finales.

O4. Desarrollo de herramientas inteligentes para el soporte al usuario en la toma de decisiones para el diagnóstico y tratamiento, así como para la formación.

O5. Evaluación de la efectividad y la usabilidad a través de evaluaciones sistemáticas y con usuarios.

Los objetivos del proyecto se encuentran en sintonía con el Reto en Economía y Sociedad Digital del Plan Estatal de I+D, así como con el reto “Health, Demographic Change and Wellbeing” del programa Horizonte 2020.

3 Metodología

3.1 Arquitectura Big Data

Una importante contribución del proyecto ipHealth es el uso de Tecnologías Big Data tanto para el tratamiento como para la integración de información heterogénea en el campo de la salud, información que puede provenir de diversas fuentes como por ejemplo historiales clínicos que aportan texto o bien de sensores fisiológicos que proporcionan otro tipo de información. El diseño de una arquitectura hardware y software que permita el almacenamiento masivo de datos así como su procesamiento debe necesariamente organizarse en diferentes capas que permitan la captura de datos, su filtrado, almacenamiento y posterior análisis y consulta de resultados.

La arquitectura hardware/software que se plantea en el proyecto debe corresponderse con las necesidades planteadas en cuanto a capturar, almacenar y procesar grandes conjuntos de datos, del orden de cientos de GBytes, lo que implica el resolver problemas como la escalabilidad, de manera que podamos realizar una monitorización y análisis de datos eficiente para un gran número de pacientes o bien una gran cantidad de datos como los que pueden provenir de los sensores fisiológicos al mismo

tiempo que se mantiene un coste asumible. En este sentido hemos optado por utilizar desde el punto de vista de hardware una infraestructira de cloud mediante el servicio EC2 de Amazon (Pandeya, 2012), esta primera capa permite implementar las funcionalidades necesarias de captura y almacenamiento de la información, sobre ella una segunda capa con utilizamos herramientas y tecnología Hadoop, bases de datos no relacionales y lenguajes de programación como Java y R (Prajapati 2013) que nos permitan a su vez el desarrollo de algoritmos modernos, complejos y específicos de análisis y modelado de datos que permitan la obtención de conocimiento a partir de grandes cantidades masivas de información y que a su vez faciliten el descubrimiento de tendencias y asociaciones, anomalías que puedan ser de interés para procesos clínicos, etc. (Sahoo, 2014). Tomando como ejemplo la información que podemos tener de algunos de los sensores considerados como un pulsioxímetro o un tensiómetro, se puede estudiar la saturación de oxígeno, obteniendo parámetros como el valor promedio, mínimo y máximo a lo largo de un intervalo temporal, el número de caídas en saturación de oxígeno etc. En el caso de la tensión arterial se pueden estudiar los índices de variabilidad y otros parámetros que combinados con la información proveniente de distintas fuentes como los historiales clínicos se refleje de forma fehaciente el estado de salud de una persona y permitir su caracterización a lo largo del tiempo. Una tercera capa de la arquitectura considerada se dedica a la implementación de interfaces de usuario para el acceso a resultados basadas principalmente en HTML5 ofreciendo un sistema (Gachet, 2014) con funciones integradas plenamente operativas para un usuario final.

3.2 Técnicas de Procesamiento del Lenguaje Natural

Las técnicas de Procesamiento del Lenguaje Natural (PLN) se utilizan para mejorar la precisión de los sistemas de información para la categorización de documentos, recuperación de información, sistemas de extracción de conocimiento, generación automática de resúmenes, etc.

Una de las técnicas más utilizadas es la aplicación de ontologías en tareas de PLN para la normalización de términos (que consiste en la transformación de los términos o palabras con el objetivo de reducirlos a formas canónicas que faciliten las correspondencias posteriores en el proceso de búsqueda.) y la organización semántica del contenido en uno o varios planos, que permita indexar el conocimiento poniendo el foco en el contexto de interés del usuario, facilitando por tanto la recuperación inteligente de información relevante para el mismo. Las discrepancias en la terminología utilizada, por diferentes autores, en textos relacionados pueden conllevar una reducción de la efectividad en la detección de la información común. En este sentido, pensamos que la integración de recursos como UMLS (Unified Medical Language System) (Demner-Fushman *et al.* 2010) puede resultar de gran utilidad.

Las técnicas desarrolladas en el ámbito de la minería de textos y de la integración de información se fusionarán a través de una plataforma de código libre que se irá desarrollando paralelamente. Para su elaboración se considerarán las fases clásicas de la Ingeniería del Software, haciendo énfasis en la modularidad y la escalabilidad de la plataforma. Estas fases de realizarán de manera cíclica, siguiendo modelos ágiles de desarrollo en espiral, así como la implementando las funcionalidades en prototipos incrementales que se corresponderán con cada uno de los objetivos parciales de desarrollo del proyecto.

Partiendo de los esquemas básicos de integración de:

a. areas de extracción de información, reconocimiento de entidades y detección de la negación y/o la especulación, la búsqueda de información cross-lingüe y la generación automática de resúmenes en la RI

b. tareas de creación de reglas, clasificación, generación de grafos, agrupamiento sobre datos y la recuperación de imágenes

se definirán esquemas de prototipos centrados en la interfaz, que se irán refinando hasta dar lugar a un conjunto de herramientas con capacidades de minería de texto e integración de información (Romero et al., 2014; L. Borrajo et al., 2015).

3.3 Entornos de evaluación y escenarios

Los escenarios de aplicación de los objetivos del proyecto son tan variados como los intereses de los profesionales de la materia, pero centramos nuestro interés en los siguientes:

- La asistencia a los profesionales sanitarios durante el proceso de toma de

decisiones en el ámbito clínico.
- El acceso a información relevante sobre su estado de salud a pacientes crónicos y dependientes.
- El soporte a la formación basada en la evidencia de los nuevos estudiantes de medicina.

Para dar soluciones a estos escenarios desarrollaremos herramientas que ofrezcan funcionalidades con capacidad para interrelacionar la información de casos clínicos, documentación científica y fuentes de conocimiento específicas, así como herramientas para la visualización de los datos monitorizados a través de los sensores conectados a pacientes. En los tres escenarios que se plantean se analizará toda la información clínica disponible de un paciente para ofrecer al profesional o estudiante información científica que pueda ser de su interés para ese caso concreto.

Los sistemas serán evaluados con usuarios en entornos controlados (hospitalarios, docentes, etc.), con el fin de validar la mejora en las tareas objetivo. Para ello, se definirán tareas concretas de acceso a la información, y se evaluará la capacidad de los usuarios para realizarlas de una manera más efectiva usando dichos sistemas, frente a grupos de usuarios de control que trabajan con sistemas más tradicionales (Villa et al., 2012).

Por otra parte, existen suficientes recursos para la evaluación de la efectividad de muchas tareas de PLN en el ámbito biomédico, muchos de ellos provenientes de evaluaciones competitivas. En los casos que sea posible, se utilizarán dichos recursos para evaluar la efectividad de las nuevas técnicas propuestas (Cruz et al., 2012; Crespo, Mata y Maña, 2013). Finalmente, se realizará una evaluación del sistema completo orientada a medir la usabilidad de la interfaz y la satisfacción de los usuarios.

Bibliografía

Borrajo, L., Seara, A., Iglesias, E.L. 2015 TCBR-HMM: An HMM-based text classifier with a CBR system, Applied Soft Computing, Volume 26, January2015 , Pages 463-473

Crespo, M., Mata, J., Maña, M.J. 2013. Improving image retrieval effectiveness via query expansion using MeSH hierarchical structure. *Journal of the American Medical Informatics Association*, 20(6): 1014-1020

Cruz, N.P., Maña, M.J., Mata, J., Pachón, V. 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.

Demner-Fushman D, Mork J, Shooshan S, Aronson A. 2010. UMLS content views appropriate for NLP processing of the biomedical literature vs. clinical text. Journal of Biomedical Informatics, 43 (4): 587-594.

Fernald, G.H., Capriotti, E., Daneshjou, R., Karczewski, K.J., Altman, R.B. 2011. Bioinformatics Challenges for Personalized Medicine. Bioinformatics, 27(13):1741-8

Gachet Páez, D., Aparicio, F., de Buenaga, M., y Ascanio, J. R. 2014. Big data and IoT for chronic patients monitoring, UCAmI 2014, pp. 416-423, Belfast, UK

Kumar, D. 2011. The personalised medicine. A paradigm of evidence-based medicine. Ann Ist Super Sanita., 47(1):31-40.

Pandeya S, Voorsluysa W, Niua S, et al. 2012 An autonomic cloud environment for hosting ECG data analysis services. Future Gener Comp Syst 2012;28:147–54

Prajapati, V. 2013. Big data analytics with R and Hadoop. Packt Publishing Ltd.

Romero, R., Seara, A., Iglesias, E.L. and Borrajo, L. 2014. BioClass: A Tool for Biomedical Text Classification. En *Proceedings of the 8th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2014)*, pp. 243-251, Salamanca.

Sahoo SS, Jayapandian C, Garg G, Kaffashi F, Chung S, Bozorgi A, et al. 2014 Heart beats in the cloud: distributed analysis of electrophysiological big data using cloud computing for epilepsy clinical research. J Am Med Inform Assoc 2014. Mar-Apr;21(2):263-71.

Villa, M., Aparicio, F., Maña, M.J. Buenaga, M. 2012. A Learning Support Tool with Clinical Cases Based on Concept Maps and Medical Entity Recognition. En *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*, pp. 61-70, Lisboa.

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 165-168 recibido 24-03-2015 revisado 27-04-2015 aceptado 14-05-2015

Termonet: Construcción de terminologías a partir de WordNet y corpus especializados

Termonet: Terminology construction from WordNet and technical corpora

Miguel Anxo Solla Portela
Universidade de Vigo
Grupo TALG
miguelsolla@uvigo.es

Xavier Gómez Guinovart
Universidade de Vigo
Grupo TALG
xgg@uvigo.es

Resumen: En esta presentación, mostraremos la metodología y los recursos utilizados en el desarrollo de Termonet, una herramienta para la consulta y verificación en corpus de los léxicos de especialidad incluidos en WordNet. Termonet realiza una identificación en WordNet de los synsets pertenecientes a un ámbito terminológico a partir de las relaciones léxico-semánticas establecidas entre los synsets, y valida los términos identificándolos en un corpus especializado desambiguado semánticamente. La construcción de esta herramienta forma parte de las tareas del proyecto de investigación SKATeR-UVigo, orientado al desarrollo y aplicación de recursos para el procesamiento lingüístico del gallego.
Palabras clave: WordNet, lexicografía computacional, terminología computacional

Abstract: In this presentation, we review the methodology and the resources used in the development of Termonet, a tool for checking and verifying in a corpus the specialty lexicons embedded in WordNet. This tool performs an identification of the synsets in WordNet belonging to a terminological domain from the lexical-semantic relations established among synsets, and validates the terms identifying them by means of a semantically disambiguated specialized corpus. The construction of this tool is part of the tasks of the SKATeR-UVigo research project, aimed at the development and application of resources for Galician language processing.
Keywords: WordNet, computational lexicography, computational terminology

1 Introducción

En este artículo[1] se describen la metodología y los recursos utilizados en el desarrollo de Termonet[2], una herramienta para la consulta de los léxicos de especialidad incluidos en WordNet[3] y para su verificación en corpus. La construcción de esta herramienta forma parte de los objetivos del proyecto de investigación SKATeR-UVigo, orientado al desarrollo y aplicación de recursos para el procesamiento lingüístico del gallego.

Termonet se centra en la explotación de WordNet para la construcción de terminologías, mediante la exploración de las relaciones semánticas codificadas entre los nodos conceptuales (o synsets) de la ontología léxica. Como se explica con detalle más adelante, el funcionamiento de la aplicación se basa en que los términos propios de un ámbito terminológico incluidos en WordNet se localizan en synsets relacionados con un nodo raíz mediante ciertas configuraciones de relaciones semánticas y a determinadas distancias máximas de este nodo.

Termonet ofrece la posibilidad de explorar los distintos conjuntos de synsets asociados a un synset de origen en función de las configuraciones definidas por el usuario para la selección de relaciones exploradas y para el nivel máximo de exploración de cada re-

[1]Esta investigación se realiza en el marco del proyecto *Adquisición de escenarios de conocimiento a través de la lectura de textos: Desarrollo y aplicación de recursos para el procesamiento lingüístico del gallego (SKATeR-UVigo)* financiado por el Ministerio de Economía y Competitividad, TIN2012-38584-C06-04.

[2]`http://sli.uvigo.es/termonet/termonet.php`

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

ISSN 1135-5948

lación. La misma aplicación permite verificar los resultados de la exploración en un corpus de textos especializados.

2 Recursos

Las funcionalidades de Termonet se fundamentan en dos recursos básicos: un léxico WordNet y un corpus textual lematizado y desambiguado con respecto a los sentidos de WordNet. En la implementación actual de Termonet, diseñada para su aplicación en tareas terminológicas relacionadas con la ampliación del WordNet del gallego en el ámbito de la medicina, estos dos recursos son el léxico Galnet y el *Corpus Técnico do Galego*.

Galnet, la versión gallega de WordNet, se distribuye como parte del MCR (González Agirre, Laparra, y Rigau, 2012). Esta versión de Galnet, de 2012, incluye los Basic Level Concepts[4], los ficheros lexicográficos de partes del cuerpo y de substancias, y la traducción parcial de los adjetivos. Además, contiene una primera ampliación realizada con el WN-Toolkit[5] a partir de la Wikipedia[6] y el *Dicionario CLUVI inglés-galego*[7].

A partir de esta versión inicial, se ha seguido ampliando Galnet con el WN-Toolkit a partir de los diccionarios de Apertium[8], Babelnet[9] 2.0, Wiktionary[10], Wikipedia, Geonames[11], Wikispecies[12] y los corpus SemCor inglés-gallego y CLUVI (Gómez Guinovart y O., 2014). También se ha realizado una expansión a partir del *Dicionario de sinónimos do galego*[13] (Gómez Guinovart y Solla Portela, 2014). Finalmente, se han efectuado ampliaciones en el ámbito de la fraseología (locuciones verbales) y de la terminología (medicina y economía). Todas estas expansiones se pueden consultar en la interfaz web de Galnet[14] utilizando la versión de desarrollo del recurso.

La implementación actual de Termonet usa la versión de desarrollo de Galnet 3.0.10 (2015), cuya extensión en número de synsets (Syns) y variantes léxicas (Vars) se recoge en la Tabla 1 en comparación con la de la versión de 2002 distribuida con el MCR.

	MCR		3.0.10	
	Vars	Syns	Vars	Syns
N	18949	14285	27825	20621
V	1416	612	4199	1564
Adj	6773	4415	8086	5104
Adv	0	0	471	370
Total	27138	19312	40581	27659

Tabla 1: Extensión léxica de Galnet

Por su parte, el *Corpus Técnico do Galego* (CTG)[15] es un corpus de orientación terminológica de 15 millones de palabras, formado por textos especializados del gallego contemporáneo en los ámbitos del derecho, informática, economía, ciencias ambientales, ciencias sociales y medicina. La sección del corpus de medicina del CTG (el subcorpus *Medigal*) utilizada en la implementación actual de Termonet totaliza 3.823.232 palabras. Para esta aplicación, se ha utilizado una versión del Medigal etiquetada mediante FreeLing[16] y UKB (Agirre y Soroa, 2009), empleando Galnet 3.0.10 como léxico para la desambiguación semántica del corpus.

3 Funcionalidades

3.1 Construcción de terminologías

La función principal de Termonet consiste en facilitar la extracción de variantes de WordNet relacionadas con un ámbito de especialidad. Con este fin, Termonet ofrece un formulario de consulta que permite elegir un synset de la ontología léxica y, a partir de él, realizar una extracción de los términos relacionados en función de la configuración de relaciones semánticas que se seleccionen. Aunque Termonet permite realizar la extracción desde cualquier synset de la ontología, dada su orientación terminológica, la aplicación trata de sugerir siempre las variantes nominales más próximas cuando se propone un synset no nominal.

Como se ilustra en la parte superior de la Figura 1, Termonet permite indicar el synset de origen que definirá el ámbito de la extracción terminológica, y seleccionar el conjunto de relaciones semánticas que se utilizarán para la identificación de los términos de ese

[4] http://adimen.si.ehu.es/web/BLC/
[5] http://sourceforge.net/projects/wn-toolkit/
[6] http://www.wikipedia.org
[7] http://sli.uvigo.es/dicionario/
[8] http://www.apertium.org
[9] http://www.babelnet.org
[10] http://www.wiktionary.org
[11] http://www.geonames.org
[12] http://species.wikimedia.org
[13] http://sli.uvigo.es/sinonimos/
[14] http://sli.uvigo.es/galnet/
[15] http://sli.uvigo.es/CTG/
[16] http://nlp.lsi.upc.edu/freeling/

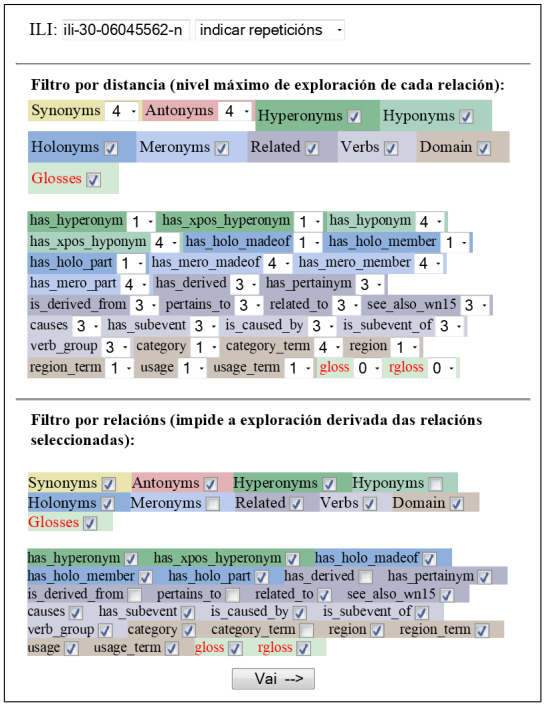

Figura 1: Consulta en Termonet.

Figura 2: Extracción de terminología.

ámbito, así como la distancia o nivel de profundidad hasta donde se desea desplegar cada tipo de relación. El concepto de distancia se refiere aquí al número de relaciones léxico-semánticas que unen dos synsets entre sí. De este modo, Termonet desplegará el árbol de relaciones desde el synset de origen a través de esa relación hasta alcanzar el nivel de profundidad determinado. Véase en la Figura 2, por ejemplo, la relación de hiponimia desplegada hasta el nivel 4 de profundidad en la terminología del ámbito de la medicina, construida a partir del synset *medical science* con los parámetros ilustrados en la Figura 1.

La aplicación cuenta también con un subformulario (parte inferior de la Figura 1) que permite restringir la extracción terminológica impidiendo la exploración derivada de las relaciones semánticas seleccionadas. Mediante este filtro, se trata de limitar la *toxicidad* de ciertas relaciónes semánticas para la selección de los términos de un ámbito de especialidad, es decir, de reducir el impacto de las relaciones que introducen synsets que se desvían del campo conceptual. Según este criterio, la hiperonimia, por ejemplo, se suele considerar una relación *tóxica*, ya que amplía la cobertura semántica inicial y tiende a introducir términos de campos conceptuales más amplios que los de partida.

Aunque la herramienta de extracción terminológica se encuentra aún en fase de desarrollo, en los experimentos se obtuvieron, con configuraciones muy simples de los parámetros de extracción, conjuntos de resultados con una congruencia mayor y cuantitativamente más significativos que la selección de variantes ligadas a un dominio de WordNet Domains[17]. Además, la extracción puede partir de cualquier synset y no está limitada a un dominio preestablecido, de modo que el procedimiento es idéntico para ámbitos conceptuales amplios, como la biología, y para campos más concisos, como la microbiología.

3.2 Verificación en corpus

Como ya se ha mencionado anteriormente, Termonet permite verificar los resultados de la extracción en un corpus textual lematizado y desambiguado con respecto a los sentidos de WordNet. En su implementación actual, permite contrastar los términos gallegos identificados en el corpus de medicina Medigal etiquetado con FreeLing y UKB.

El corpus desambiguado facilita el desarollo de estrategias de verificación con base semántica para las variantes monoléxicas procedentes de Galnet, pero no para las pluriléxicas, que no cuentan con etiquetación semánti-

[17] `http://wndomains.fbk.eu`

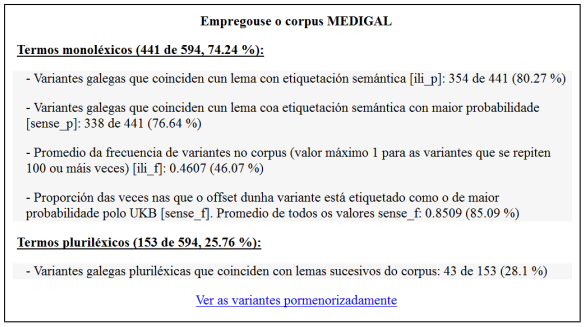

Figura 3: Verificación en corpus.

ca debido a las características de la lematización del corpus con FreeLing. Con el fin de comprobar de algún modo su presencia en el corpus, Termonet identifica las palabras léxicas de la variante en lemas sucesivos del corpus y calcula su frecuencia.

Termonet evalúa la presencia de cada término monoléxico en el corpus en base a cuatro criterios cuantificados de 0 a 1, y finalmente combina los resultados obtenidos por todos ellos en un índice general para cada criterio. Los criterios aplicados son:

1. La variante está presente (1) o no (0) como lema del corpus y con la etiqueta semántica del synset correspondiente.
2. La variante está presente como lema del corpus y con la etiqueta semántica más probable (1) o no (0) según UKB.
3. Frecuencia absoluta de la variante en el corpus, ponderando el valor máximo (1) para las variantes etiquetadas semánticamente que se repiten 100 veces o más, y el valor mínimo (0) para las variantes que no están presentes en el corpus.
4. Frecuencia con la que UKB le atribuye la mayor probabilidad a la etiqueta del synset de la variante, asignando el valor máximo (1) para la totalidad de las veces y el mínimo (0) para ninguna.

En la Figura 3 se muestran los índices globales obtenidos por la terminología construida a partir del synset *medical science* con los parámetros ilustrados en la Figura 1. A partir del análisis pormenorizado de las variantes (Figura 4), Termonet ofrece la posibilidad de comprobar sus contextos de uso en el corpus especializado (Figura 5), permitiendo así adquirir información terminológica muy valiosa sobre el uso real de los términos.

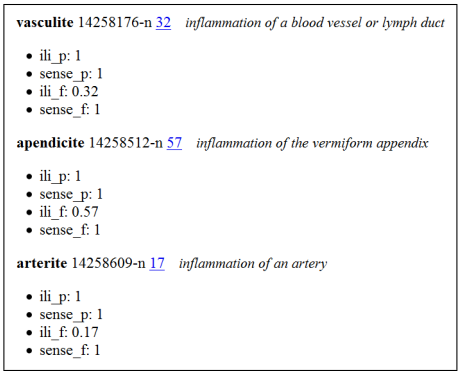

Figura 4: Evaluación de los términos.

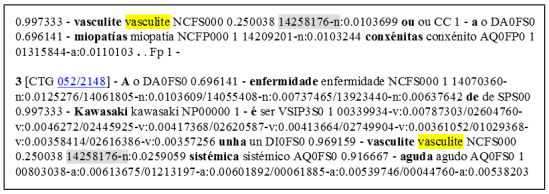

Figura 5: Término en contexto.

4 *Conclusiones*

La verificación de los términos en un corpus desambiguado permite adquirir información muy valiosa sobre su uso real y constituye una fuente de conocimiento muy relevante en la expansión de Galnet guiada por campos conceptuales. Los resultados obtenidos en la extracción, avalados por su evaluación en corpus, nos animan a continuar investigando en esta dirección y a seguir completando el WordNet del gallego desde esta perspectiva.

Bibliografía

Agirre, E. y A. Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. En *Proceedings of the 12th Conference of the European Chapter of the ACL*, págs. 33–41.

Gómez Guinovart, X. y Antoni O. 2014. Methodology and evaluation of the Galician WordNet expansion with the WN-Toolkit. *Procesamiento del Lenguaje Natural*, 53:43–50.

Gómez Guinovart, X. y M. A. Solla Portela. 2014. O dicionario de sinónimos como recurso para a expansión de WordNet. *Linguamática*, 6(2):69–74.

González Agirre, A., E. Laparra, y G. Rigau. 2012. Multilingual Central Repository version 3.0. En *6th Global WordNet Conference*.

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 169-172 recibido 24-03-2015 revisado 27-04-2015 aceptado 15-05-2015

Lexical semantics, Basque and Spanish in QTLeap: Quality Translation by Deep Language Engineering Approaches

QTLeap - Traducción de calidad mediante tratamientos profundos de ingeniería lingüística

Eneko Agirre, Iñaki Alegria, Nora Aranberri, Mikel Artetxe, Ander Barrena, António Branco (**), Arantza Díaz de Ilarraza, Koldo Gojenola, Gorka Labaka, Arantxa Otegi, and Kepa Sarasola
Ixa Taldea. Universidad del Pais Vasco /Euskal Herriko Unibertsitatea (UPV/EHU)
Manuel Lardizabal 1, -20018 Donostia
(**) Universidade de Lisboa, Departamento de Informática,Faculdade de Ciências
e.agrirre@ehu.eus

Resumen: El objetivo de este proyecto europeo FP7 es contribuir a la mejora en la calidad de la traducción automática mediante el uso de semántica, análisis sintático profundo y el uso de datos abiertos entrelazados.
Palabras clave: Traducción automática, Análisi profundo, Semántica, LOD

Abstract: The goal of this FP7 European project is to contribute for the advancement of quality machine translation by pursuing an approach that further relies on semantics, deep parsing and linked open data.
Keywords: Machine Translation, Deep language Engineering, Semantics, LOD ...

1 Summary

QTLeap project (Quality Translation by Deep Language Engineering Approaches) is a collaborative project funded by the European Commission (FP7-ICT-2013.4.1-610516) that aims to produce high-quality outbound Machine Translation (MT) using deep language engineering approaches to achieve higher quality translations (Branco and Osenova, 2014). The approach is based on deep processing and a transfer based architecture able to create hybrid systems. IXA Taldea is the partner responsible of the developments for semantics, Basque and Spanish in this project that is run by an European consortium with other seven partners: Bulgarian Academy of Sciences, Charles University in Prague, German Research Center for Artificial Intelligence, Higher Functions Lda., Humboldt University in Berlin, University of the Basque Country, University of Groningen and University of Lisbon. The project started in November 1st, 2013, and has a duration of 36 months.

The incremental advancement of research on Machine Translation has been obtained by encompassing increasingly sophisticated statistical approaches and fine grained linguistic features that add to the surface level alignment on which these approaches are ultimately anchored. The goal of this project is to contribute for the advancement of quality MT by pursuing an approach that further relies on semantics and opens the way to higher quality translation. We build on the complementarity of the two pillars of language technology, symbolic and probabilistic, and seek to advance their hybridization. We explore combinations of them that amplify their strengths and mitigate their drawbacks, along the development of three MT pilot systems that progressively seek to integrate deep language engineering approaches.

The construction of deep treebanks has progressed to be delivering now the first significant Parallel DeepBanks, where pairs of synonymous sentences from different languages are annotated with their fully-fledged grammatical representations, up to the level of their semantic representation.

The construction of Linked Open Data and other semantic resources, in turn, has progressed now to support impactful application of lexical semantic processing that handles and resolves referential and conceptual ambiguity.

ISSN 1135-5948

These cutting edge advances permit for the cross-lingual alignment supporting translation to be established at the level of deeper semantic representation. The deeper the level the less language-specific differences remain among source and target sentences and new chances of success become available for the statistically based transduction.

English is the common language for the MT systems to be built in the project, being as target or source for each one of the other 7 languages in the project: Basque, Bulgarian, Czech, Dutch, German, Portuguese and Spanish.

2 Architecture:TectoMT

All the partners use a MT transfer-based architecture, being TectoMT the architecture used for almost all the language pairs (Czech, Spanish, Portuguese, Basque, and Bulgarian). TectoMT is a highly modular, structural MT system implemented within the Treex NLP framework that allows for fast and efficient development of machine translation by exploiting a wide range of software modules already integrated in TectoMT, such as tools for sentence segmentation, tokenization, morphological analysis, POS tagging, shallow and deep syn- tax parsing, named entity recognition, anaphora resolution, tree-to-tree translation, natural language generation, word-level alignment of parallel corpora, and transfer based on deep syntactic (tectogrammatical) layer (Zeman et al., 2014). The tectogrammatical layer is based on the Prague Dependency Treebank. Figure 1 shows the architecture of the TectoMT system.

3 Deep Language Analysis for Spanish and Basque

Deep Language Analysis for Spanish and Basque will be implemented in our group via Ixa-pipes (Agerri, Bermudez, and Rigau, 2014). IXA-pipes is a modular set of Natural Language Processing tools (or pipes) which provide easy access to NLP technology for several languages. It offers robust and efficient linguistic annotation to both researchers and non-NLP experts with the aim of lowering the barriers of using NLP technology either for research purposes or for small industrial developers and SMEs. The ixa-pipes tools can be used or exploit its modularity to pick and change different components (Morpholgy, POS tagger, Chunker, Coreference, Named Entities Recognizer...). The tools are developed by the IXA NLP Group of the University of the Basque Country.

4 Semantic ways for MT improving

The overall goal of the work package on lexical semantics lead by Ixa Taldea is to enhance MT with advanced crosslingual methods for the resolution of referential and lexical ambiguity by pursuing the following objectives:

- to provide for the assembling and curation of the data sets and processing tools available to support the resolution of referential and lexical ambiguity,
- to leverage the resolution of referential and lexical ambiguity by means of advanced crosslingual named entity and word sense resolution methods,
- to proceed with the intrinsic evaluation of the solutions found in the previous task,
- to contribute for high quality machine translation by using semantic linking and resolving.

5 Real user scenario

QTLeap project successfully achieved the first year milestone last November related with the identification of a real user scenario that will show the suitability to use machine translation ('Leveraging practical multilingual helpdesk services with machine translation'). Namely in an IT helpdesk service provided by HF, Higher Functions - Intelligent Information Systems Ltd, a Portuguese SME, which is a partner of the consortium.

With this service, if a user of an IT device or service needs to solve a problem, he/she can ask a question for help through a chat channel. If there is already a similar question in the database, the associated response is immediately delivered to the user. This process helps to minimize the human operation, which becomes needed only in those cases when there is no similar question-answer pair already available in the database. The application of the machine translation system extends this support service by allowing the use of the seven languages in the project to ask a question to the helpdesk service. The

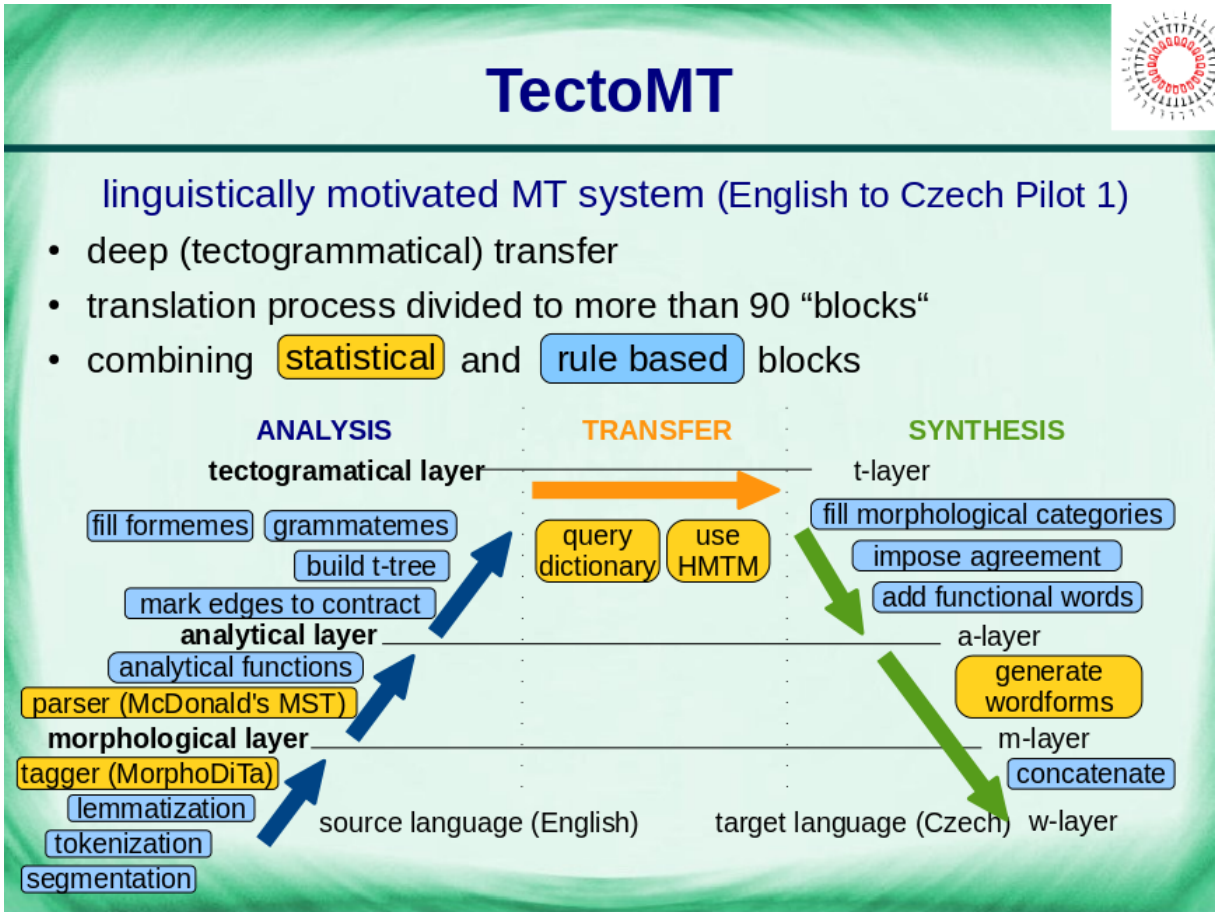

Figure 1: TectoMT architecture (Popel, 2014)

users can ask a question in their own languages, which is translated into the language of the questions and answers that are stored in the database. With the retrieved answer also translated back to the language of the original question, the helpdesk thus returns the response in the language of the user. Examples of the questions the user could ask are the following:

- Tengo un iPad. ¿Cómo abro una nueva pestana en Safari?
- ¿Cómo se qué versión de Photoshop tengo?
- En YouTube, ¿cómo puedo buscar vídeos en HD?
- ¿Qué es una IP estática?
- He instalado Linux. ¿Cómo puedo listar los archivos en una carpeta?

- Ich habe ein iPad. Wie kann ich in Safari einen neuen Tab öffnen?
- Wie kann ich auf YouTube nach Videos in HD suchen?
- Wie kann ich erfahren, welche Photoshop Version ich habe?
- Was ist ein statisches IP?
- Ich habe Linux installiert. Wie kann ich die Liste der Dateien in einem Ordner erstellen?

- iPad bat dut. Nola ireki dezaket Safari fitxa berri batean?
- Nola jakin dezaket Photoshop-en zein bertsio dudan?
- YouTube-n, nola bila ditzaket HD bideoak?
- Zer da IP estatiko bat?
- Linux instalatu dut. Nola zerrenda ditzaket fitxategiak karpeta batean?

Figure 2 shows an example of the work done by the initial translators for Spanish. Although the answer given to the user is not being to be perfect, it uses to be suitable enough to resolve the question. The evaluation of this helpdesk service expanded with

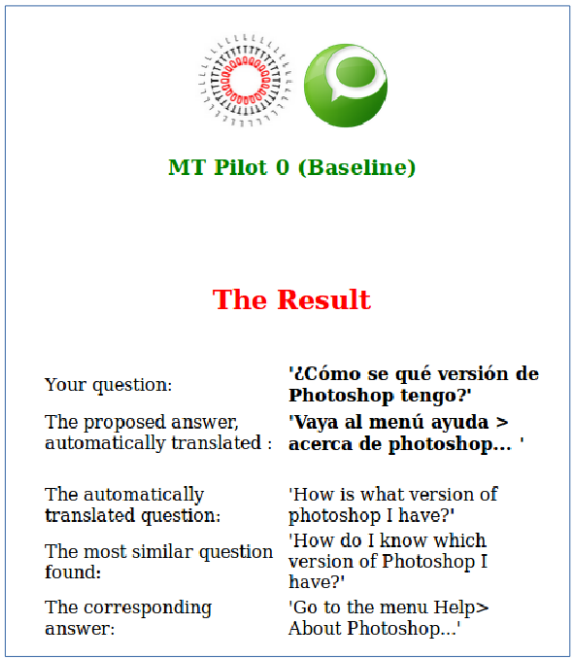

Figure 2: Example of use of PCMEDIC, the real user scenario

the machine translation functionality showed that even with a low-quality machine translation system, it was already possible to achieve a very significant reduction of human operation of about 60% on average for each new language to be covered by the service. This technical advance leverages a great deal of advantages for this kind of business with regard to its extension to the single digital market, as well for its improvement in terms of productivity and resources optimization, with the consequent effective reduction of costs.

There is a common path of progression for each pair X<->EN, ensuring comparability of the research exercise: every pair is developed along pilots 0 to 3 reinforced by complementary strengths and backgrounds of the different partners with their systems, resources and technology.

6 Advisory Board

The direction of the project is informed by the advice on strategic issues from the Advisory Board of Potential Users. This Advisory Board includes industrial participants that are ready to contribute with their advice on the strategic course of the project activities, and are interested in the innovation potential of the results targeted at by the project and will be in the first row of the potential users that will take the lead to exploit their business potential. The members of this board are:

- CA Technologies Development Spain S.A (Spain)
- Eleka Ingeniaritza Linguistikoa SL (Basque Country, Spain)
- OMQ GmBH (Germany)
- Lingea s.r.o. (Czech Republic)
- Seznam.cz, a.s. (Czech Republic)
- Higher Functions, Lda (Portugal, also partner)

Information on QTLeap project and contact details:
Website: http://qtleap.eu/
Facebook: https://www.facebook.com/qtleap
Twitter: https://twitter.com/QTLeap

7 Aknowledgements

The research leading to these results was carried out as part of the QtLeap project funded by European Community (FP7-ICT-2013.4.1-610516)

References

Agerri, R., J. Bermudez, and G. Rigau. 2014. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. 9th Language Resources and Evaluation Conference (LREC2014), Reykjavik, Iceland. pages 26–36.

Branco, A. and P. Osenova. 2014. QTLeap - Quality Translation with Deep Language Engineering Approaches. Poster at EAMT2014, Dubrovnik.

Popel, M. 2014. MT Pilot 1: Entry-level Deep MT. Internal presentation in QTLeap project Meeting. Lisbon.

Zeman, D., O. Dušek, D. Mareček, M. Popel, L. Ramasamy, J. Štěpánek, Z. Žabokrtský, and J. Hajič. 2014. Hamledt: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*, 48(4):601–637.

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 173-176 recibido 24-03-2015 revisado 27-04-2015 aceptado 13-05-2015

Sistema de diálogo basado en mensajería instantánea para el control de dispositivos en el internet de las cosas*

Instant messaging-based dialog system for device control in the Internet of things

José Ángel Noguera-Arnaldos
Proyectos y soluciones tecnológicas avanzadas, SLP (Proasistech)
Edificio CEEIM, Campus Campus de Espinardo, 30100, Murcia, España
jnoguera@proasistech.com

Mario Andrés Paredes-Valverde, Rafael Valencia-García
Universidad de Murcia
Facultad de Informática
Campus de Espinardo, 30100, Murcia, España
marioandres.paredes@um.es, valencia@um.es

Miguel Ángel Rodríguez-García
King Abdullah University of Science and Technology, 4700 Thuwal, Kingdom of Saudi Arabia

Resumen: La finalidad del proyecto im4Things es el desarrollo de una herramienta que proporcione una interfaz de comunicación entre humanos y dispositivos en la Internet de las cosas mediante diálogo en lenguaje natural escrito a través de servicios de mensajería instantánea. Esta comunicación puede ser de distintos tipos tales como el envío de órdenes, la consulta del estado e incluso se permite que sean los mismos dispositivos los encargados de alertar al usuario, si se ha producido un cambio del estado en los sensores de los dispositivos. Este proyecto está siendo desarrollado conjuntamente por la empresa Proasistech y el grupo TECNOMOD de la Universidad de Murcia y ha sido financiado por los fondos propios de la empresa Proasistech y con un contrato de I+D+i de asesoría tecnológica con el citado grupo de la Universidad de Murcia.
Palabras clave: sistemas de diálogo, interfaces en lenguaje natural, ontologías, internet de las cosas.

Abstract: The im4Things project aims to develop a communication interface to devices on the Internet of the Things (IoT) through intelligent dialogue based on written natural language over instant messaging services. This communication can be established in different ways such as order sending, status querying and even the devices themselves are responsible for alert users when a change has been produced in the devices sensors. This project is being developed by Proasistech company in cooperation with the TECNOMOD research group of the University of Murcia and it has been funded by equity capital of Proasistech company and by an R&D&i technology consulting contract with the aforementioned University of Murcia research group.
Keywords: dialog systems, natural language interfaces, ontologies, internet of things

1 Introducción y objetivos del proyecto

En la última década, los avances tecnológicos en el mercado de las aplicaciones informáticas para teléfonos móviles inteligentes o Smartphones orientadas al control de dispositivos electrónicos han sido meteóricos. Gracias a estos avances, los usuarios de estas aplicaciones pueden disfrutar, por ejemplo, de imágenes en tiempo real del interior de su casa, del estado de los toldos de su jardín, de sistemas de información deportiva, entre otras cosas.

* Este trabajo ha sido financiado por la empresa Proasistech (http://www.proasistech.com/) a través de sus fondos propios.

Por otro lado, el concepto de Internet of Things (IoT) pretende que exista una conectividad global tanto entre humanos como entre objetos físicos cotidianos a través de la Internet. Esta forma de conectividad global abre miles de posibilidades y aplicaciones a usuarios de teléfonos móviles inteligentes, tabletas y Smart TV.

Entre las aplicaciones de control de dispositivos se puede destacar el gran avance en los últimos años de sistemas que permiten interactuar con distintos aparatos mediante diálogos en lenguaje natural tanto hablado como escrito. Un ejemplo de estas aplicaciones es Mayordomo (Espejo, et al. 2010), un sistema diálogo multimodal que permite la interacción con un entorno de Inteligencia Ambiental para una vivienda a través del habla.

Sin embargo, la mayoría de estos sistemas se han centrado en el control de los dispositivos de la vivienda de manera centralizada usando para ello tecnologías de conexión domótica como EIB-KNX, X10 o sistemas propietarios.

El principal reto de este proyecto es el desarrollo de un sistema que permita el control y consulta de dispositivos tales como electrodomésticos o sistemas industriales, de manera distribuida a partir del desarrollo de un sistema de diálogo en lenguaje natural escrito a través de mensajería instantánea.

Para ello, se incluirán los dispositivos dentro de una aplicación de mensajería instantánea sobre la cual se pueden crear grupos de usuarios, grupos de dispositivos, compartir dispositivos, entre otras cosas.

2 Estado actual del proyecto

Hasta el momento se ha definido la arquitectura del sistema la cual está formada por tres módulos principales (ver Figura 1). Por un lado, el cliente de aplicación móvil implementa un chat de mensajería instantánea a través del cual se conecta con el servicio im4Things, este a su vez se encarga de enviar las órdenes y consultas a cada dispositivo. Dentro de cada dispositivo existe un sistema de diálogo que permite ejecutar órdenes y consultar el estado del dispositivo, así como detectar alguna alerta que pueda ser susceptible de avisar al usuario en lenguaje natural. Todas las posibles acciones, funciones, sensores y estados del dispositivo se configuran mediante una ontología que incorpora también información lingüística sobre estas características.

A continuación se describe brevemente la arquitectura de la plataforma y cada uno de sus módulos en el estado actual del proyecto.

2.1 Arquitectura de la plataforma im4Things.

Como se ha comentado anteriormente, el sistema im4Things (ver Figura 1) está compuesto por tres módulos principales: la aplicación móvil de chat im4Things, el servicio im4Things y el dispositivo.

Figura 1: Arquitectura de im4Things

2.2 Aplicación chat im4Things.

La app desarrollada permite una comunicación por mensajería instantánea y en tiempo real, entre usuarios y dispositivos. Esta comunicación se realiza mediante el protocolo XMPP que permite gestionar la mensajería instantánea con robustos sistemas de seguridad.

Para comenzar a utilizar esta aplicación es necesario registrarse. Entonces, el usuario podrá dar de alta a otros usuarios y a aparatos que tengan el hardware citado. La aplicación

permite registrar aparatos por medio de un código de inserción que viene con el aparato. El usuario podrá identificar a ese aparato con un nombre y contraseña. Además, permite otras opciones como envío de archivos, compartición de ubicación y contactos, creación de grupos de usuarios humanos o electrodomésticos, entre otras cosas.

La interfaz de usuario está diseñada para que su manejo sea intuitivo recordando a otras aplicaciones de mensajería instantánea similares y extendidas en la sociedad actual.

2.3 Servicio im4Things.

Este servicio es el encargado de mantener la comunicación de mensajería instantánea con los usuarios y aparatos mediante un servicio MongooseIM[1] de forma segura.

Este servicio se ha adaptado para que pueda integrarse con distintas funcionalidades tales como el registro, la mensajería, sincronización por teléfono, notificaciones push, registro de dispositivos, lista de usuarios por dispositivo, por mencionar algunas.

2.4 Dispositivo.

El dispositivo contiene también una versión de la aplicación de chat que es capturada por un módulo de diálogo, el cual se encarga de mantener la conversación con el usuario y el estado del dispositivo.

Para dotar a los dispositivos de la capacidad de procesar la información se ha diseñado un hardware específico formado por sensores y un sistema de control y comunicaciones que interactúa con el módulo de diálogo para poder procesar las órdenes y el estado del dispositivo de manera eficiente. Para el primer prototipo, este hardware está basado en Raspberry debido a que se ha utilizado en sistemas de control y reconocimiento de voz como el presentado en (Haro et al., 2014)

A continuación se explica un poco más en detalle el sistema de diálogo.

2.4.1 Sistema de diálogo.

Tradicionalmente, los sistemas de diálogo escrito llevan a cabo tres tareas que se suelen implementar en distintos módulos de la arquitectura: comprensión del lenguaje natural, la gestión del diálogo y la generación del lenguaje. En este caso también es necesario un módulo que permita la comunicación con el dispositivo para la consulta de su estado, ejecutar los comandos, o bien gestionar la comunicación que puede hacer el dispositivo con el usuario.

Como se puede observar en la arquitectura descrita en la Figura 2, el sistema de diálogo se basa en una ontología que representa la información y conocimiento sobre el dispositivo tal como las acciones que puede realizar, los sensores que contiene, los servicios que ofrecen dichos sensores, las alertas que pueden lanzarse por estos sensores, los posibles estados que puede tener, así como otros dispositivos integrados en el dispositivo.

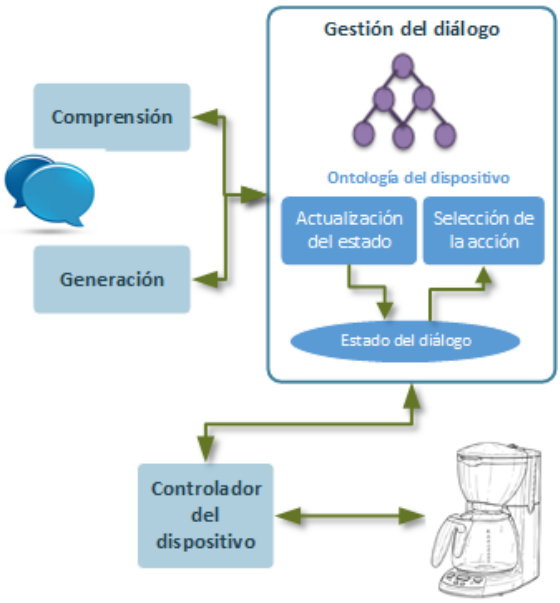

Figura 2: Arquitectura del sistema de diálogo.

El módulo de comprensión se encarga de analizar si el texto recibido corresponde a una consulta o a una orden y procesa su significado para poder actuar en consecuencia. Este módulo se basa en trabajos anteriores del grupo de investigación de la Universidad de Murcia como (Paredes-Valverde et al., 2015), donde se representa la consulta mediante una ontología de la pregunta (question model), la cual mantiene la información concerniente a los elementos contenidos en la frase, así como de los elementos de la ontología del dispositivo que se han identificado, tales como el verbo principal, el foco de la pregunta, modificadores y otra información lingüística.

[1] https://www.erlang-solutions.com/products/mongooseim-massively-scalable-ejabberd-platform

Por otro lado, la gestión del diálogo se basa en AIML (Wallace, 2003), que es un lenguaje basado en XML para crear sistemas conversacionales. Este lenguaje se utiliza para definir plantillas de patrones que producen una respuesta según una correspondencia sencilla de tokens de entrada. Sin embargo, AIML no está diseñado para procesar el lenguaje y comprender el significado de las frases. Por esta razón, en este trabajo se utiliza una versión extendida de AIML donde en vez de utilizar tokens, se hace referencia a elementos de la ontología del dispositivo e información proporcionada por el módulo de comprensión.

El módulo de gestión del diálogo obtiene los posibles patrones (comandos o consultas) que más se ajustan al texto recibido, analiza y comprueba si es posible realizar esa orden o consulta y la ejecuta en su caso cambiando el estado y comunicándose con el controlador del dispositivo y el módulo de generación.

En el caso de un comando, se ejecuta por el controlador del dispositivo quien indica si este comando se ha ejecutado correctamente en el dispositivo. Entonces el módulo de gestión del diálogo se comunicará con el módulo de generación para que pueda generar una respuesta en lenguaje natural para ser enviada al usuario.

Cabe destacar que el controlador del dispositivo, además de ejecutar los comandos dentro del mismo, monitoriza las alertas que le llegan desde el dispositivo y gestiona su estado. En este caso, el controlador avisa al sistema de gestión del diálogo para que comunique al usuario que ha ocurrido esta alerta.

3 Trabajo futuro

Actualmente existe una primera versión funcional del prototipo completo y se han desarrollado todos los componentes. Además, se ha configurado este prototipo para el control de una cafetera, lavadora y sistema de riego. Las siguientes líneas de trabajo estarán dedicadas a depurar el funcionamiento de los módulos. La mayor parte de esta tarea de depuración se centrará en el sistema de diálogo y los sistemas de sensores. Más concretamente, para la validación se seleccionarán alrededor 50 posibles usuarios para que validen el sistema, mediante un cuestionario gráfico que no los condicione sobre el lenguaje que tienen que utilizar para interactuar con los dispositivos. Con este estudio se analizará la precisión y exhaustividad del sistema. Además, se pretende también refinar este motor de diálogo para que pueda generar lenguaje ya que actualmente obtiene un conjunto de patrones parametrizados para responder al usuario.

Por otro lado, se está trabajando en un asistente de configuración de dispositivos que permita, a partir de la ontología, generar de manera semi-automática la mayor parte de la configuración y patrones definidos en la plataforma con el fin de facilitar la configuración de los dispositivos al usuario.

También se está estudiando la posibilidad de poder interactuar con los dispositivos a través del habla usando tecnologías VoiceXML de manera similar a como se realiza en (Griol et al., 2014).

Por último, se están realizando prototipos para el control y consulta de instalaciones industriales complejas a partir de las tecnologías desarrolladas en este proyecto.

Bibliografía

Espejo, G., Ábalos, N., López-Cózar Delgado, R., Callejas, Z., y Griol, D. 2010. Sistema Mayordomo: uso de un entorno de inteligencia ambiental a través de un sistema de diálogo multimodal. *Procesamiento del Lenguaje Natural*, 45:309-310.

Griol, D., García-Jiménez, M., Molina, J. M., y Sanchis, A. 2014. Desarrollo de portales de voz municipales interactivos y adaptados al usuario. *Procesamiento del Lenguaje Natural*, 53:185-188.

Haro, F. D., Cordoba, R., Rojo Rivero, J.I., Diez de la Fuente, J., Avendano Peces, D. y Bermudo Mera, J.M. 2014. Low-Cost Speaker and Language Recognition Systems Running on a Raspberry Pi. *Latin America Transactions IEEE (Revista IEEE America Latina)*, 12(4): 755-763.

Paredes-Valverde, Mario Andrés, Miguel Ángel Rodríguez-García, Antonio Ruiz-Martínez, Rafael Valencia-García, and Giner Alor-Hernández. 2015. ONLI: An Ontology-Based System for Querying DBpedia Using Natural Language Paradigm. *Expert Systems with Applications*.To appear. Accessed March 13. doi:10.1016/j.eswa.2015.02.034.

Wallace, R. 2003. The elements of AIML style. Alice AI Foundation.

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 177-180 recibido 26-03-2015 revisado 25-4-2015 aceptado 12-05-2015

Explotación y tratamiento de la información disponible en Internet para la anotación y generación de textos adaptados al usuario

Exploitation and Processing of Online Information for Annotating and Generating Texts Adapted to the User

Elena Lloret, Yoan Gutiérrez, Fernando S. Peregrino, José Manuel Gómez, Antonio Guillén, Fernando Llopis
Universidad de Alicante
Carretera San Vicente del Raspeig s/n 03690, Alicante, España
{elloret,ygutierrez,fsperegrino,jmgomez,aguillen,llopis}@dlsi.ua.es

Resumen: La gran cantidad de información disponible en Internet está dificultando cada vez más que los usuarios puedan digerir toda esa información, siendo actualmente casi impensable sin la ayuda de herramientas basadas en las Tecnologías del Lenguaje Humano (TLH), como pueden ser los recuperadores de información o resumidores automáticos. El interés de este proyecto emergente (y por tanto, su objetivo principal) viene motivado precisamente por la necesidad de definir y crear un marco tecnológico basado en TLH, capaz de procesar y anotar semánticamente la información, así como permitir la generación de información de forma automática, flexibilizando el tipo de información a presentar y adaptándola a las necesidades de los usuarios. En este artículo se proporciona una visión general de este proyecto, centrándonos en la arquitectura propuesta y el estado actual del mismo.
Palabras clave: PLN, Ontología, Paquete semántico, Generación de textos

Abstract: The great amount of available online information is making increasingly more and more difficult that users can assimilate such as volume of information, being this almost inconceivable without using Human Language Technologies (HLT) tools, for instance, information retrieval systems or automatic summarisers. The interest of this emerging project (and therefore its main goal) is precisely motivated by the need to define and create a HLT-based technological framework, able to process and semantically annotate all this information, allowing also the automatic generation of information, and making the type of information to be presented more flexible by adapting it to the users' needs. This article provides an overview of this project, focusing on the proposed architecture and its current status.
Keywords: NLP, Ontology, Semantic package, Text Generation

1 Introducción y objetivo

Actualmente, Internet cuenta con más de 2.400 millones de usuarios[1], dato que implica que más del 30 % de la población mundial está conectada. Además, desde la aparición de la Web 2.0 (o Web social), se han creado nuevos sitios Web dónde los usuarios juegan un papel más activo, a través de los que pueden participar, interactuar e intercambiar información con otros usuarios (por ejemplo, foros, blogs, redes sociales, microblogs, etc.).

Sin embargo, el principal inconveniente de toda esta gran cantidad de información disponible es la complejidad en lo que respecta a su procesamiento y tratamiento, sobre todo si el usuario desea obtener información con mayor o menor detalle acerca de un tema concreto. Dicha información se encuentra en distintas fuentes de información de distinta naturaleza y en distintos idiomas. Estos factores, junto a la redundancia existente en la Web y las opiniones y hechos contradictorios que aparecen, hacen que los usuarios inviertan mucho más tiempo de lo deseado navegando, buscando y seleccionando la información que es de su interés.

En este sentido, las Tecnologías del Len-

[1] http://www.internetworldstats.com/stats.htm

ISSN 1135-5948

guaje Humano (TLH) son clave para facilitar al usuario la gestión de toda esta información. Actualmente la investigación en esta área suele centrarse en una tarea específica e independiente, como puede ser la recuperación de información (Vila et al., 2013), minería de opiniones (Fernández, Gómez, y Martínez-Barco, 2010), desambiguación del sentido de las palabras (Gutiérrez et al., 2013) o generación de resúmenes (Vodolazova et al., 2013). Sin embargo, dadas las necesidades del contexto actual, donde la información crece a un ritmo exponencial, es necesario aunar esfuerzos en las distintas tareas hacia la creación de un marco flexible capaz de identificar el tipo de información que necesita el usuario, buscarla, procesarla y presentársela de manera adecuada, para que, por un lado, le ahorre tiempo de proceso y, por otro, le sea útil respecto a sus intereses.

El objetivo principal de este proyecto de investigación[2] es analizar, proponer y evaluar diferentes enfoques novedosos para la anotación y generación de textos adaptados al usuario, creando un marco inteligente que combine e integre distintas aplicaciones de TLH y sea de referencia para la comunidad investigadora. La generación de textos que se propone en este proyecto es flexible, puesto que el resultado no va a ser siempre un texto con el mismo formato, sino que se obtendrá un paquete de información que contendrá anotaciones a distintos niveles, y permitirá utilizar y extraer aquellas que se consideren más apropiadas según el contexto y las necesidades de los usuarios, como son resúmenes, tuits, valoraciones de opiniones, pasajes, recopilación de fuentes relevantes, etc. Esto permitirá una mejor gestión de la información disponible en Internet, proporcionando al usuario información con mayor o menor detalle sobre los temas que le interesen, que le ayudarán en multitud de tareas, incluyendo la consulta de información y/o novedades, y toma de decisiones.

Un valor añadido del proyecto es que los resultados del marco de TLH podrán ser consumidos tanto por seres humanos como por agentes informáticos. La posibilidad de compatibilizar la salida con agentes informáticos, extiende el umbral de éxitos de la propuesta de marco de TLH hacia los horizontes del mercado industrial y/o empresarial, pues posibilita que se establezcan intereses comunes entre ambas comunidades, la científica y la empresarial.

[2]http://gplsi.dlsi.ua.es/gplsi11/en/node/16595

2 *Arquitectura general para el marco de TLH*

En la figura 1 se ilustra el marco TLH propuesto, donde podemos observar cómo este marco permite a los usuarios consultar información de Internet (medios sociales, foros, noticias, etc.) y dependiendo de las necesidades que tenga el usuario, presentarle la información de una u otra manera (por ejemplo, mediante un tuit, un resumen, una valoración de un tema, etc.).

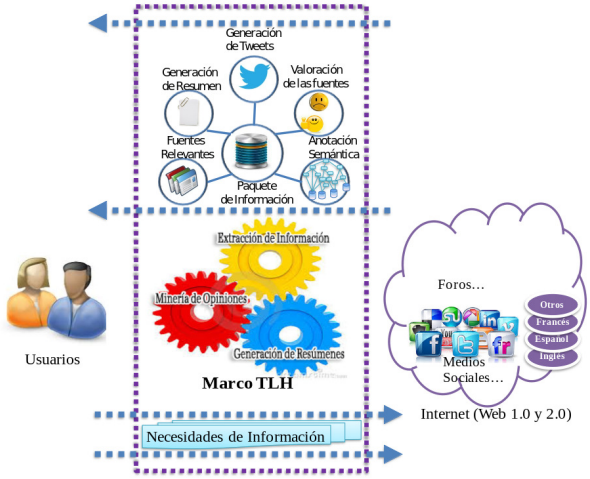

Figura 1: Ilustración del marco de TLH

Tal y como se aprecia en la figura, las herramientas y procesos de TLH son el elemento central y juegan un papel clave desde dos puntos de vista: por un lado, se utilizan para buscar y recuperar la información necesaria de la Web, y por otro, son esenciales en el procesamiento y tratamiento inteligente de dicha información para obtener conocimiento. Concretamente, se integrarán tecnologías como el análisis semántico, la recuperación y la extracción de información, la minería de opiniones y la generación de resúmenes. Aunque el marco no está limitado a la integración de estas aplicaciones, sí que es cierto que estas tareas serán las que conformen su núcleo central, y por tanto, serán cruciales para el correcto desarrollo del proyecto.

Para llevar a cabo la generación de textos debemos en primer lugar decidir qué información recuperar y seleccionarla. Posteriormente se procesará dicha información, ya sea subjetiva u objetiva, y para ello, será necesario

identificar el tipo de información, clasificarla, detectar lo realmente importante, determinar información redundante, complementaria y/o contradictoria e integrar y combinar todo el conocimiento obtenido. Todo este conocimiento obtenido quedará anotado de forma automática en un paquete de información (lo que denominaremos en nuestro proyecto "paquete semántico"), en base a una ontología previamente diseñada. Una vez poblada la ontología, la última fase sería generar un texto que cubra las necesidades de los usuarios y que pueda ser mostrado en base a diferentes formatos, a partir de las anotaciones que contenga. O bien, como se ha comentado en la sección 1 que directamente el documento anotado semánticamente pueda ser procesado y utilizado por otras aplicaciones informáticas.

3 Estado del proyecto

Durante el primer año, nos hemos centrado en la definición y creación del paquete semántico. Este proceso abarca desde el diseño de la ontología para representar distintos tipos de información que puede contener un paquete semántico, hasta el desarrollo de un proceso automático que sea capaz de integrar las herramientas de TLH a utilizar y poblar la ontología de forma automática.

3.1 Definición del paquete semántico

Como base para la definición del paquete semántico, hemos diseñado una ontología utilizando la herramienta Protégé[3]. Dicha ontología contiene tanto información léxica como semántica de cómo hemos decidido que se representen los documentos, las frases y el resumen derivado. Junto con la definición y diseño de la ontología, tenemos asociadas un conjunto de preguntas de competencia que serán las que la ontología deberá resolver (por ejemplo, *"¿qué resúmenes refieren hechos que datan del día dd/mm/aaaa?"*o *"¿qué entidades nombradas están implicadas en documentos del dominio deportivo?"*).

La figura 2 muestra la jerarquía de clases definidas en el paquete semántico.

3.2 Creación del paquete semántico

Una vez definida y diseñada la ontología, la fase de creación del paquete semántico está compuesta por tres módulos que se ejecutarán de manera secuencial, y que juntos van a constituir el núcleo central del marco de TLH. A continuación, se explicará cada uno de estos módulos con más detalle.

3.2.1 Gestor de fuentes de información

En primera instancia, necesitamos disponer del conjunto de fuentes de información de partida. Por lo tanto, este primer módulo tiene como objetivo descargar las fuentes de información con las que se desee trabajar y extraer el texto que posteriormente se procesará.

3.2.2 Integrador de procesos de TLH

Este módulo es el encargado de ejecutar los procesos de TLH deseados y determinar las entradas y salidas de cada uno. Para un primer prototipo, hemos seleccionado un conjunto de herramientas de TLH para poder procesar los documentos. Como premisa, hemos optado en la medida de lo posible reutilizar herramientas ya existentes en cada una de las áreas que han demostrado ser competitivas en su ámbito. Estas herramientas se resumen en la tabla 1 y todas ellas funcionan para el idioma inglés[4].

Proceso TLH	Herramienta	Institución
A.Semántico	ISR-Wordnet	UA
A.Sentimientos	Sentiment	UA
Gen.Resúm	GPLSICompendium	UA
Rec.NER	StandfordNER	Standford
Rec.ExprTemp	TipSem	UA

Tabla 1: Procesos de TLH integrados

3.2.3 Anotación semántica

La finalidad de este módulo es poblar la ontología previamente diseñada en base a la información proporcionada por los procesos de TLH. Como resultado de ejecutar estos tres módulos, tendremos ya creado el paquete semántico listo para poder hacer diferentes consultas en función de las necesidades de información o bien para ser integrado en otros procesos automáticos. A modo ilustrativo, un ejemplo de algunos componentes que integrarían el paquete semántico pueden verse gráficamente en la figura 3.

[3] http://protege.stanford.edu/

[4] Algunas de estas herramientas están accesibles a través de: http://gplsi.dlsi.ua.es/services/pln/doc/index.html

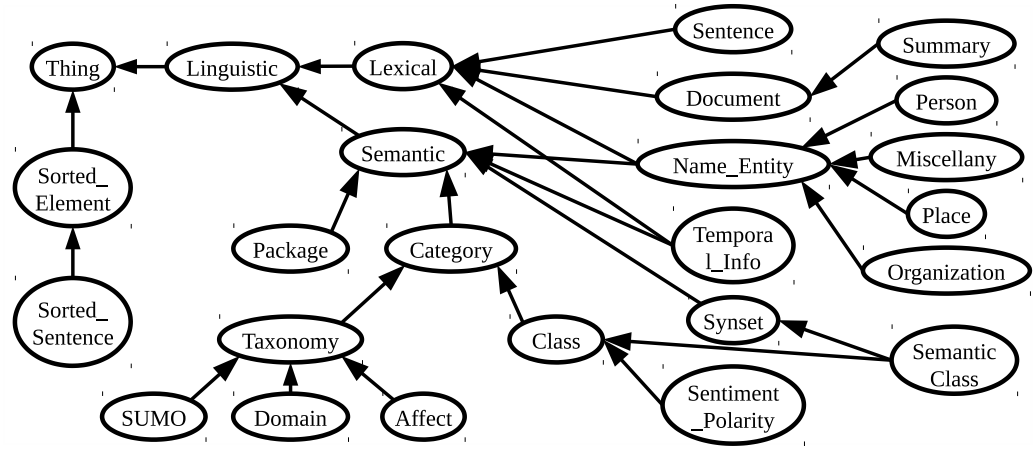

Figura 2: Jerarquía de clases de la ontología para un paquete semántico

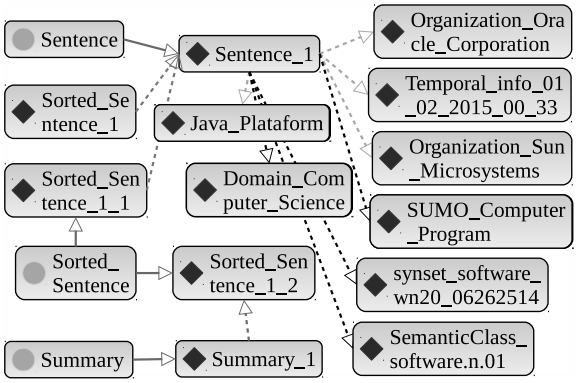

Figura 3: Ejemplo de una frase y los conceptos léxico-semánticos que contiene

4 Trabajo futuro

De cara al segundo año, nos vamos a centrar en analizar y estudiar métodos para presentar la información del paquete semántico de manera flexible y adaptada, para evaluarla mediante estudios de usuario. También estudiaremos la posibilidad de integrar y aplicar directamente el paquete semántico a otras posibles tareas de TLH. A la finalización del proyecto se pretende dejar accesible el marco de TLH desarrollado a través de la API de servicios del grupo GPLSI, así como la ontología diseñada para la creación del paquete semántico.

Agradecimientos

Este proyecto ha sido financiado por la Universidad de Alicante a través del proyecto emergente "Explotación y tratamiento de la información disponible en Internet para la anotación y generación de textos adaptados al usuario" (GRE13-15) y su temática se enmarca en el contexto de los proyectos "DIIM2.0: Desarrollo de técnicas Inteligentes e Interactivas de Minería y Generación de Información sobre la Web 2.0" (PROMETEOII/2014/001) financiado por la Generalitat Valenciana y el proyecto "Técnicas de Deconstrucción en la Tecnologías del Lenguaje Humano" (TIN2012-31224) financiado por Ministerio de Economía y Competitividad del Gobierno de España.

Bibliografía

Fernández, J., J. M. Gómez, y P. Martínez-Barco. 2010. Evaluación de sistemas de recuperación de información web sobre dominios restringidos. *Procesamiento del lenguaje natural*, 45:273–276.

Gutiérrez, Y., Y. Castaneda, A. González, R. Estrada, D. D Piug, J. I. Abreu, R. Pérez, A. Fernández Orqun, A. Montoyo, R. Muñoz, y F. Camara. 2013. UMCC DLSI: Reinforcing a ranking algorithm with sense frequencies and multidimensional semantic resources to solve multilingual word sense disambiguation. *Proc. of SemEval*, páginas 241–249.

Vila, K., A. Fernández, J. M. Gómez, A. Ferrández, y J. Díaz. 2013. Noise-tolerance feasibility for restricted-domain information retrieval systems. *Data & Knowledge Engineering*, 86:276–294.

Vodolazova, T., E. Lloret, R. Muñoz, y M. Palomar. 2013. Extractive text summarization: Can we use the same techniques for any text? En *Natural Language Processing and Information Systems*. Springer, páginas 164–175.

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 181-184 recibido 27-03-2015 revisado 27-04-2015 aceptado 13-05-2015

Socialising Around Media (SAM): Dynamic Social and Media Content Syndication for Second Screen*

Socialising Around Media (SAM): Sindicación Dinámica de Contenido Social y Multimedia para Segundas Pantallas

David Tomás, Yoan Gutiérrez, Isabel Moreno, Francisco Agulló
University of Alicante, Spain
{dtomas,ygutierrez,imoreno,fagullo}@dlsi.ua.es

Marco Tiemann
University of Reading, United Kingdom
m.tiemann@reading.ac.uk

Juan V. Vidagany
TIE Kinetix, The Netherlands
juanvi.vidagany@tiekinetix.com

Andreas Menychtas
National Technical University of Athens, Greece
ameny@mail.ntua.gr

Resumen: La actual generación de dispositivos móviles ha cambiado la forma en que los usuarios interactúan con los medios digitales, pasado de ser pasivos y unidireccionales a proactivos e interactivos. Los usuarios usan estos dispositivos para comentar y valorar programas televisivos, buscando información relacionada sobre personajes, hechos y celebridades. Este fenómeno se conoce con el nombre de *segunda pantalla*. En este artículo se describe SAM, un proyecto de investigación financiado por la EU y enfocado al desarrollo de una plataforma avanzada de distribución de contenidos digitales basada en segundas pantallas, usando la sindicación de contenidos en el contexto de los medios sociales para proporcionar maneras abiertas y estándares de caracterizar, descubrir y sindicar recursos digitales. Este trabajo resume las principales características y objetivos del proyecto, así como los retos de PLN a afrontar y las tecnologías desarrolladas para tal fin hasta el momento.
Palabras clave: Sindicación de contenidos, segundas pantallas, medios sociales

Abstract: Today's generation of Internet devices has changed how users are interacting with media, from passive and unidirectional users to proactive and interactive. Users can use these devices to comment or rate a TV show and search for related information regarding characters, facts or personalities. This phenomenon is known as *second screen*. This paper describes SAM, an EU-funded research project that focuses on developing an advanced digital media delivery platform based on second screen interaction and content syndication within a social media context, providing open and standardised ways of characterising, discovering and syndicating digital assets. This work provides an overview of the project and its main objectives, focusing on the NLP challenges to be faced and the technologies developed so far.
Keywords: Content syndication, second screen, social media

1 Introduction

The current generation of Internet devices has changed how users are interacting with media, from passive and unidirectional users to proactive and interactive. Users can comment or rate a TV show and search for related information regarding characters, facts or personalities. This phenomenon is known as *second screen*. In today's second screen environment there are no true standards, protocols or commonly used frameworks through which users can discover and access information related to consumed contents. Users have to actively perform searches using web search

* This work has been partially funded by the European Commission under the Seventh Framework Programme (FP7 - 2007- 2013) for Research and Technological Development through the SAM project (FP7-611312), and by the Spanish Government through the LEGOLANGUAGE project (TIN2012-31224).

ISSN 1135-5948

engines such as Google to "participate" in TV shows.

This paper describes SAM[1] (Socialising Around Media), an EU-funded project focusing on developing an advanced digital media delivery platform for second screen and content syndication within a social media context, providing open and standardised ways of characterising, discovering and syndicating digital assets (e.g. films, songs, books and metadata).

The potential customers of SAM are both business stakeholders (such as media broadcasters, content asset providers, software companies and digital marketing agencies) and end users. For the former, this platform will provide a number of benefits, including dynamic social and media content syndication, managing of online reputation, listening to customers, tracking real time statistics or monitoring media related social content through second screen. For the latter, SAM will offer a complete solution for people consuming media and TV programs. The platform will integrate context aware information and complex social functionalities providing contextual information about their actual and current interest. This will provide the end users with an augmented experience in which they can discover new information about the subject, and will talk and share their experience with other users that are also interested in the same topic. Users will produce and consume digital assets from different syndicated sources and synchronised devices (e.g. tablets, smartphones and connected TVs) thus creating richer experiences around the original assets.

The key innovation of SAM is that instead of users reaching for the data, it is the data that reaches the users through the syndication approach and their second screen. Core functionalities of SAM are based on Natural Language Processing (NLP) technologies—including sentiment analysis, text summarisation and semantic analysis—which are applied to both formal texts and user generated content in social networks (i.e. user comments).

The remainder of this article is organised as follows: next section provides information about the members of the SAM consortium; Section 3 describes the three core technologies supporting the development of this project; Section 4 summarises the main goals of the project; Section 5 describes the challenges faced in SAM by NLP technologies; finally, Section 6 describes the current development status of the project.

2 The Consortium

SAM involves nine partners from a pan-European consortium, presenting a well-balanced combination of research institutions, software developers and user companies (including a media broadcaster, a content provider and a smart TV manufacturer):

- *Research institutions*: National Technical University of Athens (Greece), University of Reading (United Kingdom) and University of Alicante (Spain).
- *Technology SMEs*: TIE Kinetix, coordinator (The Netherlands), Ascora GmbH (Germany) and Talkamatic (Sweden).
- *User companies*: Deutsche Welle (Germany), Bibliographic Data Services (United Kingdom) and TP Vision (Belgium).

3 The Three Pillars of SAM

The SAM platform has been designed around three pillars that highlight the main research and business directions of this project: content syndication, second screen and social media. By combining these pillars together, SAM will implement the distribution of media assets to the end users through their devices, including linked content and related information to enrich the user experience.

Content Syndication Technologies like Really Simple Syndication (RSS) have their place in the syndication world, but when an organisation needs to push more enriched and target-adapted information to partner websites or social networking sites (from product details to full microsites with rich media), approaches such as RSS are not powerful enough. Content syndication solutions today have evolved beyond RSS, allowing vendors, distributors and publishers to issue, control and track rich content experiences on third-party websites. In SAM, content syndication techniques will allow content providers to prepare their digital assets and associate them to specific media and usage context, offering mechanisms for these enriched assets to

[1] `http://www.socialisingaroundmedia.com`.

be delivered in the expected format and to be consumed by the users in a specific context.

Second Screen This concept refers to any electronic device (broadly a mobile device, such as a tablet or smartphone) that allows users to retrieve additional information about the content they are watching on the first screen (usually a TV set). The SAM platform includes a multi-device representation layer that provides syndicated information in the appropriate format to be consumed by different types of devices. This generic approach is used in order to access the asset-related syndicated information while it is being consumed, commented, or interacted with by the users, creating a second screen experience.

Social Media These technologies are changing the way in which users interact and communicate with each other, expressing their feelings, opinions and thoughts about almost anything—including products, personalities and TV shows. In social media, users not only share comments or articles, they also exchange different types of digital assets such as videos, photographs and documents. In recent years, the user activity in social networks has significantly increased, making it into a key area of interest for media business and advertisers. Decision makers try to find ways in which commercial products can make profitable use of applications such as YouTube, Facebook and Twitter. In SAM, social interaction around digital media items will provide the context in which the syndicated content will be consumed. SAM will incorporate complex context extraction mechanisms based on NLP technologies, such as sentiment and semantic analysis, for the creation of dynamic social communities based on users' actions and their context (e.g. assets consumed, demographic profile and preferences).

4 Objectives

The main objective of the SAM project is the development of an advanced digital media delivery platform that provides an open environment for defining, characterising, discovering, socially consuming, syndicating and interacting with media assets. This main goal can be decomposed into several research and technological objectives. One of these objectives is the dynamic creation of social communities based on user-consumed media, user behaviour and interests expressed while interacting with the system and with other users. NLP technologies are crucial in this context to provide advanced techniques for data extraction, analysis and characterisation. Sentiment analysis and text summarisation will help to achieve the objective of enabling companies, especially SMEs, to analyse crowdsourced reactions to the assets they publish. Another objective is the use and definition of open and standardised formats for the description of media assets, along with a framework for their configuration and use that could be exploited by third party software companies to easily build second screen social orientated applications.

Besides these research and technological objectives, the project also aims to accomplish different goals related to the exploitation and sustainability of the platform. These objectives include addressing issues paramount for commercial exploitation, such as content curation, user privacy and brand integrity. Another goal in this category is the provision of business models and exploitation plans for further development of the project results and future commercialisation activities.

5 NLP Challenges

Some of the core functionalities of SAM will be supported by NLP technologies, ranging from entity linking to sentiment analysis. This section describes the main challenges that will be faced in this project by means of human language technologies.

Ontology Exploitation One of the challenges in SAM is to define an ontology to represent the properties of the assets and the relations existing between them and with other external sources (such as Wikipedia). All the assets in SAM will be stored as instances of this ontology. In order to facilitate the process of importing external data into the SAM platform, ontology matching technologies will provide a mapping between the structure of the incoming data and the ontology defined. This task requires the implementation of CRUD operations—create, read, update and delete—over ontology instances.

Data Characterisation This task comprises different problems related to the semantic analysis of text. One of the issues to be solved here is entity linking, the task of identifying entities mentioned in text and

connecting them to instances in a knowledge base (Mihalcea and Csomai, 2007). In SAM, both assets' content and user comments will be analysed in order to identify mentions to entities existing in Wikipedia and assets already stored in SAM. The aim of this analysis is to give content providers the possibility to enrich their assets by linking them to additional internal (assets) and external (Wikipedia) sources of information, and also to identify comments on specific assets, made by the end users, for business intelligence purposes.

Another task related to this subject is asset discovery. In this case, the system will recommend assets to the users based on their context (e.g. assets consumed and user preferences). This task requires context analysis and ontology exploitation to identify suitable instances stored in the platform.

Social Mining Sentiment analysis and text summarisation techniques will be applied to social media in SAM. The purpose is to obtain suitable information from user-generated content to help the business intelligence module of the platform to create advanced reports for content providers. Sentiment analysis will extract intensity, polarity and emotions from users opinions over the assets consumed. Moreover, aspect-based sentiment analysis (Pontiki et al., 2014) will be applied to identify opinions about specific features of an asset (e.g. an actor in a film).

SAM will bring added value to this research area by using these techniques in a novel domain: the creation of dynamic social communities based on the feelings, opinions and interests expressed by the users.

Besides that, novel text summarisation techniques will be applied not only to identify the most relevant comments posted by users regarding an asset or subject, but also to detect the most salient parts of these comments. As it was the case of sentiment analysis, this information will be used by the business intelligence module to supply content providers with advanced reports on user opinions regarding their assets.

6 Current Status of the Project

The project started on September 2013 and will run for 37 months, finishing by the end of October 2016. In its first year, most of the efforts were focused on defining and specifying the SAM concept and the project vision, and carrying out the first stages of the software development life cycle: requirement analysis, global architecture definition, functional specification and technical specification.

SAM is currently in its second year, and at this stage the development of most of the modules in the platform has been started. Regarding NLP functionalities in the project, the following tasks were carried out:

- *Entity linking*: given an input text, the system identifies the occurrence of Wikipedia entities in its content. This module is based on OpenNLP[2] and DBpedia Lookup.[3]
- *Ontology definition and exploitation*: the SAM ontology has been defined using Europeana[4] as a basis to represent the properties of the assets, and including additional concepts and attributes to store specific information related to the SAM platform (such as information on the assets' owners).
- *Sentiment analysis*: a first approach has been developed to polarity and intensity detection on user generated content based on machine learning techniques.
- *Ontology matching*: a module has been defined to match incoming data structures with the SAM ontology. The mapping is based on Levenshtein distance between concept/label names, combined with a measure based on the density of the graph obtained from the structures.

News, updates and additional information on the progress of the project can be found in the official web page and in the wiki page of SAM.[5]

References

Mihalcea, R. and A. Csomai. 2007. Wikify!: Linking documents to encyclopedic knowledge. In *CIKM '07*, pages 233–242, New York, NY, USA. ACM.

Pontiki, M., H. Papageorgiou, D. Galanis, I. Androutsopoulos, J. Pavlopoulos, and S. Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *SemEval 2014*, pages 27–35.

[2] `https://opennlp.apache.org/`.
[3] `http://wiki.dbpedia.org/Lookup`.
[4] `http://www.europeana.eu/`.
[5] `http://wiki.socialisingaroundmedia.com/`.

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 185-188 recibido 29-03-2015 revisado 27-04-2015 aceptado 08-06-2015

Automatic Acquisition of Machine Translation Resources in the Abu-MaTran Project *

Adquisición automática de recursos para traducción automática en el proyecto Abu-MaTran

Antonio Toral, Tommi Pirinen, Andy Way,
ADAPT Centre, School of Computing, Dublin City University, Ireland

Raphaël Rubino, Gema Ramírez-Sánchez, Sergio Ortiz-Rojas, Víctor Sánchez-Cartagena, Jorge Ferrández-Tordera,
Prompsit Language Engineering, S.L., Elx, Spain

Mikel Forcada, Miquel Esplà-Gomis,
Dep. de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, Spain

Nikola Ljubešić, Filip Klubička,
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

Prokopis Prokopidis and Vassilis Papavassiliou
Institute for Language and Speech Processing, Athens, Greece
info@abumatran.eu

Resumen: Este artículo presenta una panorámica de las actividades de investigación y desarrollo destinadas a aliviar el cuello de botella que supone la falta de recursos lingüísticos en el campo de la traducción automática que se han llevado a cabo en el ámbito del proyecto Abu-MaTran. Hemos desarrollado un conjunto de herramientas para la adquisición de los principales recursos requeridos por las dos aproximaciones más comunes a la traducción automática, modelos estadísticos (corpus) y basados en reglas (diccionarios y reglas). Todas estas herramientas han sido publicadas con licencias libres y han sido desarrolladas con el objetivo de ser útiles para ser explotadas en el ámbito comercial.
Palabras clave: Traducción automática, adquisición de recursos lingüísticos, cooperación entre universidad y empresa

Abstract: This paper provides an overview of the research and development activities carried out to alleviate the language resources' bottleneck in machine translation within the Abu-MaTran project. We have developed a range of tools for the acquisition of the main resources required by the two most popular approaches to machine translation, i.e. statistical (corpora) and rule-based models (dictionaries and rules). All these tools have been released under open-source licenses and have been developed with the aim of being useful for industrial exploitation.
Keywords: machine translation, acquisition of language resources, industry-academia cooperation

1 Introduction

Abu-MaTran (Automatic bulding of Machine Translation)[1] is a four-year EU Marie-Curie IAPP (Industry-Academia Partnerships and Patways) project (2013–2016) that seeks to enhance industry-academia cooperation as a key aspect to tackle one of Europe's biggest challenges: multilinguality. The consortium is made up of four research institutions (Dublin City University, Universitat d'Alacant, University of Zagreb and the Institute for Language and Speech Processing in Athens) and one industry partner (Prompsit Language Engineering).

We aim to increase the hitherto low indus-

* The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran).

[1] `http://www.abumatran.eu/`

ISSN 1135-5948

trial adoption of machine translation (MT) by identifying crucial research techniques, preparing them to be suitable for commercial exploitation and finally transferring this knowledge to industry. On the opposite direction, we transfer back to academia the know-how of industry regarding management, processes, etc. to make research products more robust. The project exploits the open-source business model, all the resources produced are released as free/open-source software, resulting in effective knowledge transfer beyond the consortium.

While MT is nowadays a rather mature technology, it is still far from being widely adopted in industry. We argue that this has to do with the lack of required language resources (LRs). For example, if we look at the level of MT support for European languages, out of 30 languages, only 3 languages are considered to count with moderate to good support (English, Spanish and French), while the level of support for the remaining 27 languages ranges from fragmentary to weak or even none (Rehm and Uszkoreit, 2013).

An important strand of the Abu-MaTran project aims to alleviate this so-called LR bottleneck by providing ready-to-use tools for the automatic acquisition of the LRs required by MT systems. This paper provides an overview of the research and development actions carried out in the project in this respect. We also detail the resources that have been acquired.

While the tools we develop aim to be generic, we have a specific case study. This case study has been selected according to its strategic interest in the European context. We acquire the required resources to provide MT for the official language of a new member state of the EU (Croatian) and then extend to related South-Slavic languages official in candidate member states, such as Serbian and Bosnian. It should be noted that all these languages are considered to be under-resourced (Rehm and Uszkoreit, 2013).

The rest of the paper is organised as follows. Section 2 deals with the acquisition of resources for statistical MT (SMT) systems, namely corpora. Next, Section 3 regards the acquisition of resources for rule-based MT (RBMT) systems, namely dictionaries and rules. Finally, Section 4 derives conclusions and outlines future work directions.

2 Corpora

This section covers the acquisition of corpora, both monolingual (Section 2.1) and parallel (Section 2.2), as well as the cleaning of noisy parallel corpora (Section 2.3).

2.1 Acquisition of Monolingual Corpora

Monolingual corpora constitute a cheap (in comparison to parallel corpora) and important resource for SMT systems as they can be used to build language models for the target language.

We propose to crawl top-level domains (e.g. `.hr` for Croatia) in order to acquire vast amounts of monolingual data. The procedure has been used to crawl data for Croatian (1.9 billion tokens), Bosnian (429 million tokens) and Serbian (894 million tokens) (Ljubešić and Klubička, 2014) as well as for Catalan (779 million tokens) (Ljubešić and Toral, 2014).

While the previous approach yields general-domain data, we have also developed a novel tool to crawl tweets, given the growing importance of social media. This tool, TweetCat (Ljubešić, Fišer, and Erjavec, 2014), has been used to acquire tweets for Croatian, Serbian and other similar languages (235 million words) as well as for Slovene (38 million words).

2.2 Acquisition of Parallel Corpora

Compared to monolingual corpora, the acquisition of parallel corpora is considerably more complex. While the main aim for building these corpora is to train SMT systems, we envisage other purposes too such as assisting translators (Rubino et al., 2015).

We have built upon two parallel crawlers previously developed by project partners for research purposes, ILSP Focused Crawler (Papavassiliou, Prokopidis, and Thurmair, 2013)[2] and Bitextor (Esplà-Gomis and Forcada, 2010).[3] In Abu-MaTran we have prepared them to be ready for commercial exploitation. As a result, it is now straightforward to use these tools for crawling parallel data (Papavassiliou et al., 2014). In the project we have used these crawlers to acquire parallel corpora for the

[2] `http://nlp.ilsp.gr/redmine/projects/ilsp-fc`

[3] `http://sourceforge.net/projects/bitextor/`

tourism domain (Esplà-Gomis et al., 2014) for the language pair Croatian–English (140 thousand sentence pairs).

2.3 Cleaning of Noisy Parallel Corpora

There are vast amounts of publicly available parallel data that are not clean enough to be usable to be used for training MT systems. We have proposed a cleaning procedure so that these corpora can be useful to train MT systems (Forcada et al., 2014a). We have applied this procedure to OpenSubtitles,[4] a set of corpora made of open-domain subtitles available for several language pairs. The cleaning procedure fixes some recoverable errors and removes noisy sentence pairs (e.g. misaligned pairs). We have evaluated our procedure on the OpenSubtitles corpus for English–Croatian. An SMT system built on the clean version outperforms a system built on the original corpus by approximately 10 BLEU points absolute (Forcada et al., 2014b).

3 RBMT Resources

Part of our research is focused on RBMT systems. These systems have proven to be a sensible choice when translating between related languages, which is the case of the South-Slavic languages covered in Abu-MaTran. One of the weakest points of RBMT is that developing such systems may result expensive, since linguists have to manually encode the translation rules and dictionaries used by these systems. Our research focuses then on the automatic and semi-automatic acquisition of the main resources used by RBMT: dictionaries (Section 3.1) and rules (Section 3.2).

3.1 Dictionaries

We have proposed a novel approach to assist non-expert users to add new words to the morphological dictionaries used in RBMT systems (Esplà-Gomis et al., 2014). Our method helps the user to add unknown words and find their correct morphological paradigm by asking the user about the possible derivations of the word. A hidden Markov model is used to minimise the amount of necessary questions.

[4] `http://opus.lingfil.uu.se/OpenSubtitles.php`

3.2 Transfer Rules

Transfer rules encode the information needed to deal with the grammatical divergences between languages and they are usually developed by linguists. In order to enable the rapid and cheap building of RBMT systems we have developed a novel approach that learns shallow-transfer MT rules from very small amounts (a few hundreds of sentences) of parallel corpora (Sánchez-Cartagena, Pérez-Ortiz, and Sánchez-Martínez, 2015).

Experiments on five language pairs have shown that the translation quality significantly improves that obtained with previous approaches (Sánchez-Martínez and Forcada, 2009) and is close to that obtained with handcrafted rules.

4 Conclusion and Future Work

We have provided an overview of the research and development activities carried out in the Abu-MaTran project to alleviate the LR bottleneck in MT. To this end, we have tackled the acquisition of resources needed to build SMT (corpora) and RBMT systems (dictionaries and rules).

Regarding SMT resources, we have established a robust pipeline to crawl monolingual and parallel corpora that is ready for commercial exploitation. We have also devised a novel procedure to clean publicly available corpora that are not usable for MT as they are.

As for RBMT resources, we have proposed methodologies (i) to enable non-expert users to improve the coverage of morphological dictionaries and (ii) to learn automatically translation rules from very small parallel corpora.

In the remaining two years of the project, we will continue our work on acquisition as follows. Regarding corpora, we plan to combine the approaches for crawling of top-level domains and parallel crawling in a single tool. This will allow users to crawl both monolingual and parallel data for any language that is associated to a top-level domain by issuing a single command. As for linguistic resources, we will apply the tools presented for the acquisition of dictionaries and rules to bootstrap the development of a rule-based MT system for the pair of closely-related languages Croatian–Serbian.

References

Esplà-Gomis, M. and M. L. Forcada. 2010. Combining content-based and url-based heuristics to harvest aligned bitexts from multilingual sites with bitextor. *Prague Bull. Math. Linguistics*, 93:77–86.

Esplà-Gomis, M., V. M. Sánchez-Cartagena, J. A. Pérez-Ortiz, F. Sánchez-Martínez, M. L. Forcada, and R. C. Carrasco. 2014. An efficient method to assist non-expert users in extending dictionaries by assigning stems and inflectional paradigms to unknown words. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation Translation*, pages 19–26, Dubrovnik, Croatia, June.

Esplà-Gomis, M., F. Klubička, N. Ljubešić, S. Ortiz-Rojas, V. Papavassiliou, and P. Prokopidis. 2014. Comparing two acquisition systems for automatically building an english-croatian parallel corpus from multilingual websites. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may.

Forcada, M. L., S. Ortiz-Rojas, T. Pirinen, R. Rubino, and A. Toral. 2014a. Abu-MaTran deliverable D4.1b MT systems for the second development cycle. `http://www.abumatran.eu/?page_id=59`.

Forcada, M. L., T. Pirinen, R. Rubino, and A. Toral. 2014b. Abu-MaTran deliverable D5.1b Evaluation of the MT systems deployed in the second development cycle. `http://www.abumatran.eu/?page_id=59`.

Ljubešić, N., D. Fišer, and T. Erjavec. 2014. TweetCaT: a Tool for Building Twitter Corpora of Smaller Languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Ljubešić, N. and F. Klubička. 2014. {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden.

Ljubešić, N. and A. Toral. 2014. cawac - a web corpus of catalan and its application to language modeling and machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may.

Papavassiliou, V., P. Prokopidis, M. Esplà-Gomis, and S. Ortiz. 2014. Abu-MaTran deliverable D3.2. Corpora Acquisition Software. `http://www.abumatran.eu/?page_id=59`.

Papavassiliou, V., P. Prokopidis, and G. Thurmair. 2013. A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 43–51, Sofia, Bulgaria, August.

Rehm, G. and H. Uszkoreit. 2013. META-NET Strategic Research Agenda for Multilingual Europe 2020. `http://www.meta-net.eu/vision/reports/meta-net-sra-version_1.0.pdf`. [Online; accessed 27 March 2015].

Rubino, R., M. Esplà-Gomi, A. Toral, V. Papavasiliou, and P. Prokopidis. 2015. DIY Domain Specific Parallel Corpora for Translators. In *To appear in Proceedings of the IV International Conference on Corpus Use and Learning to Translate (CULT)*, Alacant, Spain.

Sánchez-Cartagena, V. M., J. A. Pérez-Ortiz, and F. Sánchez-Martínez. 2015. A generalised alignment template formalism and its application to the inference of shallow-transfer machine translation rules from scarce bilingual corpora. *Computer Speech and Language*, 32(1):46–90.

Sánchez-Martínez, F. and M. L. Forcada. 2009. Inferring shallow-transfer machine translation rules from small parallel corpora. *Journal of Artificial Intelligence Research*, 34:605–635.

Demostraciones

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 191-194 recibido 26-03-2015 revisado 25-04-2015 aceptado 13-05-2015

A Web-based Text Simplification System for English*

Un Sistema de simplificación de textos on-line para el inglés

Daniel Ferrés Universitat Pompeu Fabra daniel.ferres@upf.edu
Montserrat Marimon Universitat Pompeu Fabra montserrat.marimon@upf.edu
Horacio Saggion Universitat Pompeu Fabra horacio.saggion@upf.edu

Resumen: La simplificación textual consiste en reducir la complejidad léxica y sintáctica de documentos con el fin de mejorar su legibilidad y comprensibilidad. En este trabajo se presenta una demostración de un sistema *on-line* de simplificación léxica y sintáctica de textos en inglés. Nuestro sistema es modular y adaptable, lo que lo hace adecuado para diversos tipos de usuarios.
Palabras clave: Simplificación léxica, simplificación sintáctica, *demo on-line*

Abstract: Text Simplification is the task of reducing the lexical and syntactic complexity of documents in order to improve their readability and understandability. This paper presents a web-based demonstration of a text simplification system that performs state-of-the-art lexical and syntactic simplification of English texts. The core simplification technology used for this demonstration is highly customizable making it suitable for different types of users.
Keywords: Lexical simplification, syntactic simplification, web-based demo

1 Introduction

Text Simplification (Carroll et al., 1998; Siddharthan, 2006) is the task of reducing the lexical and syntactic complexity of textual documents in order to improve their readability and understandability. This paper presents a demonstration of a text simplification system that performs (sequentially) state-of-the-art lexical and syntactic simplification of documents in English. The lexical simplifier has been developed following current robust, corpus-based approaches (Biran and Brody, 2011; Bott et al., 2012). The syntactic simplifier has been built following a linguistically motivated approach implemented as transformation rules complemented with text generation techniques. The chosen approach, which uses typed dependencies as basic representation, is based on current arguments in favor of the use of such representations in order to produce correct output (Siddharthan, 2006; Siddharthan and Angrosh, 2014). These simplifiers have been built entirely in the Java programming language, with open source software and freely available lexical resources. The system is highly configurable and adaptable and the resources used can easily be changed to meet the needs of different target groups.

2 Lexical Simplifier

Lexical simplification aims at replacing difficult words with easier synonyms, while preserving the meaning of the original text segments (Carroll et al., 1998). Our lexical simplifier combines Word Sense Disambiguation (WSD) and Lexical Simplicity measures to simplify words in context. It is composed of the following processing phases (executed sequentially): Document Analysis, Complex Words Detection, WSD, Synonyms Ranking, and Language Realization. The Document Analysis phase uses default components from the GATE system (Cunningham et al., 2002) to perform tokenization, sentence splitting, part-of-speech (PoS) tagging, lemmatization, Named Entity Recognition and Classification, and co-reference resolution. In addition, only during syntactic simplification (see below), the MATE Tools dependency parser (Bohnet, 2010) adds dependency labels to sentence tokens.

2.1 Complex Word Detection

Complex word detection is carried out to identify target words to be substituted. The procedure identifies a word as complex when

* This work was funded by the ABLE-TO-INCLUDE project (European Commission Competitiveness and Innovation Framework Programme under Grant Agreement No. 621055) and project SKATER-UPF-TALN (TIN2012- 38584-C06-03) from Ministerio de Economía y Competitividad, Secretaría de Estado de Investigación, Desarrollo e Innovación, Spain.

ISSN 1135-5948

the frequency count of the word in a given psycholinguistic database is in a range determined by two threshold values (i.e. w is complex if $min \leq w_{frequency} \leq max$). The following psycholinguistic resources can be used separately: Age-of-acquisition norms (Kuperman, Stadthagen-Gonzalez, and Brysbaert, 2012)[1], Kucera-Francis[2] (Kucera and Francis, 1967) frequency counts (extracted from the Brown Corpus). For example, words such as "hand" and "sun" have 470 and 123 counts respectively in the Kucera-Francis, whereas less common words such as "manifest" and "gastronomy" have 9 and 1 counts.

2.2 Word Sense Disambiguation

Since words can have more than one meaning and in order to select an appropriate word replacement out of a list of "synonyms", a Word Sense Disambiguation (WSD) phase is applied. The WSD algorithm used is based on the Vector Space Model (Turney and Pantel, 2010) approach for lexical semantics which has been previously used in Lexical Simplification (Biran and Brody, 2011). This algorithm uses a word vectors model derived from a large text collection from which a word vector for each word in WordNet-3.1[3] is created by collecting co-occurring word lemmas of the word in N-window contexts (only nouns, verbs, adjectives, and adverbs). Then, a common vector is computed for each of the word senses of a given target word (lemma and PoS). These word sense vectors are created by adding the vectors of all words (e.g. synonyms, hypernyms) in each sense. When a complex word is detected, the WSD algorithm computes the cosine distance between the context vector computed from the words of the complex word context (at sentence or document level) and the word vectors of each sense from the model. The word sense selected is the one with the lowest cosine distance between its word vector in the model and the context vector of the complex word in the sentence or document to simplify. Two data structures were produced following this procedure: 1) one that contains 81,242 target words and 135,769 entries, 2) another version that uses only synonyms to create the word sense vectors and has 63,649 target words and 87,792 entries.

The Simple Wikipedia was used to extract the word vectors model: the plain text of its 99,943 documents was extracted using the WikiExtractor[4] tool and Freeling 3.1 (Padró and Stanilovsky, 2012) was used to extract the lemmas and PoS tags of each word, from a 11-word window (5 words to each side of the target word).

2.3 Synonyms Ranking

The Synonyms Ranking phase tries to rank synonyms by their lexical simplicity and finds the simplest and most appropriate synonym word for the given context. The simplicity measures implemented are two: 1) only the word frequency (used by default for the simplifier) is used to rank synonyms (i.e. more frequent is simpler) (Carroll et al., 1998) and 2) a metric which combines word length and word frequency proposed by (Bott et al. 2012). Frequency lists from the following corpora can be used to rank by lexical simplicity in our system: British National Corpus (BNC), Google Web 1T Corpus most frequent words[5], Simple English Wikipedia, English Wikipedia, American National Corpus, SUBTLEX-US[6], SUBTLEX-UK[7], Kucera-Francis, and Age-of-Aquisition norms.

2.4 Language Realization

The Language Realization phase generates the correct inflected forms of the final selected synonym words. The SimpleNLG[8] (Gatt and Reiter, 2009) Java API is used to convert lemmas to their correct inflectional forms according to their context and PoS tag.

3 Syntactic Simplifier

Syntactic simplification aims at transforming long and complicated sentences into their more simpler equivalents. Similar to (Aluísio and Gasperin, 2010; Bott and Saggion, 2014), our Syntactic Simplifier is linguistically motivated. Linguistic phenomena that may complicate readability are identified and appropriate transformations to generate simpler paraphases are implemented. Our simplifier targets the following syntactic constructions: *Apposition*, *Relative Clauses*, *Coor-*

[1] http://crr.ugent.be/archives/806
[2] http://www.psych.rl.ac.uk/kf.wds
[3] http://wordnet.princeton.edu/
[4] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor
[5] http://norvig.com/ngrams/count_1w.txt
[6] http://expsy.ugent.be/subtlexus/
[7] http://crr.ugent.be/archives/1423
[8] http://code.google.com/p/simplenlg/

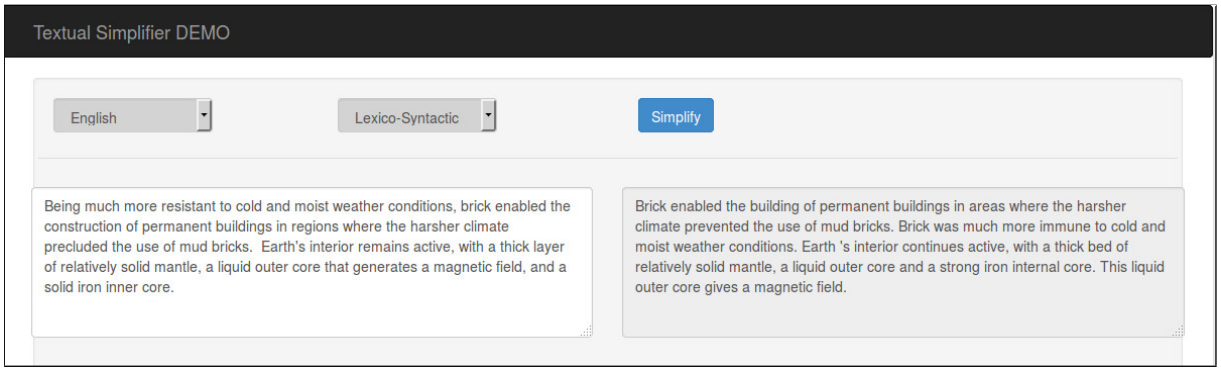

Figure 1: Screenshot of the web demonstration interface.

dination, *Coordinated Correlatives*, *Passive Constructions*, *Adverbial Clauses*, and *Subordinated Clauses*. After a process of document analysis which produces typed dependencies, the syntactic simplification is applied in two steps: 1) an analysis phase, that identifies the syntactic structures to be simplified and 2) a generation phase that produces correct simplified structures. The system recursively simplifies sentences until no more simplifications can be applied.

3.1 Syntactic Phenomena Identification and Analysis

Sentence analysis for simplification is implemented with GATE JAPE (Java Annotation Patterns Engine) grammars which detect and label the different kind of syntactic phenomena appearing in the sentences. For each of the above syntactic phenomena, a JAPE grammar contains several rules. Each rule contains a left-hand-side (LHS), which consists of an annotation pattern description, and the right-hand-side (RHS), which consists of annotation manipulation statements to produce rich simplification specific linguistic information useful for generation. These rules mainly rely on dependency information, which allows for a broad coverage of common syntactic phenomena. For example, the grammar for appositive phrases has a unique rule that identifies the apposition and its anchor using PoS and dependency labels. The LHS identifies the apposition's head by PoS and syntactic features (any common and proper noun and cardinal number which has the func(tion) appo(sition)). The RHS, first, finds out the head of the anchor (the token whose id unifies with the dependency of the apposition's head), and, then, it selects all the dependents of both the anchor's head and the apposition's head (and, recursively, the dependents of their dependents), and adds the annotations to the identified patterns. In addition, there are 17 rules for relative clauses (restrictive or non-restrictive). There are also 10 rules for coordination which deal with binary and three-conjunct coordination of sentences and VPs and 4 rules for coordinated correlatives which are distinguished by the endorsing item and the coordinator. Eight rules cover subordinated clauses expressing concession, cause, and time, both preceding and following the main clause, and 12 rules that deal with single adverbial clauses and up to three coordinated adverbial clauses, also preceding and following the modified clause. Finally, 14 rules cover passive constructions.

3.2 Sentence Generation

The generation phase uses the information provided by the analysis stage (i.e. precise annotations) to generate simple sentences. It applies a set of annotation manipulations which are specific for each phenomenon identified during analysis. These rules perform the common simplification operations, namely sentence splitting, reordering sentences, creation of new phrases, verbal tense adaption, personal pronouns transformation, capitalization and de-capitalization of some words, and word substitution.

4 Evaluation

We performed manual evaluation carried out by eight human judges, using the evaluation set used by Siddharthan and Angrosh (2014) from which we randomly selected 25 sentences. The judges assessed our system

w.r.t. fluency, adequacy, and simplicity, with a 5 point rating scale and assigned a mean score of 3.98, 4.02, and 2.86, respectivelly.

5 Web Demonstration

A web demonstration of the system can be accessed and tested in the following web address: *193.145.50.158/simplifier*. The web interface is shown in Figure 1 with an example of 2 complex sentences being simplified. The visual interface has two textual areas: one of them enabled for entering the text to be simplifier (with a white color background) and another one active only to see the output of the textual simplifier (with a grey color background). There are two selection forms that allow to change the language of simplification (currently only English) and select the type of simplification. The following types of simplifications are allowed: 1) Lexical, 2) Syntactic, and 3) Lexico-Syntactic. An execution button (with the "Simplify" label) performs the delivery of the parameters and the textual input to a back-end that performs the simplification. In the example in Figure 1 the lexical simplifier replaced words such as "construction" by "building", "inner" by "interior", etc. The syntactic simplifier transformed sentences containing subordinate and relative clauses into simpler paraphrases. The back-end of the demonstration has the following configuration options selected for the lexical simplifier: 1) the complex word detector uses the *Age Of Acquisition* norms that are complex for an age of acquisition of 7 years-old or less, 2) the WSD phase uses word vectors derived from contexts of the Simple Wikipedia and a dictionary of target words and senses derived from Wordnet 3.1 (version with only synonyms), 3) the Synonyms Ranker uses frequencies extracted from the BNC corpus.

References

Aluísio, S.M. and C. Gasperin. 2010. Fostering digital inclusion and accessibility: The porsimples project for simplification of portuguese texts. In *Proceedings of NAACL HLT 2010 YIWCALA*.

Biran, O. and N. Brody, S.and Elhadad. 2011. Putting it simply: A context-aware approach to lexical simplification. In *Proceedings of NAACL HLT 2011*.

Bohnet, B. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of COLING 2010*.

Bott, S., L. Rello, B. Drndarevic, and H. Saggion. 2012. Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish. In *Proceedings of COLING 2012*.

Bott, Stefan and Horacio Saggion. 2014. Text simplification resources for Spanish. *Language Resources and Evaluation*, 48(1):93–120.

Carroll, J., G. Minnen, Y. Canning, S. Devlin, and J. Tait. 1998. Practical simplification of English newspaper text to assist aphasic readers. In *Proceedings of the AAAI98 Workshop on Integrating AI and Assistive Technology*.

Cunningham, H., D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of ACL 2002*.

Gatt, A. and E. Reiter. 2009. SimpleNLG: A Realisation Engine for Practical Applications. In *Proceedings of ENLG 2009*.

Kucera, H. and W. N. Francis. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence, RI.

Kuperman, V., H. Stadthagen-Gonzalez, and M. Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4):978–990.

Padró, L. and E. Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of LREC 2012*.

Siddharthan, A. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.

Siddharthan, A. and M. Angrosh. 2014. Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In *Proceedings of EACL 2014*.

Turney, P. D. and P. Pantel. 2010. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188.

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 195-198 recibido 27-03-2015 revisado 27-04-2015 aceptado 14-05-2015

ElectionMap: una representación geolocalizada de intenciones de voto hacia partidos políticos sobre la base de comentarios de usuarios de Twitter

ElectionMap: a geolocalized representation of voting intentions to political parties based on twitter's user comments

Francisco Agulló, Antonio Guillén, Yoan Gutiérrez, Patricio Martínez-Barco
Carretera San Vicente del Raspeig S/N, 03690 San Vicente del Raspeig
Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante
{fagullo,aguillen,ygutierrez,patricio}@dlsi.ua.es

Resumen: ElectionMap es una aplicación web que realiza un seguimiento a los comentarios publicados en *Twitter* en relación a entidades que refieren a partidos políticos. Las opiniones de los usuarios sobre estas entidades son clasificadas según su valoración y posteriormente representadas en un mapa geográfico para conocer la aceptación social sobre agrupaciones políticas en las distintas regiones de la geografía española.
Palabras clave: Twitter, análisis de sentimientos, representación en mapas

Abstract: ElectionMap is a web application that follows, in *Twitter*, entities previously established and related to the politics. The user's opinions about the entities are classified according to its valuation by using sentiment analysis processes. Afterwards the opinions are represented in a geographic map that allows to know the social acceptance of spanish political parties in different geographical areas.
Keywords: Twitter, sentiment analysis, map representation

1 Introducción

En la actualidad las redes sociales sen han convertido en uno de los principales medios de distribución de información. La cantidad de información que se genera es tan grande que tanto empresas como gobiernos han comenzado a guiar sus campañas de promoción fijándose en las opiniones que los usuarios de las redes sociales valoran en sus perfiles. Una de las redes más utilizadas en la actualidad es *Twitter*, y la cantidad de información que proporciona esta red social no ha pasado desapercibida para las grandes empresas y gobiernos.

Otro de los problemas de las campañas de promoción es que no tienen el mismo impacto en toda la geografía. Saber en qué regiones una campaña obtiene una mayor aceptación social.

Para cubrir estas necesidades surge *ElectionMap*. Esta herramienta muestra de forma gráfica la opinión de los usuarios de *Twitter* sobre temas relacionados con la política. Como se describe en la sección 3, no todas las opiniones pueden tenerse en cuenta, sólo sirven las opiniones positivas y localizadas, por lo que el conjunto de datos utilizado no es el 100 % de la información recogida, si no que una muestra representativa del total. Finalmente tras procesar este conjunto de datos se muestra una representación gráfica de intención de voto político agrupada por áreas geográficas representando así el apoyo de los usuarios de los medios sociales. La aplicación web se divide en dos componentes, el núcleo que se explica en la sección 2 y la parte web que se detalla en la sección 3. Por último en la sección 4 se puede ver el procedimiento utilizado para la detección de opiniones.

2 Núcleo

El núcleo es el encargado de recopilar la información de los tweets en los cuales se mencionan los partidos a evaluar. Para ello, se define un conjunto de términos para cada una de las entidades que se van a representar. Cada vez que un usuario de *Twitter* escribe uno de los términos que identifican a los partidos a evaluar, se lanza una alerta hacia nuestro sistema notificando el nuevo comentario y se almacena la información relevante al mensa-

ISSN 1135-5948

je. Además de la información que *Twitter* nos proporciona, para cada uno de los tweets, se realiza un procesamiento del tweet para conocer si el texto del mensaje es positivo, negativo o neutral y se almacena para considerarlo al igual que el resto de la información.

Por otro lado, el núcleo también se encarga de filtrar la información obtenida de *Twitter* ya que se reciben muchos datos pero no todos son útiles para mostrarse geoespacialmente ya sea porque el usuario no tiene activada la geolocalización o porque en su perfil no detalla su origen. Además al ser un mapa para valorar la intención de voto existen una serie de restricciones que vienen impuestas por el objetivo:

- **Los votantes sólo suman, no restan**. Los usuarios pueden expresar tanto su conformidad como su disconformidad con los partidos políticos, ya que cuando un votante acude las urnas vota en positivo, no en negativo, todos los mensajes con valoraciones negativas son ignorados y no quedan representados en el mapa.
- **Un usuario un voto**. Un usuario puede opinar bien sobre más de un partido político, pero en unas elecciones sólo podría formalizar el voto sobre una de las formaciones políticas, por lo que se asume que un usuario votará a la formación que mayor valoración haya obtenido en el total de sus mensajes publicados en la red social, lo cual indicaría que este usuario tiene mayor inclinación sobre ese partido al que ha comentado positivamente en más ocasiones.
- **Localización del tweet**. Los datos se muestran en un mapa de España, por lo que es necesario saber a que localidad le corresponde el voto de cada usuario. Para ello en primer lugar se busca en la descripción de usuario, si ésta tiene una localización válida, se asume que esa es la localización real del usuario. En caso de no obtener ninguna localización en la búsqueda, se comprueba si el usuario tiene activada la geolocalización. En caso de tener algún tweet con localización almacenado, se tiene en cuenta la localización que más veces se ha utilizado, y si no hay ningún tweet geolocalizado, se ignoran los mensajes del usuario ya que no se podrían situar en el mapa.
- **Formaciones con restricciones**. Por último algunas de las formaciones políticas sólo se presentan en ciertas regiones, pero en *Twitter* puede haber usuarios que valore a estos partidos desde regiones en las que no sería posible su votación. Para evitar contabilizar estas opiniones, se permite añadir filtros en regiones, de forma que una formación puede estar filtrada a una o más regiones. Además un filtro por región se hereda a las divisiones territoriales de nivel inferior, es decir, un filtro de comunidad se extenderá a todas las provincias de la comunidad y éste a su vez a todas las ciudades de las provincias.

3 *Aplicación web*

La aplicación web es la parte visual de la herramienta la cual se encuentra disponible públicamente en la siguiente dirección web: http://gplsi.dlsi.ua.es/demos/electionmap. Para la representación de los datos se ha utilizado el API de Google Maps[1] al que se han añadido una serie de controles propios. En la figura 1 podemos ver las tres secciones de la aplicación que a continuación se van a detallar.

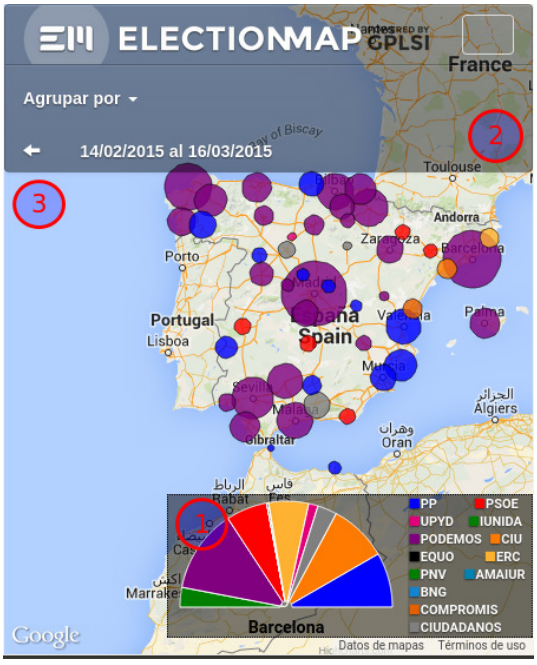

Figura 1: Vista de la aplicación.

- **Sección 1**. En este control se puede observar un gráfico (ordenado igual que

[1]https://developers.google.com/maps/?hl=es

el parlamento) con la intención de voto para cada partido en una determinada área geográfica. El gráfico se actualiza automáticamente cuando se selecciona alguna de las burbujas de la *Sección 3*.

- **Sección 2**. Menú de la aplicación, en el podemos editar tanto el modo de visualización como el periodo de tiempo que queremos ver representado en el mapa. Ya que se trata de un mapa español, podemos distinguir 4 modos para visualizar los datos:
  - Agrupados por ciudades.
  - Agrupados por provincias.
  - Agrupados por comunidades.
  - Agrupados por país.
- **Sección 3**. El último de los controles es en el que se muestran las burbujas. Una burbuja representa el color de la formación política que más valoraciones positivas tiene en una área geográfica. Además el tamaño de cada una de las burbujas es proporcional al porcentaje de usuarios con intención de voto hacia esa formación política obtenidos así como al número de habitantes del área representada. Por último si se quiere dejar una burbuja fijada en el gráfico de la *Sección 1* basta con hacer *click* sobre dicha burbuja y la información quedará fijada por un intervalo de 10 segundos. Si se desea desfijar antes de que este intervalo finalice se puede hacer *click* en cualquier zona del mapa.

4 Análisis de opiniones

Como se ha comentado anteriormente, la aplicación realiza un proceso de recogida y de almacenamiento de datos, pero entre estos dos procesos se realiza un procesamiento intermedio en el cual se evalúa si un texto expresa apoyo o rechazo hacia las entidades (ii.e. partidos politicos) a valorar. El método de evaluación utilizado es el comentado en "*The OpAL System at NTCIR 8 MOAT*" (Balahur et al., 2010a). Este método ya se demostró con éxito cuando se aplicó a otras tareas de la minería de opiniones como en el caso del Opinion Question Answering (Balahur et al., 2010b). En los artículos anteriormente citados se puede encontrar una explicación detallada sobre los léxicos que se combinan así como el uso de modificadores léxicos y elementos de potenciación de características utilizados.

El método se basa en un conjunto de lexicones que contienen palabras que expresan opiniones positivas o negativas agrupados según su intensidad. Cuando se evalúa un texto, se buscan las palabras contenidas en los lexicones y se les asigna una valoración base según el lexicón en el que se encuentren. El método además utiliza *stemmer* para valorar familias de palabras que pudieran estar relacionadas con las distintas palabras de los lexicones. Las palabras que no aparecen en ninguno de los lexicones se consideran neutrales. Este valor puede ser modificado posteriormente según los potenciadores y los modificadores que se encuentren en la frase. *OpAL* tiene en cuenta distintos tipos de modificadores (ii.e muy, mucho, poco, menos, no, ningún, etc) que inciden en el cambio o modificación de la polaridad de las distintas palabras con carga sentimental implicadas en los mensajes procesados. Finalmente se suman las valoraciones de cada una de las palabras y el valor resultante es el que se utiliza para etiquetar el texto como positivo (valoraciones mayores que 0), neutral (valoraciones iguales a 0) o negativo (valoraciones inferiores a 0).

5 Agradecimientos

ElectionMap es una aplicación web desarrollada por el Grupo de Procesamiento del Lenguaje Natural y Sistemas de Información (GPLSI)[2] de la Universidad de Alicante[3]. Esta aplicación ha sido parcialmente financiada por el Gobierno Español y la Comisión Europea a través de los proyectos: ATTOS (TIN2012-38536-C03-03), LEGOLANG (TIN2012-31224), SAM (FP7- 611312) y FIRST (FP7-287607) y por la Universidad de Alicante a través del proyecto emergente "Explotación y tratamiento de la información disponible en Internet para la anotación y generación de textos adaptados al usuario" (GRE13-15).

Bibliografía

Balahur, A., E. Boldrini, A. Montoyo, y P. Martínez-Barco. 2010a. The opal system at ntcir 8 moat. En *Proceedings of*

[2] http://gplsi.dlsi.ua.es/
[3] http://www.ua.es/

NTCIR-8 Workshop Meeting, Tokyo, Japan, páginas 241–245.

Balahur, A., E. Boldrini, A. Montoyo, y P. Martínez-Barco. 2010b. Opinion question answering: Towards a unified approach. En *ECAI*, páginas 511–516.

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 199-202 recibido 27-03-2015 revisado 27-04-2015 aceptado 15-05-2015

Social Rankings: análisis visual de sentimientos en redes sociales

Social Rankings: Visual Sentiment Analysis in Social Networks

Javi Fernández, Yoan Gutiérrez, José M. Gómez, Patricio Martínez-Barco
Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante
Carretera San Vicente del Raspeig S/N, 03690 San Vicente del Raspeig
{javifm,ygutierrez,jmgomez,patricio}@dlsi.ua.es

Resumen: *Social Rankings* es una aplicación web que realiza un seguimiento en tiempo real de entidades en las redes sociales. Detecta y analiza las opiniones sobre estas entidades utilizando técnicas de análisis de sentimientos para generar un informe visual de su valoración y su evolución en el tiempo.
Palabras clave: análisis de sentimientos, minería de opiniones, redes sociales

Abstract: *Social Rankings* is a web application that follows different entities in the social networks in real time. It detects and analyses the opinions about these entities using sentiment analysis techniques, to generate a visual report of their reputation and evolution in time.
Keywords: sentiment analysis, opinion mining, social networks

1 Introducción

En la actualidad millones de personas expresan cada día sus opiniones públicamente a través de las redes sociales. Empresas y organizaciones analizan manualmente esta información subjetiva para realizar estudios y obtener una visión global de la valoración de diferentes marcas, productos o personajes públicos. Pero la cantidad de información disponible es tan grande que es necesario crear herramientas que realicen esta tarea de manera automática y en tiempo real. Este es el objetivo de la aplicación web *Social Rankings*.

Social Rankings es una aplicación web que realiza un seguimiento en tiempo real de entidades en las redes sociales. Detecta y analiza las opiniones sobre estas entidades utilizando técnicas de análisis de sentimientos para generar un informe visual de su valoración y su evolución en el tiempo.

2 Configuración

El único paso a realizar por el usuario para poder utilizar *Social Rankings* es elegir las entidades a seguir. Pueden ser tanto personas como empresas, organizaciones, temas o, en general, cualquier conjunto de palabras de interés. Para iniciar su seguimiento, estas deben ser definidas por el usuario, indicando su nombre, descripción y palabras clave. A partir de ese momento, las redes sociales son rastreadas periódicamente, obteniendo todas las publicaciones que mencionen a estas entidades, es decir, que contengan sus palabras clave.

3 Funcionamiento

Una vez elegidas las entidades y sus palabras clave, comienza a realizarse el seguimiento en las redes sociales. En este momento realizamos el seguimiento en Twitter[1], por la facilidad de uso de sus APIs[2] y la disponibilidad de herramientas para su manipulación[3]. En el futuro planeamos añadir nuevas redes sociales, además de blogs y foros.

A medida que se van encontrando diferentes publicaciones mencionando a las entidades especificadas, un sistema de análisis de sentimientos detecta aquellas que contienen una opinión y las clasifica como positivas o negativas. En la actualidad utilizamos una combinación dos sistemas diferentes.

El primero está basado en una de las aproximaciones de (Balahur, 2011). Este sistema utiliza unos diccionarios de palabras que contienen sentimientos, obtenidos combinando otros recursos de opinión como *WordNet Affect* (Strapparava y Valitutti, 2004) o *SentiWordNet* (Esuli y Sebastiani, 2006). La polaridad de un texto dependerá de la suma total de los pesos de las palabras de opinión que

[1] http://twitter.com
[2] https://dev.twitter.com/streaming/overview
[3] http://twitter4j.org/

ISSN 1135-5948

contiene ese texto.

El segundo está basado en la aproximación híbrida de Fernández et. al (2013) y (2014). En esta aproximación se utilizan *skipgrams* como unidades de información (Guthrie et al., 2006). A partir de un corpus se genera un diccionario de skipgrams con pesos de positividad y negatividad asociados. Utilizando el corpus y este diccionario, se entrena un modelo de aprendizaje automático que aprende a combinar los pesos de los skipgrams para obtener la polaridad de nuevos textos. En este momento el corpus que hemos utilizado es el proporcionado en la edición de 2013 del workshop TASS[4] (Villena Román et al., 2013).

Estas aproximaciones han sido elegidas por sus buenos resultados en competiciones (Fernández et al., 2013) y su velocidad de procesamiento, apta para aplicaciones en tiempo real.

4 Valoración

Con el fin de generar un informe visual en el tiempo para cada entidad, necesitamos obtener una puntuación numérica que tenga en cuenta la cantidad de opiniones positivas y negativas detectadas, y la cantidad de gente a la que han llegado estas opiniones, en un periodo de tiempo concreto. A esta puntuación le hemos llamado *valoración* y se ha calculado utilizando la ecuación 1.

$$v_{e,t} = \frac{\sum_{p \in P_{e,t,+}} 2a_p + \sum_{p \in P_{e,t,0}} a_p - \sum_{p \in P_{e,t,-}} 3a_p}{d_t + \sum_{p \in P_{e,t}} 3a_p} \quad (1)$$

Donde $v_{e,t}$ es la valoración de la entidad e en el período de tiempo t; $P_{e,t}$ es el conjunto de todas las publicaciones que contienen una mención a la entidad e en el período de tiempo t (y p es una publicación dentro de ese conjunto); $P_{e,t,+}$, $P_{e,t,0}$ y $P_{e,t,-}$ son subconjuntos de $P_{e,t}$ cuyas publicaciones han sido clasificadas como positivas, neutrales o negativas respectivamente; a_p es la audiencia o número de usuarios a los que ha llegado la publicación p; y d_t es la duración en milisegundos del periodo t. Utilizando esta ecuación obtenemos un valor dentro del intervalo $[-1, +1]$, donde -1 sería la peor valoración y $+1$ sería la mejor valoración dada por el sistema.

Consideramos una mención neutral ($p \in P_{e,t,0}$) como algo positivo, ya que el hecho de que una entidad sea mencionada en las redes sociales aumenta su valoración. Por eso las menciones neutrales influyen de manera positiva en la fórmula. Las menciones positivas ($p \in P_{e,t,+}$) indican que no sólo se está mencionando a la entidad sino que se está diciendo algo bueno sobre ella. Por eso consideramos que deben influir en mayor medida que las menciones neutrales, aumentando su valor en un factor de 2. Finalmente, a las menciones negativas ($p \in P_{e,t,-}$) les aumentamos su valor en un factor de 3, ya que para los usuarios suele ser más interesante conocer los malos comentarios sobre una entidad (por ejemplo, para encontrar una solución). Estos parámetros han sido establecidos manualmente según lo descrito anteriormente.

La suma de la duración del período en el denominador de la ecuación (d_t) es una forma de dar una mayor valoración a las publicaciones que hayan llegado a más seguidores. Por ejemplo, si una entidad tiene menciones negativas que ha llegado a 100 personas en un minuto, su valoración sería de $-3 \cdot 100/(3 \cdot 100 + 60000) = -0{,}005$. Sin embargo, si las publicaciones ha llegado a 10.000 personas, la valoración sería de $-3 \cdot 10000/(3 \cdot 10000 + 60000) = -0{,}333$.

5 Informe visual

En la Figura 1 podemos ver un ejemplo de la interfaz visual de *Social Rankings*. Las entidades que comparamos en este ejemplo son los cuatro principales partidos políticos en España en este momento. Concretamente, se muestra el valor de reputación entre el 17 y el 24 de marzo de 2015.

Esta interfaz podemos dividirla en cuatro partes, que describiremos a continuación:

- *Selector de entidades.* Es un campo desplegable que contiene todas las entidades a las que se les está realizando el seguimiento. Se deben seleccionar una o varias entidades para obtener el informe visual. En la Figura 2 podemos ver el proceso de selección de entidades en funcionamiento.
- *Selector de vista.* Es otro campo desplegable mediante el cual se puede cambiar la vista actual. Las vistas disponibles son *número de tweets*, *número de tweets positivos*, *número de tweets negativos* y *valor de reputación*. En las primeras tres vistas simplemente se realiza un conteo

[4] http://www.daedalus.es/TASS2013

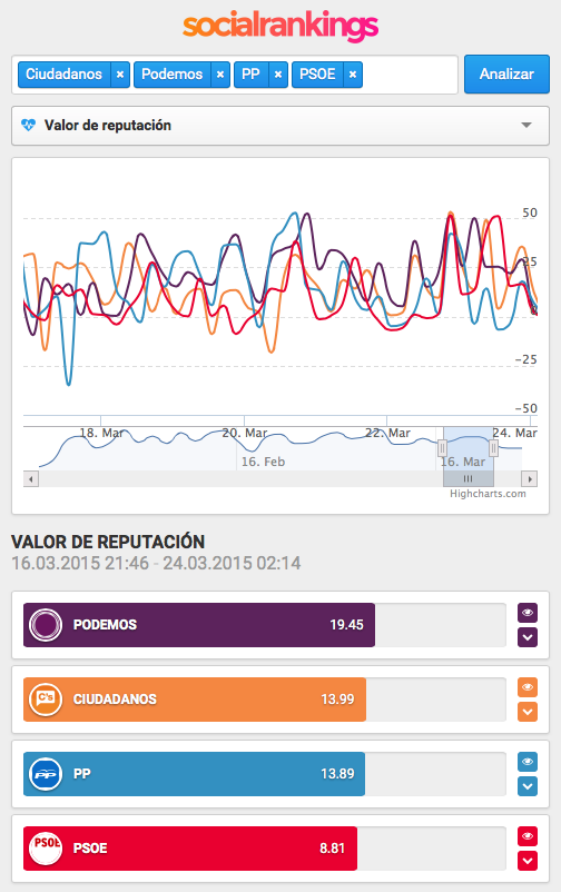

Figura 1: Interfaz visual de Social Rankings

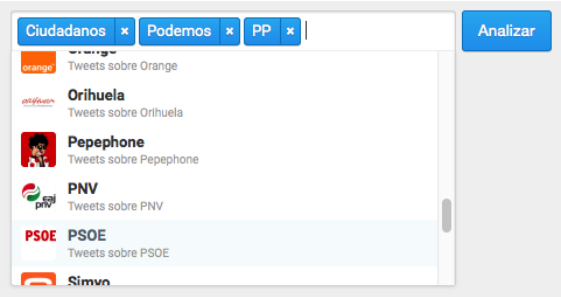

Figura 2: Selector de entidades

del número de publicaciones total, publicaciones positivas y publicaciones negativas respectivamente. La última vista muestra la valoración descrita en la sección 4. En la Figura 3 se puede observar el desplegable con las vistas disponibles.

- *Ranking de entidades.* En esta sección se muestra una lista de las entidades elegidas con un valor asociado para un rango de tiempo dado, que depende de la vista elegida. En el ejemplo de la Figura 4 la vista elegida es el número de tweets positivos. Para que sea más sencillo compa-

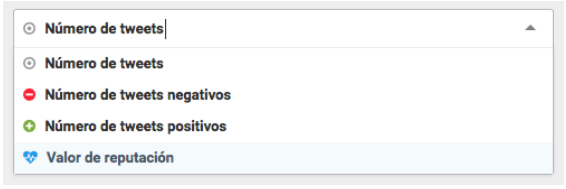

Figura 3: Selector de vista

rar los valores para las distintas entidades, se muestra una barra con el color de la entidad cuyo tamaño depende del valor asignado. Para elegir seleccionar un rango de fechas diferente, se utiliza la gráfica de evolución, que explicamos a continuación.

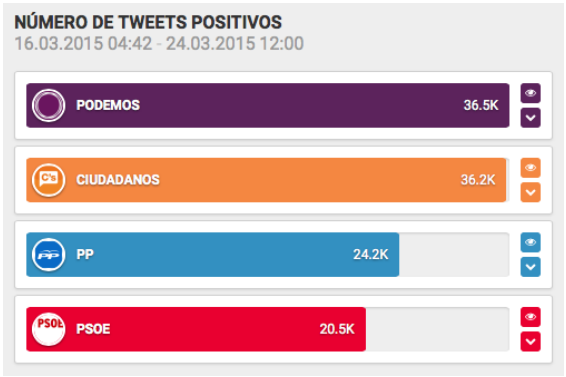

Figura 4: Ranking de entidades

- *Gráfica de evolución.* Es una gráfica en la que se muestra la evolución de los valores de la vista elegida para las entidades seleccionadas durante todo el periodo de seguimiento. Por defecto sólo se muestra el último día, pero es posible seleccionar un rango de fechas diferente a lo largo de toda la línea de evolución. Al cambiar el rango de fechas, el ranking de entidades explicado anteriormente también actualizará sus valores. En la Figura 5 podemos ver un ejemplo de evolución del valor de reputación de las entidades elegidas. Cada punto representa un periodo de tiempo concreto y, al pasar el ratón por encima, es posible ver los datos para ese periodo.

También se ha añadido una funcionalidad adicional experimental cuando se visualiza una única entidad. Podremos ver los términos y expresiones más utilizados en cada período de tiempo. De esta forma es posible obtener una visión general de lo que ha ocurrido y lo que más se ha comentado en un momento

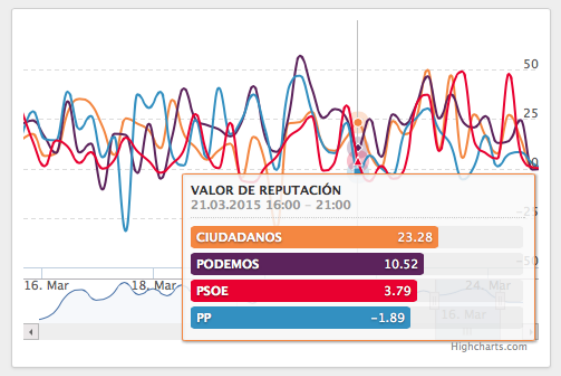

Figura 5: Gráfica de evolución

concreto. Esto se ha realizado mediante técnicas de clustering de texto utilizando la herramienta Carrot2[5]. En la Figura 6 se muestra un ejemplo.

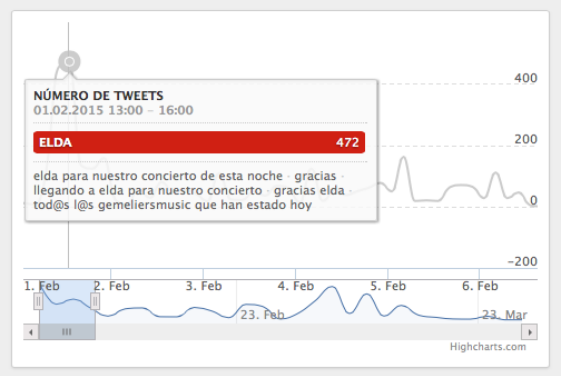

Figura 6: Términos y expresiones más relevantes

Esta herramienta se puede utilizar públicamente[6]. En esta versión pública sólo se visualizan los datos de las entidades predefinidas, no es posible añadir nuevas entidades. Si desea añadir alguna entidad contacte con los autores de este artículo.

6 Agradecimientos

Social Rankings ha sido desarrollada por el Grupo de Procesamiento del Lenguaje Natural y Sistemas de Información (GPLSI)[7] de la Universidad de Alicante[8]. Esta aplicación ha sido financiada parcialmente por el Gobierno Español a través de los proyectos *ATTOS* (TIN2012-38536-C03-03) y *LEGOLANG* (TIN2012-31224), la Comisión Europea a través del proyecto *SAM* (FP7-611312), la Generalitat Valenciana a través del proyecto *DIIM2.0* (PROMETEOII/2014/001) y la Universidad de Alicante a través del proyecto emergente *"Explotación y tratamiento de la información disponible en Internet para la anotación y generación de textos adaptados al usuario"* (GRE13-15).

[5]http://project.carrot2.org
[6]http://gplsi.dlsi.ua.es/demos/socialrankings
[7]http://gplsi.dlsi.ua.es/
[8]http://www.ua.es/

Bibliografía

Balahur, A. 2011. *Methods and resources for sentiment analysis in multilingual documents of different text types*. Universidad de Alicante.

Esuli, A. y F. Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. En *Proceedings of LREC*, volumen 6, páginas 417–422. Citeseer.

Fernández, J., Y. Gutiérrez, J. M. Gómez, y P. Martınez-Barco. 2014. A Supervised Approach for Sentiment Analysis using Skipgrams. En *Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC 2014)*.

Fernández, J., Y. Gutiérrez, J. M. Gómez, P. Martınez-Barco, A. Montoyo, y R. Munoz. 2013. Sentiment Analysis of Spanish Tweets Using a Ranking Algorithm and Skipgrams. En *XXIX Congreso de la Sociedad Espanola de Procesamiento de Lenguaje Natural (SEPLN 2013)*, páginas 133–142.

Guthrie, D., B. Allison, W. Liu, L. Guthrie, y Y. Wilks. 2006. A closer look at skip-gram modelling. En *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006)*, páginas 1–4.

Strapparava, C. y A. et al. Valitutti. 2004. WordNet Affect: an Affective Extension of WordNet. En *LREC*, volumen 4, páginas 1083–1086.

Villena Román, J., S. Lana Serrano, E. Martínez Cámara, y J. C. González Cristóbal. 2013. TASS-Workshop on Sentiment Analysis at SEPLN.

Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 203-206 recibido 27-03-2015 revisado 25-04-2015 aceptado 21-05-2015

Summarization and Information Extraction in your Tablet*

Resumen y extracción de información en tu Tablet

Francesco Barbieri
Universitat Pompeu Fabra
C/Tànger 122 - Barcelona
francesco.barbieri@upf.edu

Francesco Ronzano
Universitat Pompeu Fabra
C/Tànger 122 - Barcelona
francesco.ronzano@upf.edu

Horacio Saggion
Universitat Pompeu Fabra
C/Tànger 122 - Barcelona
horacio.saggion@upf.edu

Resumen: En este artículo describimos la demonstración de una serie de aplicaciones de resumen automático y extracción de informaciones integradas en una tableta. Se presentan funcionalidades para resumir las últimas noticias publicadas en la Web, extraer información sobre eventos concretos, y resumir textos en ingés y español ingresados por el usuario. La aplicación está disponible en un Web-browser y una tableta con sistema operativo Android.
Palabras clave: Resumen automático, extracción de informaciones, aplicaciones Web

Abstract: In this article we present a Web-based demonstration of on-line text summarization and information extraction technology. News summarization in Spanish has been implemented in a system that monitors a news provider and summarizes the latest published news. The possibility to generate summaries from user's provided text is also available for English and Spanish. The demonstrator also features event extraction functionalities since it identifies the relevant concepts that characterize several types of events by mining English textual contents.
Keywords: Text summarization, Information Extraction, Web-based Applications

1 Introduction

Two Natural Language Processing (NLP) technologies which can help people assess content relevance or quickly skim textual content are text summarization (Lloret and Palomar, 2012; Saggion and Poibeau, 2013) and information extraction (Piskorski and Yangarber, 2013). We have integrated our summarization and information extraction technology into the Web-based application whose architecture is outlined in Figure 1. Our application allows a user to monitor the latest published news by having access to different types of summaries automatically generated; it also features a functionality to extract information about specific events from textual sources such as Wikipedia articles. NLP demonstrators integrated in mobile applications or Web browsers can be particularly effective to introduce NLP related topics such as classification, summarization, and information extraction to students. In particular, the applications and demonstrator we are showcasing here constitute the core platform of the hands-on practice of a NLP-related course at Universitat Pompeu Fabra where students learn and adapt summarization technology to different languages and input documents. Besides, all the components demonstrated here are made freely available for research and teaching[1].

2 Tools

In order to develop our demonstrator we use available NLP tools: The SUMMA library which is an easy-to-customize summarization software distributed free of charge[2] for research purposes (Saggion, 2008) and the GATE system (Cunningham et al., 2002), a Java library and open source NLP development environment which is freely available. We also rely on the ROME[3] set of Really Simple Syndication and Atom Utilities for Java to access the latest news from one or more news providers.

* We acknowledge support from the Spanish research project SKATER-UPF-TALN TIN2012-38584-C06-03, the EU project Dr. Inventor FP7-ICT-2013.8.1 611383, and UPF projects PlaQUID 65 2013-2014 and PlaQUID 47 2011-2012.

[1] http://taln.upf.edu/content/resources/699
[2] http://www.taln.upf.edu/pages/summa.upf/
[3] http://rometools.github.io/rome/

ISSN 1135-5948

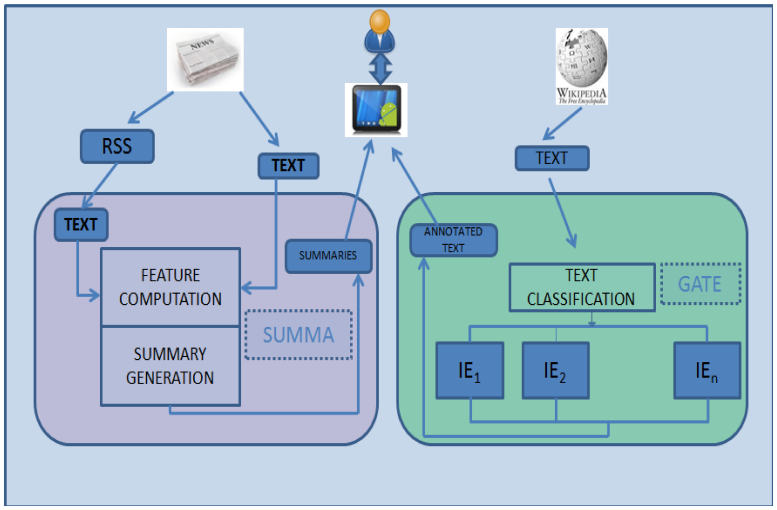

Figure 1: Overview of the Demo's Architecture

2.1 Summarization Application

Two summarization pipelines one for English and one for Spanish are implemented as GATE applications (i.e. gapp files) and integrated in the demonstrator. They are made up of the following SUMMA components (refer to Saggion (2014) for a description of these modules):

1. a term weighting computation algorithm based on term frequency and inverted document frequency (i.e. inverted document frequency table),
2. a vector computation module which represents each document sentence as a vector of terms and weigths,
3. a sentence position scorer,
4. a term frequency scorer based on term frequency * inverted document frequency,
5. a document vector computation module to represent the whole document as a vector,
6. a sentence-document similarity scorer,
7. a first-sentence similarity scorer, and
8. various sentence rankers to provide different summarization functionalities to the users.

Because the summarizers are based mainly on superficial and statistical features, they are *quasi* language-independent. The inverted document frequency table constitutes the sole language-specific component which is different between the two applications. Each application computes different types of summaries based on content relevance critera and at different compression rates. One type of summary just considers the relevance of sentences according to their position in the document, another type of summary considers the relevance of the sentence according to the distribution of words it contains. A third type of summary is based on scores given to sentences based on their similarity to a sentence-centroid computed out of all document's sentences.

2.2 Event Extraction Application

Based on the availability of the annotated CONCISUS corpus of event summaries in Spanish and English, we have developed a multi-domain information extraction application which targets three domains: airplane accidents, train accidents, and earthquakes (Saggion and Szasz, 2012). The event extraction application identifies different concept types that characterize the events. For example, for an airplaine accident some target concepts are the date of the accident, the place of the accident, the airline, the flight number, etc. The extraction procedure is implemented as follows: after a linguistic analysis of the input document for which we use ANNIE (Cunningham et al., 2002), a text classification algorithm is applied to the document to identify the target domain. Then, given the identified domain, an appropriate domain-specific information extraction component is invoked. The whole process is implemented as a conditional pipeline to prop-

erly control the execution of each information extraction component. The text classification algorithm to identify one of the 3 target domains is a Support Vector Machines classifier (Joachims, 1998) trained over unigrams (lemmas). During development the classifier achieved perfect F1 (training on 83 documents and testing on 29 unseen documents). The demonstrator includes a classifier trained on the whole CONCISUS corpus of 112 articles.

Since each domain requires the extraction of different concepts, three information extraction systems are implemented using a Support Vector Machines token classification approach with a context window of size five. All IE systems use exactly the same features for concept identification computed by ANNIE: word identity, lemmas, gazetteer list information, named entities, and POS tags. Cross-validation performance of the system varies from domain to domain with an F1 of 0.63 for extracting 28 aviation accident concepts, an F1 of 0.61 for extracting 20 train accident concepts, and an F1 of 0.40 for the extraction of 30 earquake-related concepts. The number of concepts to be learnt, the small size of the corpus, and the skewed distribution of concepts in the corpus partially explain the results.

3 Deployment

The Web application has been built relying on widespread Web technologies and libraries. The core component is made up of its Server Module that enables clients to invoke the summarization and information extraction services by accessing on-line REST endpoints that deliver their output in JSON format[4]. The Server Module is implemented as Java Web Application that exposes RESTful Web Services by relying on Jersey, a popular open source Java framework[5]. Jersey implements the Java API for RESTful Web Services specifications[6], thus supporting a modular and versatile development of RESTful Web Services in Java. The following set of functionalities can be invoked by querying the application's REST endpoints:

- retrieve and summarize on-line RSS feeds, like the feeds exposed by on-line newspapers (see Figure 2 for the automatic summaries about a recent aviation accident);
- summarize a pasted text, like for instance a news article;
- extract useful information from a textual excerpt (i.e., the concepts associated to the identified news event) – see Figure 3 for information extracted from the recent GermanWigs airplane crash described in Wikipedia.

When a REST endpoint of the Server Module is queried, a properly configured instance of SUMMA is exploited in order to process one or more texts, thus generating the results that are serialized as JSON and sent back to the client. The Server Module implements a thread-based service of RSS feeds retrieval: this service is responsible for maintaining a in-memory copy of the contents of the RSS feeds that can be processed by the application. In particular the service periodically performs asynchronous downloads of the contents of one or more RSS feed URLs in order to refresh a local copy of such information.

The Client Module is implemented by relying on HTML and Javascript. In particular, the Javascript framework JQuery[7] is exploited to implement the client (browser) logic. The REST endpoints exposed by the Server Module are queried by AJAX calls issued by the Client Module in order to retrieve and process data in response to user interactions. Our application can be deployed on any Java Web Application Container; in our current deployment we use the version 7 of Apache Tomcat[8].

References

Cunningham, H., D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02).*

Joachims, T. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In Claire Nédellec and Céline Rouveirol, editors,

[4]http://json.org/
[5]https://jersey.java.net/
[6]JAX-RS, https://jax-rs-spec.java.net/
[7]https://jquery.com/
[8]http://tomcat.apache.org/

Figure 2: RSS Summarization Application

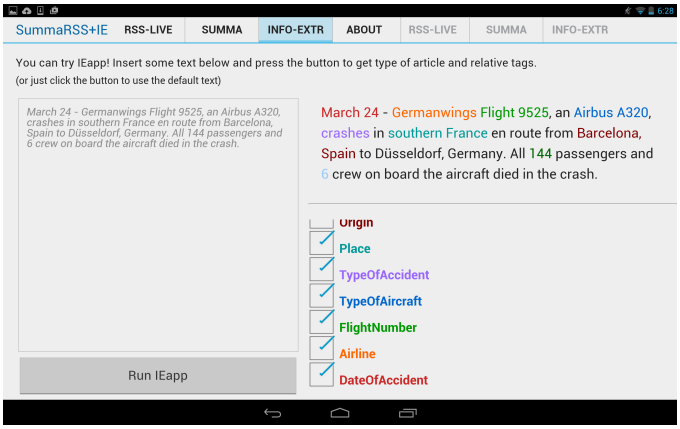

Figure 3: Event Identification in a Wikipedia Article

Proceedings of ECML-98, 10th European Conference on Machine Learning, number 1398 in Lecture Notes in Computer Science, pages 137–142, Chemnitz, Germany. Springer Verlag, Heidelberg.

Lloret, E. and M. Palomar. 2012. Text summarisation in progress: a literature review. *Artif. Intell. Rev.*, 37(1):1–41.

Piskorski, J. and R. Yangarber. 2013. Information extraction: Past, present, and future. In T. Poibeau, H. Saggion, J. Piskorski, and R. Yangarber, editors, *Multi-source, Multilingual Information Extraction and Summarization*, Theory and Applications of Natural Language Processing. Springer.

Saggion, H. 2008. SUMMA: A Robust and Adaptable Summarization Tool. *Traitement Automatique des Langues*, 49(2):103–125.

Saggion, H. 2014. Creating summarization systems with SUMMA. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 4157–4163.

Saggion, H. and T. Poibeau. 2013. Automatic text summarization: Past, present, and future. In T. Poibeau, H. Saggion, J. Piskorski, and R. Yangarber, editors, *Multi-source, Multilingual Information Extraction and Summarization*, Theory and Applications of Natural Language Processing. Springer.

Saggion, H. and S. Szasz. 2012. The CONCISUS corpus of event summaries. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012*, pages 2031–2037.

Información General

Información para los Autores

Formato de los Trabajos

- La longitud máxima admitida para las contribuciones será de 8 páginas DIN A4 (210 x 297 mm.), incluidas referencias y figuras.
- Los artículos pueden estar escritos en inglés o español. El título, resumen y palabras clave deben escribirse en ambas lenguas.
- El formato será en Word o 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*: 25 €* (residentes en España) o 30 *€* (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)
Ramón López-Cózar Delgado	Universidad de Granada (España)
Bernardo Magnini	Fondazione Bruno Kessler (Italia)

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/

Proyectos

Demostraciones

Información General