Sociedad Española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT 2015

ISSN: 1135-5948

Artículos

Tesis

Información General

© 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

Sociedad Española para el Procesamiento del Lenguaje Natural

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

ISSN: 1135-5948

Comité Editorial

Consejo de redacción

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

ISSN: 1135-5948
ISSN electrónico: 1989-7553
Depósito Legal: B:3941-91
Editado en: Universidad de Jaén
Año de edición: 2018
Editores: Mariona Taulé Delor, Universidad de Barcelona, mtaule@ub.edu
M. Teresa Martín Valdivia, Universidad de Jaén, maite@ujaen.es
Eugenio Martínez Cámara, Universidad de Granada, emcamara@decsai.ugr.es
Publicado por: Sociedad Española para el Procesamiento del Lenguaje Natural
Departamento de Informática. Universidad de Jaén
Campus Las Lagunillas, EdificioA3. Despacho 127. 23071 Jaén
secretaria.sepln@ujaen.es

Consejo asesor

Manuel de Buenaga	Universidad Europea de Madrid (España)
Sylviane Cardey-Greenfield	Centre de recherche en linguistique et traitement automatique des langues (Francia)
Irene Castellón Masalles	Universidad de Barcelona (España)
Arantza Díaz de Ilarraza	Universidad del País Vasco (España)
Antonio Ferrández Rodríguez	Universidad de Alicante (España)
Alexander Gelbukh	Instituto Politécnico Nacional (México)
Koldo Gojenola Galletebeitia	Universidad del País Vasco (España)
Xavier Gómez Guinovart	Universidad de Vigo (España)
Ramón López-Cozar Delgado	Universidad de Granada (España)
José Miguel Goñi Menoyo	Universidad Politécnica de Madrid (España)
Bernardo Magnini	Fondazione Bruno Kessler (Italia)
Nuno J. Mamede	Instituto de Engenharia de Sistemas e Computadores (Portugal)
M. Antònia Martí Antonín	Universidad de Barcelona (España)
M. Teresa Martín Valdivia	Universidad de Jaén (España)
Patricio Martínez-Barco	Universidad de Alicante (España)
Paloma Martínez Fernández	Universidad Carlos III (España)
Raquel Martínez Unanue	Universidad Nacional de Educación a Distancia (España)

© 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

Ruslan Mitkov	University of Wolverhampton (Reino Unido)
Manuel Montes y Gómez	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)
Lluís Padró Cirera	Universidad Politécnica de Cataluña (España)
Manuel Palomar Sanz	Universidad de Alicante (España)
Ferrán Pla Santamaría	Universidad Politécnica de Valencia (España)
German Rigau Claramunt	Universidad del País Vasco (España)
Horacio Rodríguez Hontoria	Universidad Politécnica de Cataluña (España)
Paolo Rosso	Universidad Politécnica de Valencia (España)
Leonel Ruiz Miyares	Centro de Lingüística Aplicada de Santiago de Cuba (Cuba)
Emilio Sanchís Arnal	Universidad Politécnica de Valencia (España)
Kepa Sarasola Gabiola	Universidad del País Vasco (España)
Encarna Segarra Soriano	Universidad Politécnica de Valencia (España)
Thamar Solorio	University of Houston (Estados Unidos de América)
Maite Taboada	Simon Fraser University (Canadá)
Mariona Taulé Delor	Universidad de Barcelona (España)
José Antonio Troyano Jiménez	Universidad de Sevilla (España)
L. Alfonso Ureña López	Universidad de Jaén (España)
Rafael Valencia García	Universidad de Murcia (España)
Felisa Verdejo Maíllo	Universidad Nacional de Educación a Distancia (España)
Manuel Vilares Ferro	Universidad de la Coruña (España)
Luis Villaseñor-Pineda	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)

Revisores adicionales

Alberto Díaz	Universidad Complutense de Madrid (España)
Salud María Jiménez Zafra	Universidad de Jaén (España)

© 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

Sociedad Española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA 2015

ISSN: 1135-5948

Preámbulo

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

Las áreas temáticas tratadas son las siguientes:

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

El ejemplar número 60 de la revista *Procesamiento del Lenguaje Natural* contiene trabajos correspondientes a dos apartados diferentes: comunicaciones científicas y resúmenes de tesis. Todos ellos han sido aceptados mediante el proceso de revisión tradicional en la revista.

© 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

Queremos agradecer a los miembros del Comité asesor y a los revisores adicionales la labor que han realizado.

Se recibieron 19 trabajos para este número, de los cuales 16 eran artículos científicos y 3 correspondían a resúmenes de tesis. De entre los 16 artículos recibidos, 7 han sido finalmente seleccionados para su publicación, lo cual fija una tasa de aceptación del 43,75%.

El Comité asesor de la revista se ha hecho cargo de la revisión de los trabajos. Este proceso de revisión es de doble anonimato: se mantiene oculta la identidad de los autores que son evaluados y de los revisores que realizan las evaluaciones. En un primer paso, cada artículo ha sido examinado de manera ciega o anónima por tres revisores. En un segundo paso, para aquellos artículos que tenían una divergencia mínima de tres puntos (sobre siete) en sus puntuaciones, sus tres revisores han reconsiderado su evaluación en conjunto. Finalmente, la evaluación de aquellos artículos que estaban en posición muy cercana a la frontera de aceptación ha sido supervisada por más miembros del comité editorial. El criterio de corte adoptado ha sido la media de las tres calificaciones, siempre y cuando hayan sido iguales o superiores a 5 sobre 7.

Marzo de 2018
Los editores

© 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

Sociedad Española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA 2015

ISSN: 1135-5948

Preamble

The *Natural Language Processing* journal aims to be a forum for the publication of high-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 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 60th issue of the *Procesamiento del Lenguaje Natural* journal contains scientific papers and doctoral dissertation summaries. All of these were accepted by a peer review process. We would like to thank the Advisory Committee members and additional reviewers for their work.

© 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

Nineteen papers were submitted for this issue, from which sixteen were scientific papers and three dissertation summaries. From these sixteen papers, we selected seven (43.75%) for publication.

The Advisory Committee of the journal has reviewed the papers in a double-blind process. Under double-blind review the identity of the reviewers and the authors are hidden from each other. In the first step, each paper was reviewed blindly by three reviewers. In the second step, the three reviewers have given a second overall evaluation of those papers with a difference of three or more points out of seven 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 criterion adopted was the mean of the three scores given.

March 2018
Editorial board

© 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

Sociedad Española para el Procesamiento del Lenguaje Natural

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

ISSN: 1135-5948

Artículos

Tesis

Información General

© 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

Artículos

Procesamiento del Lenguaje Natural, Revista nº 60, marzo de 2018, pp. 13-20 recibido 03-11-2017 revisado 05-01-2018 aceptado 09-02-2018

Ensembles for clinical entity extraction

Agrupaciones para la extracción de entidades clínicas

Rebecka Weegar[1], **Alicia Pérez**[2], **Hercules Dalianis**[1]
, **Koldo Gojenola**[2], **Arantza Casillas**[2], **Maite Oronoz**[2]
[1]Clinical Text Mining group; DSV; Stockholm University
[2]IXA (`http://ixa.eus`); Euskal Herriko Unibertsitatea (UPV-EHU)
Corresponding author: `rebeckaw@dsv.su.se`

Abstract: Health records are a valuable source of clinical knowledge and Natural Language Processing techniques have previously been applied to the text in health records for a number of applications. Often, a first step in clinical text processing is clinical entity recognition; identifying, for example, drugs, disorders, and body parts in clinical text. However, most of this work has focused on records in English. Therefore, this work aims to improve clinical entity recognition for languages other than English by comparing the same methods on two different languages, specifically by employing ensemble methods. Models were created for Spanish and Swedish health records using SVM, Perceptron, and CRF and four different feature sets, including unsupervised features. Finally, the models were combined in ensembles. Weighted voting was applied according to the models individual F-scores. In conclusion, the ensembles improved the overall performance for Spanish and the precision for Swedish.
Keywords: Clinical entity recognition, ensembles, Swedish, Spanish

Resumen: Los informes médicos son una valiosa fuente de conocimiento clínico. Las técnicas de Procesamiento del Lenguaje Natural han sido aplicadas al procesamiento de informes médicos para diversas aplicaciones. Generalmente un primer paso es la detección de entidades médicas: identificar medicamentos, enfermedades y partes del cuerpo. Sin embargo, la mayoría de los trabajos se han desarrollado para informes en Inglés. El objetivo de este trabajo es mejorar el reconocimiento de entidades médicas para otras lenguas diferentes a Inglés, comparando los mismos métodos en dos lenguas y utilizando agrupaciones de modelos. Los modelos han sido creados para informes médicos en Español y Sueco utilizando SVM, Perceptron, CRF y cuatro conjuntos diferentes de atributos, incluyendo atributos no supervisados. Para el modelo combinado se ha aplicado votación ponderada teniendo en cuenta la F-measure individual. En conclusión, el modelo combinado mejora el rendimiento general y para posibles mejoras debemos investigar métodos más sofisticados de agrupación.
Palabras clave: Reconocimiento de entidades médicas, agrupaciones, sueco, castellano

ISSN 1135-5948. DOI 10.26342/2018-60-1 © 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

1 Introduction

Natural language processing has been applied to health records for tasks as diverse as detecting adverse drug reactions (Henriksson et al., 2015), surveillance of nosocomial infections (Haas et al., 2005) and for assigning ICD codes to health records (Crammer et al., 2007). To many of the tasks utilizing natural language processing on health records, a well-functioning named entity recognition module is central (Demner-Fushman, Chapman, and McDonald, 2009).

There are European and also national projects that focus on the automatic extraction of valuable information from patient records. Three on-going projects are: firstly, CrowdHEALTH, a European project that attempts at gathering and processing multi-modal data from member states, conform ethical regulations, and exchange important information; secondly, the Spanish Ministry has involved a multi-disciplinary team to tackle natural language processing in the clinical domain among others in the so called "Plan de impulso de las tecnologías del lenguaje"; a third example is the Nordic Center of Excellence in Health-Related e-Sciences (NIASC) which is funded by NORDFORSK, the Nordic council of ministers, with one aim to detect early symptoms of cancer in patient records. Being so different from one another, the aforementioned three projects include, to different extents, the detection of key entities. While CrowdHEALTH shall incorporate languages from European states, English is still the dominating language in research articles in the clinical domain. Moreover, patient records is a type of data seldom explored due to confidentiality issues.

Motivated by this gap and shared interest, the Clinical Text Mining group at Stockholm University and the IXA research group at the University of the Basque Country cooperate with the aim to extract information from patient records and build robust methods for languages other than English.

The goal of this work is to extract medical entities from Electronic Health Records (EHRs) focusing on patient records in Swedish and Spanish, from Karolinska University Hospital and Galdakao-Usansolo Hospital respectively.

Swedish is a Germanic language with about 10 million speakers. A challenge for processing Swedish, as well as other Germanic languages, is that compounds are very common. For Swedish, a rich variety of noun compounds are possible, an example is the word *huvudvärkstablett* (*huvud*-head, *värk*-ache, *s*-, *tablett*-tablet). Spanish is a Romance language and about 360 million people has Spanish as their first language. Regarding the object of this paper, clinical entities, some examples of specific features of the Spanish language are given in (Reynoso et al., 2000). For instance, medical terms in English expressed by gerunds tend to take the form of subordinate clauses or prepositional phrases in Spanish. Some examples from SNOMED CT are as follows: *Conditions causing complications in pregnancy* that takes the form of a subordinate *"condiciones que causan [that cause] complicaciones en el embarazo"*; *dispatching and receiving clerk* takes the form of the prepositional phrase *"empleado de despacho y recepción de mercadería"*.

Text in patient records tend to pose characteristics that are not shared with other kind of texts (such as journal abstracts, social media etc.) which make them challenging to process. These characteristics include a rich vocabulary with many possible forms for the same concept and domain specific terminology, many abbreviations and acronyms which may be ambiguous, and few complete sentences. Besides, it has been found that up to 10% of tokens in health records are misspelled (Ruch, Baud, and Geissbühler, 2003; Lai et al., 2015; Ehrentraut et al., 2012).

Conditional Random Fields (CRFs) is a probabilistic model for labelling sequences of data (Lafferty, McCallum, and Pereira, 2001), which makes it suitable for named entity recognition. CRFs were previously applied in the clinical domain with good results (Skeppstedt et al., 2014). As with CRFs, Support Vector Machines (SVMs) have proven useful for entity recognition.

The works explored so far that made use of CRFs or SVMs to detect entities have the drawback of relying on vast discrete feature-spaces built up on the basis of n-grams of words. Named entity recognition has been recently shifted from symbolic representations (words, lemmas, POS, etc.) to dense representations.

In Tang et al. (2014) biomedical entity recognition was carried out on Biocreative II GM corpus making use of CRFs. With regard to the features, they used basic features

(stemmed words and POS), Brown clusters, distributional word representations and word embeddings extracted with word2vec.

Regarding entity recognition for Spanish, word representations (Turian, Ratinov, and Bengio, 2010) have been incorporated as external features to infer a CRF (Zea et al., 2016; Agerri and Rigau, 2016). To this end, the entity recognition system was inferred from in-domain annotated data, however, large out-domain unannotated data were used to infer continuous word-representations. This strategy can yield results comparable to those obtained with approaches based on deep learning strategies (Zea et al., 2016), possibly due to the semantic relatedness associated to continuous spaces that lead to generalization (Faruqui and Padó, 2010).

In addition to more robust feature representations, ensembles of classifiers have previously been shown capable of improving Entity Recognition. Florian et al. (2003) applied Named Entity Recognition to English and German texts using an ensemble of four different classifiers achieving improved results. Saha and Ekbal (2013) created an ensemble of seven base-learners, including CRF and SVM, and performed Named Entity Recognition on Hindi, Bengali and Telugu. The performance of the ensemble of classifiers using weighted voting was better than that of any of the individual classifier and the weights were determined using genetic algorithms. Ensembles of classifiers have also been used on clinical texts, Kang et al. (2012) combined seven existing system for clinical entity recognition for English texts. A threshold – the number of systems needed to agree on an entity to include it – was decided by evaluating the systems on the training set. An ensemble of systems was found to give a higher performance than any of the individual systems

Exploring patient records is a challenging task. Moreover, given that this is a joint-project on Swedish and Spanish, the aim is to use robust cross-lingual techniques. The contribution of this work is the exploration of the use of ensemble techniques in a comparable task on both languages. We mean *ensemble* in two ways: on the one hand, we explored a simple combination of three base-learners (a perceptron, a CRF and an SVM); on the other hand, each base-learner was trained on ensembles of semantic spaces. The system rests on classical supervised classification techniques combined taking advantage of features derived from dense representations.

The focus of this paper is clinical entity recognition following the criteria in Pérez et al. (2017). That is, first, the decision space is set by means of semi-supervised representations that include ensembles of features derived from distributional semantics and also from classical symbolic representations (section 2 is devoted to the representation). The characterisation relied upon a big unannotated data-set, next, with an annotated set of much smaller size supervised classifiers can be inferred to decide whether a phrase is a clinical entity or not.

2 *Ensembles of features for clinical entity representation*

Classical entity recognition systems rested mainly upon word-forms (W) as a surface representation and lemmas with POS as a representation with linguistic (L) connotations. The linguistic features conveys helpful information, but to generate such features an analyser adapted to the medical domain is required, which is not available for all languages. Here, Freeling-Med (Oronoz et al., 2013) was used for Spanish and Stagger (Östling, 2013) paired with terminology matching following Skeppstedt et al. (2014) were used for Swedish.

In this work, the linguistic features were complemented with unsupervised (U) features. Current trends in language processing are shifting from symbolic representations based on words to distributional semantics. The benefits are multiple: while word-based representations tend to be scattered, continuous representations embed semantic information in a vector space. Classical symbolic representations (e.g. bag of words) entail a big number of components, and close vectors are rarely related. By contrast, distributional semantics keeps the dimension of the space manageable and permits a quantitative interpretation of word-relatedness. Word representations are obtained from big corpora by means of unsupervised techniques based on co-occurrences of words. The representation achieved depends not only on the corpus but also on a set of hyper-parameters influencing the training of the model. With a given corpus and different hyper-parameters,

different spaces are obtained. Yet, currently there is no conclusive fine-tuning technique to decide on the parameter setting. It has also been shown that when combining different spaces, rather than being redundant, the ensembles of semantic spaces enhanced the word-representation and improved information extraction techniques (e.g. entity recognition) (Henriksson, 2015).

In the clinical domain and, particularly, working with EHRs, available data tend to be scarce. The question arising is if distributional semantics can cope, in a robust and reliable way, with data sparseness. A typical method of dealing with sparsity of data given a continuous variable is clustering. Clustering regards as equivalent close values of a given variable as if we zoomed out our variable and could not make distinguishable close values.

On this account, the semantic spaces were clustered using k-means clustering. Again, k is a key parameter that changes the representation.

All in all, two semantic spaces were built from a given unannotated corpus, each of which with different hyper-parameters. The semantic spaces were clustered using two different numbers of clusters (k) in an attempt to combine fine-grained and coarse-grained clustering.

As an additional effort to handle data sparsity, features were also generated using Brown clustering (Brown et al., 1992). In this case, the information conveyed and the approach to get it is notably different. Brown clustering is a hierarchical clustering which arranges words found in a corpus into a tree with the words at the leaf nodes and where clusters corresponds to sub trees (Liang, 2005).

3 Ensemble classifier for entity recognition

This work started from the hypothesis that a simple ensemble learner would beat the individual base-learners, the contribution of three state of the art supervised classifiers was explored and next they were combined in a simple way. All the classifiers were trained using the ensemble representation-space features described in section 2.

3.1 Base-learners

Three approaches for supervised learning of medical entities were selected. The selected learners are all discriminative classifiers that perform sequential tagging, with different characteristics:

Perceptron This algorithm performs Viterbi decoding of the training examples combined with simple additive updates, trying to find the sequence of tags with the maximum score. The algorithm is competitive to other options such as CRFs (Carreras, Márquez, and Padró, 2003).

Support Vector Machine SVMs make use of kernel functions, which provide a similarity metric between two instances and, hence, a way to get a model suitable for discriminative tasks.

Conditional Random Fields CRF is a machine learning algorithm that makes use of feature functions representing the relationships between the features and the output. To assign the current output, it takes into account both earlier and later parts of the input and, also, the previous output tag.

3.2 Ensembles

The rationale of ensemble or committee models is quite intuitive: if many estimates are averaged together the variance of the estimate is reduced (Murphy, 2012). Regarding the ensembles, there are two key-issues:

1. The **diversity** of the base-models to be combined, since there is no point in combining models that make similar decisions. In this case, two kind of combinations were explored:
   (a) Combining models obtained with the same learning approach but different input representations or parameters, in this case with four different feature sets.
   (b) Combining models obtained with different learning approaches, namely the CRFs, Perceptrons and SVMs.
2. The **combination strategy**. There are a wide variety of combination strategies: linear opinion pools (or simple voting); weighted voting; stacking or stacked generalization learns a classifier from the

predictions of the base-learners; and others. Weighted voting was selected for its simplicity, and the weights were set to the average F-score of each base-model.

4 Experimental layout

4.1 Task and corpus

Two data sets were used for each language, a smaller annotated set and a larger set used for the unsupervised features. For Spanish, diseases (4,296 instances) and drugs (1,862 instances) were annotated, and for Swedish the annotated entities were body parts (2,082 instances), disorders (981 instances) and findings (3,759 instances) from HEALTH BANK[1] (Dalianis et al., 2015). The annotated data was divided into training sets containing about 60% of the annotations, development sets with 20% of the data and a test set with the remaining 20%. The unannotated data sets were of similar size for both languages, 52×10^6 tokens for Spanish and 51×10^6 for Swedish. More details about the data sets can be found in (Pérez et al., 2017).

4.2 Results

Table 1 shows the results for the ensemble tagger. The ensemble model was built up of 3 base-learners (CRF, Perceptron and SVM) each of which was trained in 4 alternative spaces using different sets of features:

1. **W:** Word-forms.
2. **WL:** Word-forms and Linguistic information (lemmas and part of speech)
3. **WLU:** the previous WL and Unsupervised features (ensembles of semantic spaces clusterized and Brown clusters).
4. **WU:** just word-forms and unsupervised features.

The composition of the 12 base-models consisted of a simple weighted voting strategy where the weights associated with each base-learner were set according to their individual F-scores on the development set. To be precise, the votes were weighted by the average F-score over all the the classes (the different entities in each set), giving 12 votes in total. This means that a model that proved more successful on the development data was given a stronger influence over the final tagging.

Other strategies are possible, for example using only the three base-models trained on the feature set that was most successful (WLU), and relying only on 3 votes for the ensemble, nevertheless, these results were slightly lower than the ensemble of all the models provided in Table 1.

Spanish				
Set	**Entity**	**P**	**R**	**F**
Dev	Disease	69.98	60.61	64.96
	Drug	94.95	82.76	88.43
	Average	78.55	67.46	72.58
Test	Disease	69.92	55.82	62.08
	Drug	94.38	84.46	89.15
	Average	78.68	65.22	**71.32**

Swedish				
Set	**Entity**	**P**	**R**	**F**
Dev	Body part	88.03	76.27	81.73
	Disorder	70.81	57.00	63.16
	Finding	63.89	58.33	60.98
	Average	72.68	63.98	68.02
Test	Body part	86.14	81.45	83.73
	Disorder	70.47	55.51	62.10
	Finding	68.28	65.35	66.78
	Average	74.39	69.24	**71.65**

Table 1: Results of the ensemble tagger comprising 12 base-models for each language. Evaluation metrics: Precision (P), Recall (R) and F-score (F)

4.3 Discussion

Not all the entities are equally easy to recognize for the system. Finding drugs or body parts is by far simpler than recognizing diseases, disorders or findings. Drugs, substances and brand-names tend to follow similar patterns and the same applies to body parts. By contrasts, in EHRs diseases, disorders and findings are expressed in a variety of ways that hardly ever follow their corresponding standard term in clinical dictionaries (e.g. ICD) or ontologies (e.g. SNOMED-CT). In medical records the same disease could be described in diverse and very different ways: either formal, or colloquial, either in a specific way or in a general way. In addition, there are variations in the way of expressing numbers (e.g. *"diabetes mellitus type II"*, *"diabetes mellitus type 2"*) and abbreviations are used fre-

[1]This research has been approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2014/1882-31/5.

quently (e.g. *"DM2"*). For example, Pérez et al. (2015) showed the case of the *"Malignant neoplasm of prostate"* disease that appeared in the EHRs with the variants *"Adenocarcinoma of the prostate"*, *"prostate adenocarcinoma"*, *"prostate Ca."* and *"PROSTATE CANCER"*.

The models were trained on the training set, fine-tuned on the development set and, finally, re-trained on a joint training and development set to assess the system on the test set. The results achieved in both development and test sets are comparable.

With regard to the difference of the performance across languages, while the results in the development set were better for Spanish than for Swedish, it was the other way around in the test set. Our intuition is that the differences in the performance on the development and test sets stand on the way the split was carried out. The sets were split at document level and given that the documents are much longer in the Spanish set, it might have made the inference tougher.

Previous work on medical entity recognition in this task showed that the aforementioned base-learners (SVM, CRF and Perceptron) were useful for the clinical domain. The best results for an individual model were achieved by the Perceptron using the WLU feature set. For Swedish, the average F-score in the test set for this configuration was 71.72 and for Spanish, the average F-score was 70.30 (Pérez et al., 2017). This work investigated the capability of ensemble techniques and explored diverse sets of features. The results of each of the 12 base-models involved were combined following a weighted voting strategy. We found that the ensemble approach was robust and that the overall trend, for both languages, and on both the development sets and the tests, was an improvement of the precision scores. On the test set this improvement was 1.54 points for Swedish and 4.3 points for Spanish. However, in most cases, the ensemble approach decreased the average recall. The recall was only improved on the Spanish development set. Altogether, the average F-scores were improved for both development and test data for Spanish, but only on the development data for Swedish.

The p-value given by the McNemar's test (McNemar, 1947) on the improvements achieved with respect to the best base-model on the development set (i.e. the Perceptron on WLU space) show statistical significance (p-value $\ll 0.01$) for Swedish, however, not for Spanish (p-value < 0.08). By contrast, for the best performing model in the test set, the difference with respect to the ensemble model is statistically significant with p-value $\ll 0.01$ for both languages.

It is debatable whether this increment in precision is worth the combination of 12 models. However, within the clinical domain precision tends to be crucial. Given the improvements achieved by this simple combination technique, the plans forfuture work include to test other methods for combining the learners, for example, stacked generalization could prove more efficient than the weighted voting approach.

5 Concluding remarks

Text in health records is challenging to process, and one of the biggest challenges for further work has to do with the variability associated to the spontaneous expressions found in medical records, particularly when it comes to express multi-word entities regarding the diseases, disorders and findings. A strength of this work stands on the comparable framework achieved which allows for evaluations of clinical entity recognition on clinical texts in two different languages A step ahead is made with respect to previous works combining three base-learners (CRF, Perceptron, SVM) inferred on four alternative spaces. Influenced by previous works these spaces, including unsupervised features such as clusterized ensembles of semantic spaces, brown clusters and also linguistically motivated features (word-forms, POS and lemmas), were built. All together, we constructed an ensemble that combined 12 base-models using a weighted voting paradigm where the weights were set as the average F-score of each model. The ensemble model achieved an average F-score above 71%. The combination increased the performance in terms of precision for both languages. It seems as if the upper threshold was not achieved yet and that there is room for improvement, specifically for recall. Therefore, ensemble techniques other than weighted voting should be explored for future work.

Acknowledgements

This work has been partially funded by the Spanish ministry (PROSAMED: TIN2016-

77820-C3-1-R, TADEEP: TIN2015-70214-P), the Basque Government (DETEAMI: 2014111003), the University of the Basque Country UPV-EHU (MOV17/14) and the Nordic Center of Excellence in Health-Related e-Sciences (NIASC).

References

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

Brown, P. F., P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Carreras, X., L. Márquez, and L. Padró. 2003. Learning a perceptron-based named entity chunker via online recognition feedback. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 156–159. Association for Computational Linguistics.

Crammer, K., M. Dredze, K. Ganchev, P. P. Talukdar, and S. Carroll. 2007. Automatic code assignment to medical text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136. Association for Computational Linguistics.

Dalianis, H., A. Henriksson, M. Kvist, S. Velupillai, and R. Weegar. 2015. HEALTH BANK–A Workbench for Data Science Applications in Healthcare. In *Proceedings of the CAiSE-2015 Industry Track co-located with 27th Conference on Advanced Information Systems Engineering (CAiSE 2015), J. Krogstie, G. Juel-Skielse and V. Kabilan, (Eds.), Stockholm, Sweden, June 11, 2015, CEUR, Vol-1381*, pages 1–18.

Demner-Fushman, D., W. W. Chapman, and C. J. McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5):760–772.

Ehrentraut, C., H. Tanushi, H. Dalianis, and J. Tiedemann. 2012. Detection of Hospital Acquired Infections in sparse and noisy Swedish patient records. In *Proceedings of the Sixth Workshop on Analytics for Noisy Unstructured Text Data*.

Faruqui, M. and S. Padó. 2010. Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In *KONVENS*, pages 129–133.

Florian, R., A. Ittycheriah, H. Jing, and T. Zhang. 2003. Named entity recognition through classifier combination. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 168–171. Association for Computational Linguistics.

Haas, J. P., E. A. Mendonça, B. Ross, C. Friedman, and E. Larson. 2005. Use of computerized surveillance to detect nosocomial pneumonia in neonatal intensive care unit patients. *American journal of infection control*, 33(8):439–443.

Henriksson, A. 2015. *Ensembles of semantic spaces: On combining models of distributional semantics with applications in healthcare*. Ph.D. thesis, Department of Computer and Systems Sciences, Stockholm University.

Henriksson, A., M. Kvist, H. Dalianis, and M. Duneld. 2015. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. *Journal of Biomedical Informatics*, 57:333–349.

Kang, N., Z. Afzal, B. Singh, E. M. Van Mulligen, and J. A. Kors. 2012. Using an ensemble system to improve concept extraction from clinical records. *Journal of biomedical informatics*, 45(3):423–428.

Lafferty, J. D., A. McCallum, and F. C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lai, K. H., M. Topaz, F. R. Goss, and L. Zhou. 2015. Automated misspelling detection and correction in clinical free-text records. *Journal of Biomedical Informatics*, 55(Supplement C):188–195.

Liang, P. 2005. *Semi-Supervised Learning for Natural Language*. Ph.D. thesis, Massachusetts Institute of Technology.

McNemar, Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Murphy, K. P. 2012. *Machine Learning: A Probabilistic Perspective*. MIT Press.

Oronoz, M., A. Casillas, K. Gojenola, and A. Perez. 2013. Automatic annotation of medical records in Spanish with disease, drug and substance names. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Springer, pages 536–543.

Östling, R. 2013. Stagger: An open-source part of speech tagger for Swedish. *Northern European Journal of Language Technology*, 3:1–18.

Pérez, A., K. Gojenola, A. Casillas, M. Oronoz, and A. D. a. de Ilarraza. 2015. Computer aided classification of diagnostic terms in Spanish. *Expert Systems with Applications*, 42:2949–2958.

Pérez, A., R. Weegar, A. Casillas, K. Gojenola, M. Oronoz, and H. Dalianis. 2017. Semi-supervised medical entity recognition: A study on Spanish and Swedish clinical corpora. *Journal of Biomedical Informatics*, 71:16–30.

Reynoso, G. A., A. D. March, C. M. Berra, R. P. Strobietto, M. Barani, M. Iubatti, M. P. Chiaradio, D. Serebrisky, A. Kahn, O. A. Vaccarezza, J. L. Leguiza, M. Ceitlin, D. A. Luna, F. G. B. de Quirós, M. I. Otegui, M. C. Puga, and M. Vallejos. 2000. Development of the Spanish Version of the Systematized Nomenclature of Medicine: Methodology and Main Issues. In *Proceedings of the AMIA Symposium*, pages 694–698.

Ruch, P., R. Baud, and A. Geissbühler. 2003. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. *Artificial intelligence in medicine*, 29(1):169–184.

Saha, S. and A. Ekbal. 2013. Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition. *Data & Knowledge Engineering*, 85:15–39.

Skeppstedt, M., M. Kvist, G. Nilsson, and H. Dalianis. 2014. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of Biomedical Informatics, 49*, pages 148–158.

Tang, B., H. Cao, X. Wang, Q. Chen, and H. Xu. 2014. Evaluating word representation features in biomedical named entity recognition tasks. *BioMed research international*, 2014.

Turian, J., L. Ratinov, and Y. Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics.

Zea, J. L. C., J. E. O. Luna, C. Thorne, and G. Glavaš. 2016. Spanish NER with Word Representations and Conditional Random Fields. In *Proceedings of the Sixth Named Entity Workshop*, pages 34–40.

Procesamiento del Lenguaje Natural, Revista nº 60, marzo de 2018, pp. 21-28 recibido 03-11-2017 revisado 08-01-2018 aceptado 14-02-2018

Not all the questions are (equally) difficult. An hybrid approach to CQA in Arabic

No todas las preguntas son (igualmente) difíciles, una aproximación híbrida a la CQA en árabe

Imane Lahbari[1], **Horacio Rodríguez**[2], **Said Ouatik El Alaoui**[1]
[1]Laboratory of Informatics and Modeling, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco
[2]Dep. of Computer Science, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
imane.lahbari@usmba.ac.ma, horacio@cs.upc.edu, s_ouatik@yahoo.com

Abstract: In the past we faced the problem of Community Question Answering using an unified approach. Some of the questions, however, are easier to be approached by a conventional rule-based system. In this paper we explore this direction.
Keywords: Community question answering, arabic processing

Resumen: En el pasado hemos abordado la búsqueda de respuestas en comunidades usando un enfoque uniforme. Sin embargo, algunas preguntas pueden ser respondidas utilizando métodos basados en reglas. En este tratajo exploramos esta dirección.
Palabras clave: Búsqueda de respuestas, procesamiento del árabe

1 Introduction

Community Question Answering (*CQA*) has become increasingly popular in the last years. It is seen as an alternative to both classical Information Retrieval and more specific Question Answering (*QA*) tasks. Both general purpose, as Yahoo!Answers (*Y!A*)[1], and topic-specific communities, such as StackOverflow (*SO*)[2], have got an impressive growth.

CQA purpose is to provide to users pertinent answers to their questions by identifying, among a set (sometimes a thread) of question answer pairs, questions that are similar to the original one.

The SemEval Task 3 subtask D (Nakov et al., 2017) asks, given a question in Arabic, denoted the original question, and a set of the first 9 related questions (retrieved by a search engine), each associated with one correct answer, to re-rank the 9 question-answer pairs according to their relevance with respect to the original question. Figure 1 presents a fragment of a query thread containing an Arabic query (a), and its English translation, carried out using Google Translate API (b). It is worth noting from this example: i) the high density of medical terms as seen in (c), ii) the occurrence of English terms embedded within the Arabic texts, that could complicate the linguistic process of Arabic texts, iii) the relatively high overlapping of terms between the query texts and the texts of query/answer pairs in both Arabic texts and English translations, and iv) The relatively low quality of English translations, that could result on low accuracy of the linguistic process of English texts.

We developed in the past a CQA system based on the combination of a number of atomic classifiers, which was evaluated in the framework of SemEval 2017 Task 3 subtask D, getting good results. Some of the questions, however, seem to be easier to be approached by a conventional rule-based system. In this paper we explore this direction.

2 Related works

QA, i.e. querying a computer using Natural Language, is an old objective of Natural Language Processing. Though initially *QA* systems focused on factual questions (who, where, when, Y/N, etc.), increasingly, the scope of *QA* has become wider, facing complex questions, list questions, definitional, why questions, etc. In parallel the *QA* sys-

[1]http://answers.yahoo.com/
[2]http://stackoverflow.com/

ISSN 1135-5948. DOI 10.26342/2018-60-2
© 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

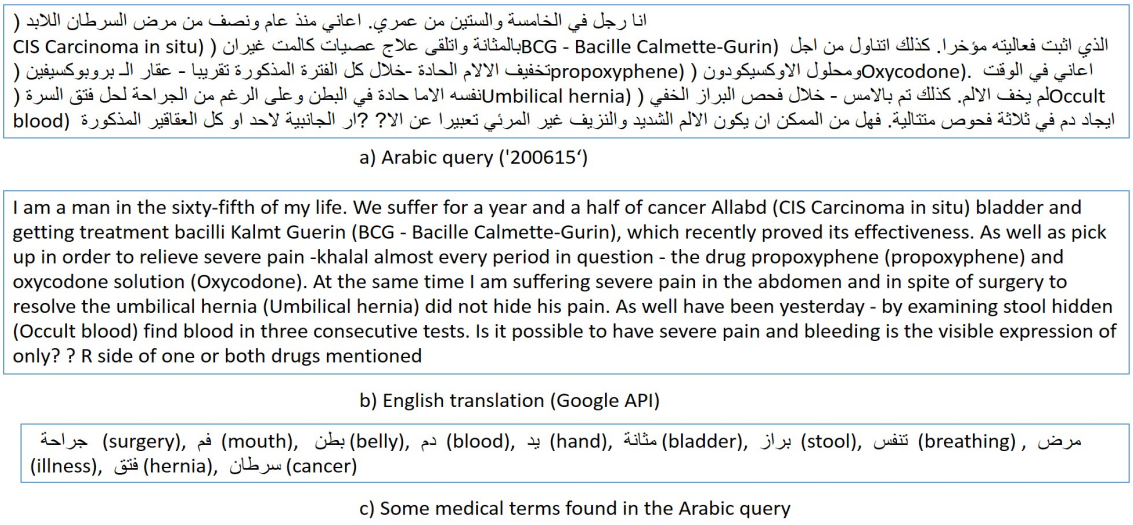

Figure 1: Query thread fragment

tems have suffered a process of specialization: domain-restricted *QA*, *QA* for comprehension reading, *QA* over Linked Data, or *CQA*.

CQA differs from conventional *QA* systems basically on three aspects:

- The source of the possible answers, that are threads of queries and answers activated from the original query. So, the document or passage retrieval components, needed in other *QA* systems can be avoided or highly simplified;
- The structure of the threads and the available metadata can be exploited for the task
- The types of questions include the frequent use of complex questions, as definitional, why, consequences, how_to_proceed, etc.

Many approaches have been applied to the task (Nakov et al., 2016; Nakov et al., 2017; El Adlouni et al., 2016; El Adlouni et al., 2017) for details and references.

Most of the systems use, as core features or combined with others, textual features, superficial (string-based), syntactic, and, less frequently, lexico-semantic (knowledge-based), usually reduced to similarity or relatedness measures between the textual components of the thread (query, query/answer pairs), Gomaa and Fahmy (2013) present an excellent survey of these class of features;

Most of the research on QA has been applied to English language. There are, however, interesting examples in other languages, including Arabic. Two of the most useful references for Arabic *QA* are the thesis of Yassine Benajiba (Benajiba, 2009) and Lahsen Abouenour (Abouenour, 2014). Focusing on rule-based approaches, interesting systems are: QARAB, (Hammo, Abu-Salem, and Lytinen, 2002), for Factoid questions, DefArabicQA, (Trigui, Belguith, and Rosso, 2010), for Definitional questions, and, EWAQ, (Al-Khawaldeh, 2015), an Entailment-based system.

3 Our SemEval 2017 system

In this section we present our previous system, (El Adlouni et al., 2017; El Adlouni et al., 2016), evaluated in the framework of SemEval-2017 Task 3 D, on which we will further include our rule-based module described in this article. Our official results in his contest were rather good, second (but from only 3 teams) in *MAP* and first in *accuracy*. Our system combined different basic classifiers in several ways.

The overall architecture of our system is shown in Figure 2. As can be seen, the system performs in four steps:

- A preparation step, aiming to collect domain (medical) specific resources;
- A learning step, for getting the models of the classifiers;

- A classification step, for applying them to the test dataset. These two steps are applied independently for each of the basic classifiers;
- A last step combines the results of the atomic classifiers for obtaining the final results.

We describe each step next.

3.1 Overall description

A core component of our approach is the use of a medical terminology, covering both Arabic and English terms and organized into three categories: *body parts*, *drugs*, and *diseases*. We decided to use this resource taking into account the origin of the datasets for task D: *medical fora*. The terminology was automatically collected as reported in El Adlouni et al. (2017). The process of collecting it was performed in a multilingual setting (7 languages were involved). Some of the languages provided available medical resources (as SnomedCT for English, French, and Spanish, DrugBank, and BioPortal for English, and other), while translingual links were obtained from DBPedia (English, French, German, and Spanish) through the use of *same_as* and *label* properties (Cotik, Rodriguez, and Vivaldi, 2017). The final figures for Arabic and English can be found in Table 1.

After downloading the training (resp. test) Arabic dataset we translated into English all the Arabic query texts and all the Arabic texts corresponding to each of the query/answers pairs. For doing so we have used the Google Translate API[3].

For processing the English texts we have used the Stanford CoreNLP toolbox[4] (Manning et al., 2014). For Arabic we have used *Madamira*[5] (Pasha et al., 2015).

The results obtained were then enriched with WordNet synsets both for Arabic (Rodríguez et al., 2008) and English (Fellbaum, 1998). Also the sentences extracted were enriched with Named Entities corresponding to the medical terminologies collected in the preparation step[6].

[3]translate.google.com

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

[5]http://nlp.ldeo.columbia.edu/madamira/

[6]Some of these terms are classified as ORG or MISC, by the linguistic processors, others are simply not recognized as Named Entities.

Medical Category	English	Arabic
Body Part (BP)	25,607	1,735
DISEASE	292,815	3,352
DRUG	87,254	2,149

Table 1: Medical terms datasets

No WSD was attempted. As detection of medical multiword terms is poor in StanfordCore and Madamira, a post process for locating them when occurring in our medical terminologies or WordNets was carried out.

After that, a process of feature extraction was performed. This process is different for each atomic classifier and will be described in next sections. Finally, a process of learning (resp. classification) is performed. Also these processes differ depending on the involved classifier and will be described next.

Our approach for learning consists on obtaining a set of N classifiers. Besides classifying, a score or credibility value is provided by the classifier that can be used in fact as a ranker[7].

3.2 Atomic classifiers

The set of atomic classifiers was selected in order to deal with the different aspects that seem relevant and have been successfully applied to similar tasks: textual features, IR, topics, dimensionality reduction, etc.). The atomic classifiers used by our system are the following:

- Basic lexical string-based classifiers, i.e. *Basic_ar* and *Basic_en*, see details in El Adlouni et al. (2017);
- A simple *IR* system, using *Lucene* engine;
- A *LSI* system, learned from different datasets;
- A topic-based *LDA* system;
- A *Embedding* system.

3.2.1 Basic classifiers

We use two equivalent basic atomic classifiers, one applied to Arabic (*basic_Ar*) and the other to English (*basic_En*). The basic classifiers use three sets of features: shallow linguistic features, vectorial features, and domain-based features. As shallow linguistic features we use most of the 147 features proposed in Felice, M. (2012).

[7]Because the task we face consists on both classifying and ranking.

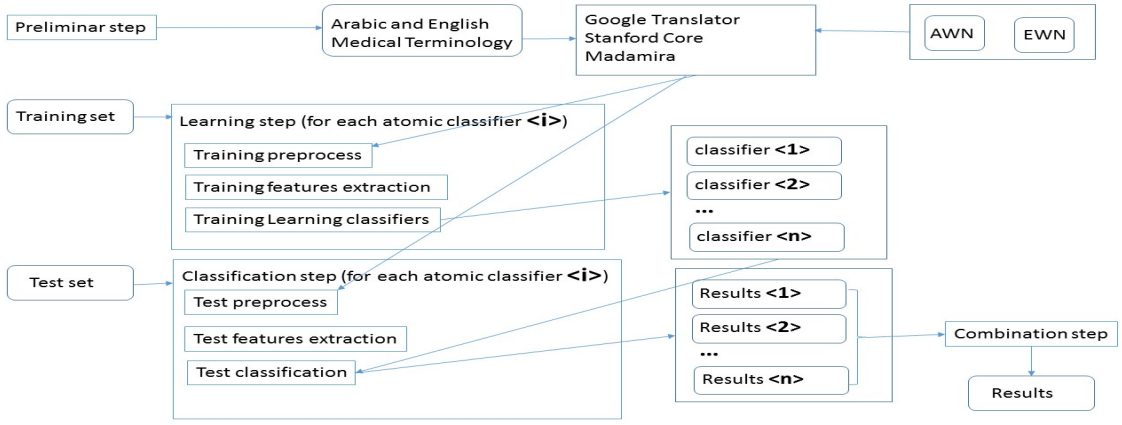

Figure 2: Train and testing pipelines

It is worth noting the importance of the medical features: While only 57 out of 147 basic features were used by the classifier, the whole set of medical features (16) were used. Ranking the features by decreasing accuracy 4 medical features (i.e. 25% of them) occur between the 20 highest ranked features.

We have used for learning the *Logistic Regression* classifier in the Weka toolkit[8].

3.2.2 Lucene classifier

Using Lucene[9], we index the pairs by using all possible combinations (q, q^i, a^i, and $q^i \oplus a^i$) searching thereafter for each pair $<q, q^i>$ for obtaining a set of hits or document with their respective relevance to the query.

3.2.3 LSI and Embedding classifiers

For dealing with dimensionality reduction we have used two techniques, LSI and embeddings. LSI was used to have dense representations of our sentences by using SVD. Various corpora were used for that matter including Wikipedia latest dump (January 2017), Webteb.com, altibbi.com and dailymedicalinfo.com which are specialized Arabic websites for medical domain articles. For embeddings we used the *doc2vec* approach described in Le and Mikolov (2014).

3.2.4 LDA classifier

As for LSI, LDA is used to produce dense representations for our sentences using the implementation included within *Gensim* (Hoffman, Bach, and Blei, 2010). Our aim is to capture topics implicitly occurring within the questions.

[8]http://www.cs.waikato.ac.nz/ml/weka/
[9]https://lucene.apache.org/

3.3 Combinations

Our combiner receives as input a set of atomic result and scores and returns an overall result and score.

The combiner is driven by a set of hiperparameters:

- *scoring form*, i.e. 'max' or 'ave';
- *thresholding form*, i.e. None, 'global' or 'local';
- *thresholding level*, i-e. 0.2, 0.4, 0.6, 0.8;
- *result form*, i.e. 'max', 'voting', 'coincidence'.

We have used grid search for setting the best combination of the hiper-parameters, using the development dataset.

4 Experimental framework

The scorer made available by SemEval organizers provides a range of evaluation metrics to assess the quality of the proposed model, the two most important are *MAP* and *Accuracy*. The former which stands for the Mean Average Precision is the official score of the competition and is based on the top ranked question-answer pairs for each original question leveraging the value computed for our score in our dataset on a scale from 0 to 1. The latter is based on the binary result (relevant or not).

4.1 Results

In Table 2 a summary of the Official results of Semeval 2017 Task 3 Subtask D, are presented (all but last row).

Team	MAP	Acc
GW_QA-primary	0.6116	0.6077
UPC-USMBA-primary	0.5773	0.6624
QU_BIGIR-primary	0.5670	0.4964
UPC-USMBA-with rules	0.5786	0.6747

Table 2: Official results of the task

ruleset	accuracy: m	a	overall
Arabic	0.757	0.559	0.635
English	0.763	0.549	0.652
union	0.755	0.54	0.629
intersection	0.921	0.842	0.875

Table 3: Accuracy of rule-based on test set

Regarding *MAP*, and so looking at the official rank, we were placed in the middle (2nd from 3 participants). Regarding *accuracy* we are placed on the top of the rank. We analyzed the results in the test dataset of our atomic classifiers (with different parameterization) and combinations. The *MAP* for the atomic classifiers (using the best parameters got in training) range from 55 to 58.32. All the atomic results were outperformed by our combiner run but *Lucene*, which obtained our best result, 58.32.

5 Including a rule-based model

A careful examination of errors in our previous approach revealed that some apparently easy cases, as those shown in Table 4, failed to be correctly classified. We saw that some of the original queries, though not the majority, corresponded to factoid questions and could be approached by a conventional rule-based system. So, we developed a rule-based model for facing factoid questions, i.e. cases where a clear, although possibly not unique, objective can be extracted from the text. This rule-based model will be later included into our combination approach.

Consider, for instance, a question beginning with "What is the cause of", and containing close to it a disease name. This question can be intuitively classified as *CAUSE_DISEASE* and parameterized with the tag *DISEASE* with the extracted name as value.

Our rule-based approach consists of the following steps:

- We build a set of question types (QT), *QTS*. QT are domain-restricted semantically-driven tags. *QTS* consists of 27 QT, including DEFINITION_DISEASE, CAUSE_DISEASE, SIDE_EFFECTS_DRUG, etc. The later can be paraphrased as "given an instance of a DRUG, what are its possible side effects (clinical findings)?";
- For each QT we have build 4 sets of classification rules, for Arabic and English, manually and automatically built. For building the rules the training material of SemEval was used. The process of building these rulesets is detailed in section 5.1. The process resulted in 27 Manual Arabic rules, 29 Automatic Arabic rules, 52 Manual English rules, and 83 Automatic English rules. An average of 8 rules per QT have been built;
- Extraction rules are straightforward and language independent. We have manually built one for each QT;
- We have built a rule-based classifier that can be applied to the original question q or and to any of the questions in the pairs of the thread (q^i). The same classifier is used for both languages using the corresponding rule-set. The classifier returns for each case zero (in most of the cases) or more QT from *QTS*. We have built, too, a rule-based extractor that can be applied to the answers in the pairs of the thread (a^i).

I suffer from psoriasis since a long time I want ...
Is there a cure for psoriasis in Jordan ?
Is there a cure for psoriasis ?
Do Hnal cure for psoriasis
I have psoriasis in the top of the ...

Table 4: Some questions in the thread of "What the treatment of psoriasis ?"

The application of the classifiers/extractors is as follows: The process of *classification rules* consist of obtaining the QT, deriving from it the *Expected Answer Type* (EAT), and set the *Mandatory Constraints* (MC) and *Optional Constraints* (OC)[10]. For example, for the case of QT SIDE_EFFECTS_DRUG, the MC is reduced to the tag 'DRUG' associated with the specific name quoted in the question.

[10] Although both MC and OC are generated, only the former are considered in this paper.

After applying the classifiers to all the cases a pair $<q,\ q^i>$ is considered relevant when:

- q and q^i are classified into the same QT (not necessarily by the same rule or language);
- The involved MC are compatible;
- An extraction rule can be applied to a^i using the same MC.

The sets of rules have been evaluated in terms of accuracy over the test set. The results are shown in Table 3. We depict the accuracy of the Arabic and English rulesets, their union and their intersection for manual, automatic, and overall rulesets. It is worth noting the serious degradation of accuracy from manual to automatic rules and the relative similarity of performance for Arabic and English.

5.1 Building the rulesets

Classification rules perform on all the questions, both q and q^i.

A rule consists of a sequence of conditions followed by a sequence of actions (usually only an action is included into the rule). Actions are executed only when all the conditions are satisfied. Each condition (and action) returns a value that can be used by the following ones. The action part of the rule is in charge of building the constraints that will be evaluated by the extraction rules. Extraction rules are associated to the MC and OC obtained by classification rules, Usually they are reduced to check whether the entities (diseases, drugs, body parts) contained in MC occur on the answer text. There are basically three kind of conditions in classification rules (see some examples just below):

- Those checking for the occurrence of textual patterns referring to words, lemmas, pos, NEs, ... on the text of the question;
- Those looking for the occurrence of medical entities (DISEASE, DRUG, BP) from our medical vocabularies;
- Those establishing distance constraints between the tokens located in 1 and 2.

A total of 22 condition predicates have been built to be used within the rules. In average each rule contains 5 conditions. Some of these predicates are the following:

```
def CQARule_SYMPTOMS_DISEASE_en_2(lang,qT):
# 20006, 100739, What are the symptoms of bird-pig disease ...
pattern = [u'l(what)', 'sk(0:2)', u'l(symptom)', 'sk(0:2)', 'n(DISEASE)', 'sk(*)']
id = 'CQARule_SYMPTOMS_DISEASE_en_2'
p = QtclassRegularPattern(id, pattern)
p.prepare()
r = QTclassrule('SYMPTOMS_DISEASE_en_2',qT,lang)
r.addCondition(QTclasscondition("c0","thereAreTriggers(l, qT, s)"))
r.addCondition(QTclasscondition("c0","noStigmas(l, qT, s)"))
r.addCondition(QTclasscondition("c0","noYNQuestion(l,s)"))
r.addCondition(QTclasscondition(
   "c1","applyComplexPattern('CQARule_SYMPTOMS_DISEASE_en_2',s)"))
r.addAction(QTclassAction("a1",'addInvolvedToMandatory_1([("c1",-1)],[])'))
return r
```

Figure 3: Example of rule

- thereAreTriggers: Checks whether the question contains at least one of the triggers of the QT, i.e. terms heavily pointing to this QT.
- noStigmas: Checks whether the question contains stigma terms, i.e. terms forbidden for the QT, usually triggers of the other QT;
- noYNQuestion: Checks whether a pattern for a YN question occurs;
- applySimplePattern: Checks whether the regular expression in *pattern* is satisfied by the question;
- applySortedPatterns: Sorts the list of strings in *patterns* by decreasing length and checks their occurrence in the question;
- existInInstancesInOntology: Checks whether instances of the elements of *involved* occur in the question;
- checkDistanceConstraint: Checks the distance constraints, contained in *constrains* between the tokens located in previous conditions.

We tried to build rules for the most used patterns. Within the training data set, people use to ask about their own issues. We studied this data set and we extract the most used expressions. In general, people ask about diseases, drug or body parts (BP) which are automatically detected by our system. The interrogative patterns, *IP*, are the first component for building any rule, then we describe the whole expression. For each expression, we define a few tokens after the *IP*, then we add the extracted diseases (or drug, or BP).

An example of Python function for building a manual English rule is shown in Figure 3. The function for creating

the rule has two parameters, the language and the *QT*, "English" and "SYMPTOMS_DISEASE" in this case. The identification of the rule is defined as id = "CQARule_SYMPTOMS_DISEASE_en_2". The rule includes as a comment an example of application: "What are the symptoms of bird-pig disease ..." The rule owns an internal parameter, *pattern*, that can be paraphrased as: The question starts with a token whose lemma should be "what", next zero to two tokens could be skipped, the next token has be have a lemma "symptom", new skipping of up to two tokens and a token corresponding to a NE of type "DISEASE". Finally the rest of tokens could be skipped.

This rule contains 4 conditions and 1 action. The first three conditions apply "thereAreTriggers", "noStigmas", and "noYNQuestion". The results of all these three conditions are assigned to the variable "c0", not later used. The fourth condition checks whether the complex pattern is satisfied. The list of tokens, "what", "symptom", and the disease, is assigned to variable "c1". The only action simply builds the *MC* including the last member of "c1", i.e. the name of the disease.

Building manually the set of classification rules resulted on 27 rules for Arabic and 52 for English. Although these rules offered a nice precision, the recall was very low. So, we decided to complement these rulesets by means of a semi-automatic procedure involving a very low human intervention. This process is as follows:

For each *QT* and for both languages, all the manual rules were applied to all the questions (q and q^i) in the training set. We collected all the cases of success. We obtained in this way a set of question texts (444 for Arabic and 746 for English). For each of these texts we collected the occurring n-grams (up to 5-grams including up to 2 skips). Using a tf*idf weighting, the most frequent n-grams were obtained. This resulted on 3,958 n-grams for Arabic and 1,702 for English. From this information we built two matrices of 27 rows corresponding to *QTS* and 3,958 (1,702) columns, number of selected n-grams. This matrices were manually revised and some of the columns were removed. Then for each row the involved medical entities (DISEASE, DRUG, BP, ANY) and their distance constraints were manually added. After this process the set of automatic rules is easily generated.

In Table 3 global accuracy of the set of rules obtained on the test set are presented.

The rule-based classifier has been incorporated to our combiner getting the result shown in the last row of Table 2. Both MAP and accuracy got an improvement though only the later is significant.

6 Conclusions and future work

Our official results on the contest have been rather good, second (but from only 3 teams) in *MAP* and first in *accuracy*. The inclusion of our rule-based classifier has consistently outperformed *accuracy*. *MAP* has also improved but the improvement is not significant. This is due to the fact that a very limited number of cases has changed, so, although the binary results (classification) have improve, the scores (ranks) have changed only marginally.

Our next steps will be:

- Performing an in depth analysis of the performance of our two rulesets, analyzing the accuracy of each rule and cross comparing the rules fired in each language. It is likely that if a rule has been correctly applied to a pair for a language a corresponding rule in the other language should be applied as well, so modifying an existing rule or including a new one could be possible. This line of research can be followed for both manual and automatic approaches.
- Using a final ranker (not a simple classifier) over the results of our atomic classifiers for trying to improve our *MAP*.
- Trying others NN models as CNN and LSTM that have provided good results for English.
- Extending the coverage of our medical terminologies to other medical entities (procedures, symptoms, clinical signs and findings).

Acknowledgments

We are grateful for the suggestions from three anonymous reviewers. Dr. Rodríguez has been partially funded by Spanish project "GraphMed" (TIN2016-77820-C3-3R).

References

Abouenour, L. 2014. *Three-levels Approach for Arabic Question Answering Systems.* Ph.D. thesis, University Mohammed V Agdal, R, (Morocco).

Al-Khawaldeh, F. T. 2015. Answer extraction for why arabic question answering systems: Ewaq. In *World of Computer Science and Information Technology Journal (WCSIT)*, pages 82–86.

Benajiba, Y. 2009. *Arabic Named Entity Recognition.* Ph.D. thesis, UPV Valencia (Spain).

Cotik, V., H. Rodriguez, and J. Vivaldi. 2017. Arabic medical entity tagging using distant learning in a Multilingual Framework. In *Journal of King Saud University-Computer and Information Sciences, vol. 29*, pages 204–211.

El Adlouni, Y., I. Lahbari, H. Rodríguez, M. Meknassi, S. O. El Alaoui, and N. Ennahnahi. 2016. Using domain knowledge and bilingual resources for addressing community question answering for arabic. In *4th IEEE International Colloquium on Information Science and Technology, CiSt 2016, Tangier, Morocco, October 24-26, 2016*, pages 368–373.

El Adlouni, Y., I. Lahbari, H. Rodríguez, M. Meknassi, S. O. El Alaoui, and N. Ennahnahi. 2017. Upc-usmba at semeval-2017 task 3: Combining multiple approaches for cqa for arabic. In *Proceedings of SemEval 2017.*

Felice, M. 2012. Linguistic Indicators for Quality Estimation of Machine Translations. In *Master's thesis. University of Wolverhampton, UK.*

Fellbaum, C. 1998. WordNet: An Electronic Lexical Database. In *MIT Press, Cambridge, Mass (USA).*

Gomaa, G. and A. Fahmy. 2013. A Survey of Text Similarity Approaches. In *International Journal of Computer Applications 04/2013; 68(13).*

Hammo, B., H. Abu-Salem, and S. Lytinen. 2002. Qarab: A question answering system to support the arabic language. In *Workshop on Computational Approaches to Semitic Languages*, pages 1–11.

Hoffman, M., F. R. Bach, and D. M. Blei. 2010. Online learning for latent dirichlet allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23.* Curran Associates, Inc., pages 856–864.

Le, Q. and T. Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1188–1196.

Manning, C., M. Surdeanu, B. J., J. Finkel, S. Bethard, and D. McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *ACL System Demonstrations*, pages 55–60.

Nakov, P., L. Màrquez, A. Moschitti, W. Magdy, H. Mubarak, A. A. Freihat, J. Glass, and B. Randeree. 2016. Semeval-2016 task 3: Cqa. In *Proceedings of SemEval '16*, San Diego, California. ACL.

Nakov, P., D. Hoogeveen, L. Màrquez, A. Moschitti, H. Mubarak, T. Baldwin, and K. Verspoor. 2017. SemEval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada. ACL.

Pasha, A., M. Al-Badrashiny, M. Diab, N. Habash, M. Pooleery, O. Rambow, and R. Roth. 2015. Madamira 2.1. In *Center for Computational Learning Systems Columbia University*, pages 55–60.

Rodríguez, H., D. Farwell, J. Farreres, M. Bertran, M. Alkhalifa, M. A. Martí, S. Elkateb, W. Black, J. Kirk, A. Pease, P. Vossen, and C. Felbaum. 2008. Arabic WordNet: Current state and future extensions. In *Proceedings of the Fourth Global WordNet Conference, Szeged, Hungary*, pages 387–405.

Trigui, O., L. H. Belguith, and P. Rosso. 2010. DefArabicQA: Arabic Definition Question Answering System. In *7th Workshop on Language Resources and Human Language Technologies for Semitic Languages*, pages 40–45.

Procesamiento del Lenguaje Natural, Revista nº 60, marzo de 2018, pp. 29-36 recibido 05-11-2017 revisado 07-01-2018 aceptado 02-02-2018

A Supervised Central Unit Detector for Spanish

Un detector de la unidad central para textos en castellano

Kepa Bengoetxea and Mikel Iruskieta
IXA Group. University of the Basque Country
{kepa.bengoetxea,mikel.iruskieta}@ehu.eus

Resumen: En este artículo presentamos el primer detector de la Unidad Central (CU) de resúmenes científicos en castellano basado en técnicas de aprendizaje automático. Para ello, nos hemos basado en la anotación del *Spanish RST Treebank* anotado bajo la Teoría de la Estructura Retórica o *Rhetorical Structure Theory* (RST). El método empleado para detectar la unidad central es el modelo de bolsa de palabras utilizando clasificadores como Naive Bayes y SVM. Finalmente, evaluamos el rendimiento de los clasificadores y hemos creado el detector de CUs usando el mejor clasificador.
Palabras clave: Unidad central, RST, clasificación, minería de datos, Naive Bayes, SVM

Abstract: In this paper we present the first automatic detector of the Central Unit (CU) for Spanish scientific abstracts based on machine learning techniques. To do so, learning and evaluation data was extracted from the *RST Spanish Treebank* annotated under the *Rhetorical Structure Theory* (RST). We use a bag-of-words model based on Naive Bayes and SVM classifiers to detect the central units of a text. Finaly, we evaluate the performance of the classifiers and choose the best to create an automatic CU detector.
Keywords: Central unit, RST, classification, data mining, Naive Bayes, SVM

1 *Introduction*

Knowing what is the most important sentence of a text and the intention in which this was uttered is a crucial task for language learners to understand a text.

Following Iruskieta, Diaz de Ilarraza, and Lersundi (2014) the central unit (CU) is an elementary discourse unit (EDU) and the most salient text-span of a rhetorical structure. Rhetorical structures or the RST diagrams are represented as trees (RS-trees) and there is at least one text-span[1] that is not modified by any other EDU through any mononuclear relation. On the contrary, this text span functions as the central node of the tree.

Determining first the most important segment of a discourse in a text is crucial also to annotate the rhetorical structure of a text (Iruskieta, de Ilarraza, and Lersundi, 2014), but also for some advanced NLP tasks such as sentiment analysis, summarization tasks and question answering, among others.

Automatic classification is a learning process, during which a program recognizes the features that distinguish each category from others and constructs a classifier when given a set of training examples with class labels. Application of this approach to the CUs can help in automatic detection on the basis of similarity of their content. In this research we classify CUs using the bag of words model. Algorithms used in classification are Naive Bayes (NB) (McCallum, Nigam, and others, 1998) and Support Vector Machine (SVM) (Cortes and Vapnik, 1995) that were successfully used in previous researches in text classification (Schneider, 2005).

Some CU's detectors were developed for Basque (Bengoetxea, Atutxa, and Iruskieta, 2017) and for Brazilian Portuguese (BP),[2] but there is no tool to detect the CU for Spanish.

To fulfill this gap, the main aim of this paper is to built an automatic Central Unit detector for Spanish scientific abstracts.

[1] If the relation at the top is a multinuclear one, there are more than one EDU functioning as CU.

[2] The demos of these two tools can be tested at http://ixa2.si.ehu.es/CU-detector/ for Basque (reliability of 0.57 F1) and http://ixa2.si.ehu.es/clarink/tools/BP-CU-detector/ for BP (reliability of 0.657 F1).

ISSN 1135-5948. DOI 10.26342/2018-60-3
© 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

Although this tool can be used in different approaches, it was developed under Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) that is a descriptive, language-independent theory of the organization of texts, which characterizes the text structure primarily in terms of the hierarchical relations that hold between discourse segments (EDUs).

Following Iruskieta et al. (2013), the rhetorical analysis of a text includes three phases: *i*) text segmentation (EDUs), *ii*) CU annotation and *iii*) description of relations between EDUs and groups of EDUs linked to the CU, building a hierarchical tree (RS-tree).

The results shows that we can build an entirely automatic CU detector with a good performance if we depart from an annotated RST corpus.

2 Related Work

Using different techniques some CU detectors were developed following the findings by Iruskieta, Diaz de Ilarraza, and Lersundi (2014) that show the importance of annotating the CU before rhetorical relations: *i*) Rule based detectors use features that were design by linguists for Basque[3] and Brazilian Portuguese (Bengoetxea, Atutxa, and Iruskieta, 2017; Iruskieta, Antonio, and Labaka, 2016). *ii*) Machine learning techniques using features developed by linguists for Basque (Bengoetxea, Atutxa, and Iruskieta, 2017). In both approaches, evaluation measures are based in annotated data, where the CU was considered in annotation guidelines.

In these works, authors found that annotating or detecting the CUs is a genre and domain oriented classification task. Some features which work very well for scientific abstracts genre, do not work for argumentative answer texts genre, or vice versa. Therefore, developing a general good CU detector is a complicated task, because following these approaches a linguist is needed to annotate the corpus and to extract the features manually for each genre or domain (and language).

The work presented here is different from the previous works, because of these two reasons: 1) the features to detect the CU are extracted automatically and 2) the corpus, the RST Spanish Treebank, employed in this work was annotated with rhetorical relations, following typical two step annotation methodology: *i*) EDU segmentation and *ii*) rhetorical relation labeling (da Cunha, Torres-Moreno, and Sierra, 2011). Therefore, it was annotated without taking into account the CU constraints in the annotation guidelines.[4]

The method employed in this work will be useful to detect the CU in other languages, genres and domains with less effort, if RST annotated data is available. The CU detector can be useful in several NLP tasks, such as sentiment analysis (to identify the most important evaluative sentence (Alkorta et al., 2017)), annotation of the rhetorical RS-trees (Iruskieta, de Ilarraza, and Lersundi, 2014) or to improve some parsers or prototypes (da Cunha et al., 2012).

3 Methodology

As we noted previously, there is not an annotated corpus with CUs for Spanish, but we extract the root of the rhetorical trees and label as CU. So, in order to build the CU detector we follow the subsequent phases.

3.1 Source for corpus compilation

The corpus we have used for such task is the RST Spanish Treebank (da Cunha, Torres-Moreno, and Sierra, 2011) which is the first corpus annotated with rhetorical relations for Spanish. The corpus is annotated with specialized texts of 9 domains: *i*) Astrophysics, *ii*) Earthquake Engineering, *iii*) Economy, *iv*) Law, *v*) Linguistics, *vi*) Mathematics, *vii*) Medicine, *viii*) Psychology and *ix*) Sexuality.

3.2 Selected corpus

To ensure the compilation of the corpus we check if every text was organized as follows: *i*) If all the text has a title at the beginning of the document. *ii*) If the text was long enough (most of the texts of the same domain has more than 4 EDUs). *iii*) If the extracted CU from the RS-tree is reliable.

We found that a lot of texts of different domains do not fulfill these constraints, so we selected the best two domains that fulfill these constraints: *i*) Psychology and *ii*) Linguistics.

[3]Basque corpus is composed with different domains, in the same genre.

[4]In the studies previously mentioned, the CU constraints were considered in the annotation process.

In one of these domains, we detect that the linguistic texts lack the title (4 of them) and some CU (7 text of 45) were wrongly annotated (and wrongly extracted),[5] once we compared with our CU annotation guidelines (Iruskieta, de Ilarraza, and Lersundi, 2014).[6]

Therefore, when an inconsistency in the annotation was found, the entry was fully examined, the title was added and the extracted CU was changed in our database.

After this process, the corpus description used in this paper is presented in Table 1 describing the two domains (Dom.): Psychology (PS) and Linguistics (LI), texts (T), words (W), Elementary Discourse Units (EDU) and Central Units (CU).

Dom.	T	W	EDU	CU
PS	28	4409	274	36
LI	45	11176	599	51
Total	73	15585	873	87

Table 1: Corpus description

The gold standard we created contains 873 EDUs and 73 texts, each with its CU.

The amount of texts of this study is smaller than previously used for similar tasks (Bengoetxea, Atutxa, and Iruskieta, 2017; Iruskieta, Antonio, and Labaka, 2016; Burstein et al., 2001).

3.3 Preprocessing

The steps to preprocess the data are the following:

i) Data. We extract EDUs and CUs from the annotated Spanish RST Treebank (da Cunha, Torres-Moreno, and Sierra, 2011). The gold standard segmented corpus was annotated automatically with morphosyntactic information using FreeLing (Carreras et al., 2004).

ii) Database. The database was created with the gold standard files.

iii) Data-sets. This corpus was divided into 2 non-overlapping datasets as we show in Table 2: 60 texts as a training dataset (Train) and 13 texts as test dataset (Test). So, we used 20% of the data for testing and rest of the 80% for training. To estimate the performance of our systems and to select the best classifier, we use a 10-fold cross-validation procedure: the 60 texts of the train dataset were partitioned randomly into 10 groups and we train 10 times on 9/10 of the labeled data and we evaluate the performance on the other 1/10 of the data.
Table 2 reports some information about the 2 non-overlapping datasets, measures (T for texts, EDUs, CUs) and difficulty (Diff.), multiple CUs (M) and texts where the CU is in the first EDU (F).[7]

Set	T	EDU	CU	Diff.	M	F
Train	60	621	69	0.111	8	25
Test	13	183	14	0.076	1	6
Total	73	804	83			

Table 2: Data-set information

The task's difficulty to find the CU has been calculated as follows: $Difficulty = \frac{CUs}{EDUs}$ where the nearer it is from 1 the easier it is to determine the CU.
Test dataset is more difficult, because difficulty is farther from 1 to determine the CU. While the proportion of multiple CUs (M) and the EDU position of the CUs (F) are similar in both dataset.

iv) Classification tasks. All the data we prepare was performed using Perl scripts and Weka workbench (automatic feature extraction with bag of words).

- We converted each segment words into a set of attributes representing word occurrence information and we created a set of 1000, 5000 and 15000 words (attributes) using the training data. We represented each segment by an array of lemmas.
- We convert all letters to lower case.
- We followed bag of words approach and used tokens (unigrams, bigrams and trigrams) as features, where a classification instance is a vector of tokens appearing in the segmented text.[8]
- We also added EDU position and title word occurrence information to the feature vector. Thus, there was

[5] We think that this is due to that the CU was not considered in the annotation guidelines.

[6] The psychology texts were formated as well as we need.

[7] Multiple CUs (M) are the most difficult to detect by automatic means, whereas texts that the CU is in the first EDU (F) are the easiest to detect.

[8] We tried removing all words without linguistic meaning using a list of Spanish stop words (this list can be consulted at http://members.unine.ch/jacques.savoy/clef/), but the results were worse.

no attempt to remove or normalize them. Using weka's "string to word vector", text was converted into feature vector using TF-IDF (Manning, Raghavan, and Schtze, 2008) as feature value.
– Finally, the training set dictionary obtained using this scheme contains 1000 features; the same dictionary was used for the test set. TF-IDF feature valued representation was selected for Sequential Minimal Optimization (SMO) (Platt, 1998) and Multinomial Naive Bayes (MNB) (McCallum, Nigam, and others, 1998) systems, and boolean feature valued representation for Bernoulli Naive Bayes (BNB) (John and Langley, 1995) system.

3.4 Automatic feature selection

Feature selection is classic refinement method in classification. It is an effective dimensionality reduction technique to remove noise feature. In general, the basic idea is to search through all possible combinations of attributes in the data to find which subset of features works best for prediction. Removal is usually based on some statistical measures, such as segment frequency, information gain, chi-square or mutual information.

In this research, we have tested the two most effective feature selection methods: *i*) chi-square and *ii*) information gain using different set of attributes: 50, 100, 500 and 1000. Finally we performed all the classifiers using chi-square with a set of 100 attributes.

3.5 Classification

Classification was perform using WEKA workbench, to choose the best system. In our experiment we used 3 types of classifiers: *i*) Sequential Minimal Optimization (SMO),*ii*) Multinomial Naive Bayes (MNB) and *iii*) Bernoulli Naive Bayes (BNB).

In machine learning, Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between features.

The reasons to choose Naive Bayes models are:

– They only require a small amount of training data to estimate the parameters necessary for classification.
– They have been used successfully in similar tasks: for identifying thesis statements (Burstein et al., 2001) or for classifying short texts (McCallum, Nigam, and others, 1998).
– They can be used as predictive and descriptive model.

We have implemented three different ML methods:

– MNB. Multinomial Naive Bayes implements the naive Bayes algorithm for multinomially distributed data, and it is one of the two classic naive Bayes variants used in text classification (where the data is typically represented as word vector counts, although TF-IDF vectors are also known to work well in practice).
– BNB. Bernoulli Naive Bayes approach is the other classic naive Bayes variant. BNB trains classifiers on the absence and presence of features and using this information we can build a model to classify or select from a text the EDU that is the most likely candidate to be labeled as CU.
– SMO. Sequential Minimal Optimization is an optimization technique for solving quadratic optimization problems, which arise during the training of SVM and it has better generalization capability. Another reason for SMO is the high classification accuracy on different tasks reported in the literature (Schuller et al., 2012; Mairesse et al., 2007; Kermanidis, 2012) on personality traits recognition.

3.6 Evaluation

As a performance measure we used the average performance of our classifier using traditional recall (Rec.), precision (Prec.), and F-score (F_1) metrics. F-score was calculated with the standard measures as follows:

$$Prec. = \frac{correct_{CU}}{correct_{CU} + excess_{CU}}$$

$$Rec. = \frac{correct_{CU}}{correct_{CU} + missed_{CU}}$$

$$F_1 = \frac{2 * Prec. * Rec.}{Prec. + Rrec.}$$

where $correct_{CU}$ is the number of correct central units, $excess_{CU}$ is the number of over-predicted central units and $missed_{CU}$ is the

System	Data	C	E	M	P	R	F_1
Baseline	Train	34	26	35	0.492	0.566	0.527
	Test	6	7	8	0.428	0.461	0.444
BNB	Cross	51	39	18	0.566	0.739	0.641
	Test	11	6	3	0.647	0.785	0.709
MNB	Cross	58	22	11	0.725	0.841	0.779
	Test	11	3	3	0.785	0.785	0.785
SMO	Cross	50	5	19	0.909	0.725	0.806
	Test	11	4	3	0.733	0.786	0.759

Table 3: Results obtained on cross-validation and test sets

number of central units the system missed to tag.

We have compared the results of 3 systems against a simple baseline to detect the CU. This baseline is based on the position of the given EDU into the whole document.[9] The position is an important indicator, because we found that the likelihood of a CU occurring at the beginning of the text was 49.27% in the training set. So we consider that the first segment is the only CU of the text as our baseline.

The choice of algorithms is driven by their different properties for classification. Results were calculated as average of 10 experiments using 10-fold cross-validation and we compare the results of all the system in a box plot.[10] After that, we use the best system to extract the CUs of the test dataset. Results and error analysis are evaluated in this test dataset (see Subsection 4.2).

4 Results

Table 3 shows the results obtained using *i*) a baseline, *ii*) three different machine learning methods: BNB, MNB and SMO.

We can observe that SMO and MNB systems are better than baseline and BNB systems. The best model of the Table 3 is SMO which provides 0.806 in cross-validation and 0.759 in test.

Table 3 shows that SMO system is better than MNB system in 0.027 points in cross-validation, but in test dataset SMO system is worse than MNB system in 0.026 points. In the next subsection we compare all the systems in more detail.

4.1 A comparison using box plot

To show how robust the systems are on the dataset we run 10-fold cross-validation 10 times. The training dataset was randomly broken into 10 partitions using 10 random seeds. We have calculated 10 means of the F-score value for each 10-fold cross-validation (see Figure 1).

To visualize the performance of the 4 systems (Baseline, BNB, MNB and SMO), we have summarized the distribution of F-score values using box plots (Chambers, 1983).

Figure 1 shows the following main results:

- SMO and MNB classifiers show a greater F-score median value than BNB and Baseline F-score value.
- The best systems are SMO and MBM systems which has the same median value of F-score.
- And finally we can see that SMO is slightly better than MBM system because the upper and lower quartiles are slightly upper.

To understand how the CU detector works, we present the results obtained in the test dataset and an error analysis in the following subsection.

4.2 Error analysis

We analyze in Table 4 the results obtained with the best system from manual segmentation (SMO Gold) extracted from RST Spanish Treebank and from automatic segmentation (SMO Auto) performed with DiSeg (da Cunha et al., 2010) and we describe why SMO does not detect correctly some of these CUs from the test dataset.

The SMO Gold system has selected 5 TP (true positive) at the beginning of the text,

[9] Other baselines with linguistic features can be tested but we excluded then, because this is out of the objectives assigned to the study.

[10] A box plot consists of a box summarizing 50% of the data. The upper and lower ends of the box are the upper and lower quartiles, while a thick line within the box encodes the median. Dashed appendages summarize the spread and shape of the distribution, and dots represent outside values (see Figure 1).

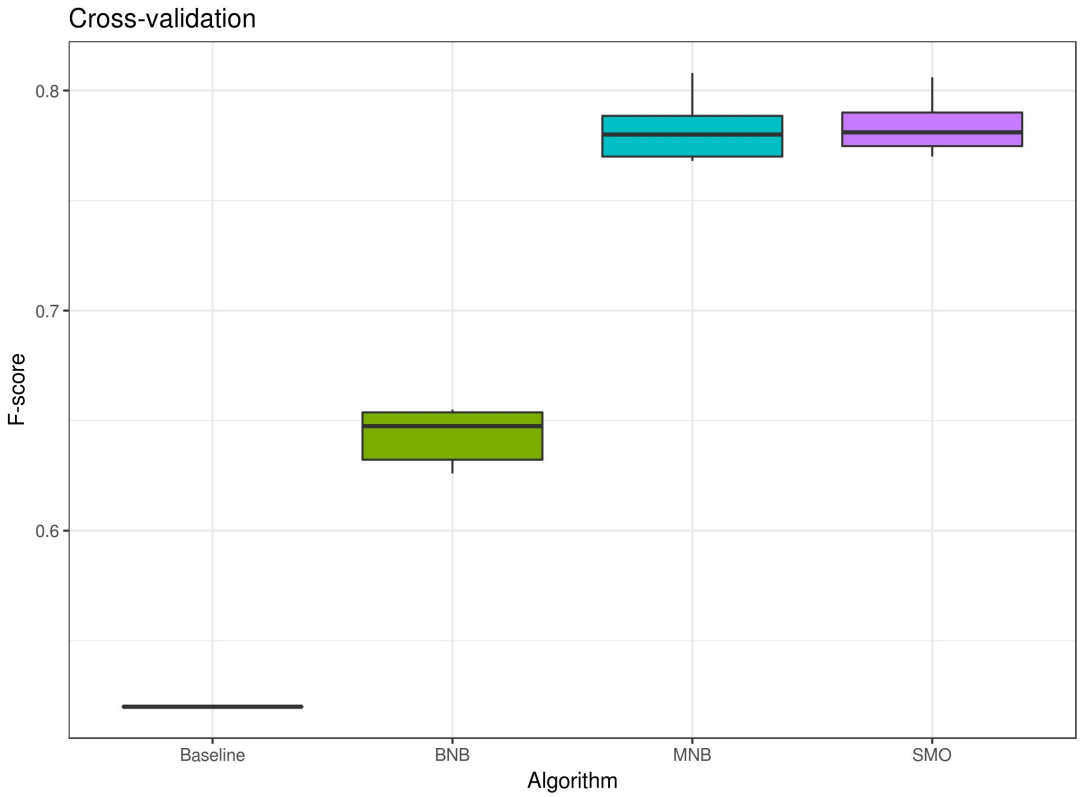

Figure 1: Exploring F-score distribution on the 10-fold cross-validation using 10 random seeds with Box Plot

System	Correct	Something wrong	
	Total agreem.	Partial agreem.	Total disagr.
SMO Gold	8	3	2
SMO Auto	8	2	3

Table 4: SMO's error analysis of the test dataset

2 TP at the middle and 1 TP at the end. Example 1 shows one CU that was found by the system.

(1) [el propósito de esta comunicación es hacer una reflexión sobre los retos a que se está enfrentando la neología terminológica en la realidad actual ;]$_{EDU2}$

Regarding the partial agreements, *i*) the system did not detect properly a CU at the end of the text, because it has selected 1 TP and another EDU as CU candidate (1 FP, false positive), that was some EDUs before, towards the middle of the text. *ii*) Another example that the system did not detect properly was a CU at the beginning of the text, because it has selected the CU (1 TP) and also other two false candidates (EDUs) at the end of the text (2 FPs). *iii*) The last one that the system has detected 1 TP of a multiple CU and did not detect the other EDU as a CU candidate (1 TN, true negative). This example of a partial agreement is presented in Example 2, in where the $EDU4$ was detected, but not the $EDU5$, which is in a clear conjunction.

(2) [el objetivo de el presente artículo es ; a_través_de un instrumento de evaluación de papel y lápiz ; evaluar el tipo de vínculo en la adolescencia]$_{EDU4}$ [y hacer correlaciones entre las calificaciones de la niñez y la adolescencia con_respecto_a el tipo de vínculo y las relaciones de pareja ;]$_{EDU5}$

Finally, the total disagreements were because the system could not detect a CU that was not indicated or written in a proper way. *i*) One of them, is at the end of the text and the CU is an intrasentential EDU. *ii*) The other has to objectives and the CU is an intrasentential EDU. Example 3 shows an example where the ML techniques ($EDU3$) dis-

agree with the Gold Standard ($EDU9$).

(3) [el objetivo de nuestro proyecto es crear herramientas de aprendizaje de la lengua para estudiantes de formación profesional en las áreas de informática ; secretariado y electrónica]$_{EDU3}$ (...)
[nuestro artículo propone una metodología para la creación de una terminología plurilingüe]$_{EDU9}$

The results with the segmenter (SMO Auto) are only slightly worse (Table 4) and, therefore, we think that are acceptable. The small difference is that the system could not choose one CU that was partially correct in SMO Gold.

5 Discussion

An interesting point of this work is that in the annotation process of the Spanish RST Treebank (similar to the annotation of other RST Treebanks, such as Marcu (2000)) the CU was not considered during the annotation process. This will support, in such a sense, the claim that the CU is crucial point in RS-tree annotation, even when it is not considered in annotation guidelines.

In this paper we have introduced the first CU detector for Spanish[11] using SMO machine learning techniques without any linguistic design of features or rules in two subcorpus of the Spanish RST Treebank. The limitation of this work is that we could not use all the Spanish RST Treebank, due to some corpus formating constraints we think that are necessaries to develop CU detectors: *i*) text size and *ii*) title-body format of texts.

The experiments carried out on the corpus show competitive and promising results given the simplicity of the proposed method, which can be applied to different domains, if we have annotated RST treebank or a corpus partially annotated with discourse segments (EDUs) and CUs.

We are currently working to achieve the following aims:

- To reuse these techniques with other annotated data in different languages.
- To integrate the segmenter Diseg (da Cunha et al., 2010) and the CU detector for Spanish and follow up to detect some signaled discourse relations, to parse plain texts in Spanish and other languages as Basque, for example.

[11] A demo of the system can be tested here: http://ixa2.si.ehu.es/clarink/tools/ES-CU-detector/.

References

Alkorta, J., K. Gojenola, M. Iruskieta, and M. Taboada. 2017. Using lexical level information in discourse structures for basque sentiment analysis. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms, pages 39–47, Santiago de Compostela, Spain, September 4 2017. ACL.*

Bengoetxea, K., A. Atutxa, and M. Iruskieta. 2017. Un detector de la unidad central de un texto basado en técnicas de aprendizaje automático en textos científicos para el euskera. *Procesamiento del Lenguaje Natural*, 58:37–44.

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

Carreras, X., I. Chao, L. Padró, and M. Padró. 2004. Freeling: An open-source suite of language analyzers. In *LREC*.

Chambers, J. M. 1983. *Graphical methods for data analysis*. Wadsworth Belmont, CA.

Cortes, C. and V. Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

da Cunha, I., E. SanJuan, J.-M. Torres-Moreno, M. T. Cabré, and G. Sierra. 2012. A symbolic approach for automatic detection of nuclearity and rhetorical relations among intra-sentence discourse segments in spanish. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 462–474. Springer.

da Cunha, I., E. SanJuan, J.-M. Torres-Moreno, M. Lloberas, and I. Castellón. 2010. Diseg: Un segmentador discursivo automático para el español. *Procesamiento del Lenguaje Natural*, 45:145–152.

da Cunha, I., J.-M. Torres-Moreno, and G. Sierra. 2011. On the Development of the RST Spanish Treebank. In *5th*

Linguistic Annotation Workshop (LAW V '11), pages 1–10, Portland, USA, 23 June. ACL.

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

Iruskieta, M., M. Aranzabe, A. Diaz de Ilarraza, I. Gonzalez, M. Lersundi, and O. L. 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. D. de Ilarraza, and M. Lersundi. 2014. The annotation of the central unit in rhetorical structure trees: A key step in annotating rhetorical relations. In *COLING*, pages 466–475.

John, G. H. and P. Langley. 1995. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc.

Kermanidis, K. L. 2012. Mining authors' personality traits from modern greek spontaneous text. In *Proc. of Workshop on Corpora for Research on Emotion Sentiment & Social Signals, in conjunction with LREC*, pages 90–93. Citeseer.

Mairesse, F., M. A. Walker, M. R. Mehl, and R. K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.

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

Manning, C. D., P. Raghavan, and H. Schtze. 2008. Relevance feedback and query expansion. *Introduction to Information Retrieval. Cambridge University Press, New York.*

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

McCallum, A., K. Nigam, et al. 1998. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Madison, WI.

Platt, J. 1998. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report. MSR-TR-98-14. Microsoft Research.

Schneider, K.-M. 2005. Techniques for improving the performance of naive bayes for text classification. *Computational Linguistics and Intelligent Text Processing*, pages 682–693.

Schuller, B. W., S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. Van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss. 2012. The interspeech 2012 speaker trait challenge. In *Interspeech*, volume 2012, pages 254–257.

Procesamiento del Lenguaje Natural, Revista nº 60, marzo de 2018, pp. 37-44 recibido 20-11-2017 revisado 04-01-2018 aceptado 02-02-2018

The democratization of deep learning in TASS 2017

La democratización del aprendizaje profundo en TASS 2017

Manuel C. Díaz-Galiano[1], Eugenio Martínez-Cámara[2], M. Ángel García-Cumbreras[1], Manuel García-Vega[1], Julio Villena-Román[3]
[1]Advanced Studies Center in Information and Communication Technologies (CEATIC), University of Jaén, Jaén, Spain
[2]Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain
[3]MeaningCloud, Madrid, Spain
[1]{mcdiaz, magc, mgarcia}@ujaen.es,
[2]emcamara@decsai.ugr.es, [3]jvillena@meaningcloud.com

Abstract: TASS 2017 has brought advances in the state-of-the-art in Sentiment Analysis in Spanish, because most of the systems submitted in 2017 were grounded on Deep Learning methods. Moreover, a new corpus of tweets written in Spanish was released, which is called InterTASS. The corpus is composed of tweets manually annotated at document level. The analysis of the results with InterTASS shows that the main challenge is the classification of tweets with a neutral opinion and those ones that do not express any opinion. Likewise, the organization exposed the project of extending InterTASS with tweets written in different versions of Spanish.
Keywords: TASS, sentiment analysis, deep learning, linguistic resources

Resumen: TASS 2017 ha vuelto a suponer un avance en el estado del arte de análisis de opiniones en español, debido a la exposición de sistemas mayoritariamente fundamentados en métodos de Deep Learning. Además, en esta edición se ha presentado un nueva colección de tuits en español manualmente etiquetados a nivel de documento y que se llama InterTASS. El análisis de los resultados con InterTASS demuestra que en el futuro el esfuerzo investigador se tiene que centrar en la distinción del nivel de intensidad de opinión neutro y la ausencia de opinión. Asimismo, se presentó el proyecto de ampliar el nuevo corpus con tuits escritos en el español que se habla en España y en algunos países de América.
Palabras clave: TASS, análisis de opiniones, aprendizaje profundo, recursos lingüísticos

1 Introduction

After sixth editions, the Workshop Sentiment Analysis at SEPLN (TASS) has become the reference workshop for the research community on Sentiment Analysis (SA) for the Spanish language in microblogs, specifically in Twitter. The main contribution of TASS is the progress of the state-of-the-art as can be read in Villena-Román et al. (2013), Villena-Román et al. (2014), Villena Román et al. (2015) and Martínez Cámara et al. (2016).

The success of TASS may be attributed to: 1) the generation and release of newly annotated corpora in every edition; 2) the organization of competitive evaluations in which the participants submit their systems, which are ranked according to their performance; and 3) the active involvement of the research community in the discussion about the main features of the submitted systems and the state-of-the-art in SA in Spanish, and setting up the challenges for the next edition.

Spanish is the second most widely-spoken language in the world, and it is mainly spoken in Spain and America. Although the language is the same, there exist several varieties with specific lexical and semantic differences among different geographical areas, namely Spain and American countries. Consequently, we set up the project of generating a new corpus of tweets for SA with the novelty of including tweets written in the different varieties of Spanish.

ISSN 1135-5948. DOI 10.26342/2018-60-4
© 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

In this paper, the first release of the International TASS Corpus, called InterTASS, is presented. The first version is only composed of tweets written in the Spanish spoken in Spain, but, in contrast to the General Corpus of TASS, InterTASS was manually annotated. Further details about the annotation of the corpus are described in section 2.

TASS 2017 proposed two subtasks: Task 1, polarity classification at document (tweet) level; and Task 2, polarity classification at aspect level (see section 3). Eleven teams from Spain and America submitted several systems and a description paper. Most of the systems are based on the use of Deep Learning methods. Some of them attempted to improve the results using ensemble classifiers. In this paper, we also depict the main features of the best submitted systems and analyse their results (see section 4).

2 Resources

TASS 2017 provided four datasets to the participants for the evaluation of their systems. Three of them were already used in previous editions: General Corpus, Social-TV Corpus, and STOMPOL. A new dataset, InterTASS, was created for Task 1 in TASS 2017.

2.1 InterTASS

The International TASS Corpus (*InterTASS*) is a new corpus released in TASS 2017 for the polarity classification at tweet level in Task 1. This is the first version and includes tweets posted in Spain, all of them are written in the Spanish variety spoken in Spain. The final version of the corpus will be composed of tweets written in the variety of Spanish spoken in different Spanish-speaking countries in America.

In order to prepare this version, over 500,000 tweets were downloaded between July 2016 and January 2017 using some keywords. These tweets were filtered according to the following requirements: 1) tweets should be written in Spanish,[1] 2) each tweet should have at least one adjective, 3) the minimum length of tweets should be four words.

Eight subsets were prepared, sorting the tweets according to their number of words. Using these subsets, the final collection was created by randomly selecting a homogeneous number of tweets from each subset, 3,413 tweets in total.

[1] `langdetect` Python library was used to check.

The annotation process was made by five annotators using a scale of four levels of polarity for the global sentiment of the tweet: positive (P), negative (N), neutral (NEU) and no sentiment (NONE). Tweets were evenly distributed, so that each tweet was annotated by at least three annotators. The annotation guidelines regarding the assignment of the label of each tweet were:

- A label is assigned to a tweet when at least two annotators are agree.
- In case the three annotators are not agree, the other two ones, who are different from the first three, annotate the tweet.
- If the tie persisted, the conflicting annotator decided the label of the tweet.

Each tweet includes its ID (`tweetid`), the creation date (`date`) and the user ID (`user`). Restrictions in the Twitter API Terms of Service,[2] do not allow to release a corpus that includes text contents or information about users. The actual message content of tweets can be obtained by making queries to the Twitter API using the `tweetid`. The corpus is in XML, and Figure 1[3] shows a sample tweet[4].

```
<tweet>
  <tweetid>[ID]</tweetid>
  <user>[USER NAME]</user>
  <content>y lo peor de todo es que
      funcionaba maldita Jaco como te
      quiero </content>
  <date>[DATE]</date>
  <lang>es</lang>
  <sentiment>
    <polarity>
      <value>NEU</value>
    </polarity>
  </sentiment>
</tweet>
```

Figure 1: A tweet in the InterTASS corpus

Finally, the corpus was divided into three datasets: Training, Development and Test. The Training and Development sets were released with the annotations, so the participants could train and validate their models.

[2] `https://dev.twitter.com/terms/api-terms`

[3] The `tweetid`, the `user`, the `date` fields are hidden because of the Twitter term of service.

[4] In English: The worst is that it worked, fucking Jaco I love you too much.

The test corpus was provided without any annotation and was used to evaluate systems. Statistics are shown in Table 1.

	Training	Dev.	Test
P	317	156	642
N	416	219	767
NEU	133	69	216
NONE	138	62	274
Total	1,008	506	1,899

Table 1: Number of tweets per dataset and class in the InterTASS corpus

2.2 General corpus

The General Corpus has been used since the first edition of TASS. It has about 68,000 tweets, written in Spanish by about 150 well-known personalities and celebrities of the world of politics, economy, communication, mass media, and culture, between November 2011 and March 2012. The details of the corpus are described in Villena-Román et al. (2015) and García-Cumbreras et al. (2016).

This corpus was divided into training set (10%) and test set (90%). Each tweet in the training set was annotated with its global polarity in a scale of six intensity levels: strong positive (P+), positive (P), neutral (NEU), negative (N), strong negative (N+) and no sentiment (NONE). The test set was annotated by a meta-classifier based on majority voting, using as base classifiers the submitted systems of previous editions of TASS.

In addition, a selected subset containing 1,000 tweets with a similar class distribution than the training set was extracted in 2015 edition and manually annotated for an additional evaluation of the systems (1k test set).

2.3 Social-TV Corpus

The Social-TV corpus was released in TASS 2014. The tweets were gathered during the 2014 Final of Copa del Rey championship in Spain between Real Madrid and F.C. Barcelona. After filtering out useless tweets, a subset of 2,773 tweets was selected. Further details in Villena-Román et al. (2015), and García-Cumbreras et al. (2016).

The sentiment was manually annotated at aspect level (31 aspects), using only 3 levels of opinion: positive (P), neutral (NEU) and negative (N). The corpus was randomly split into two subsets: training and test (1,773 and 1,000 tweets, respectively).

2.4 STOMPOL

STOMPOL (Spanish Tweets for Opinion Mining about POLitics), released in TASS 2015, is a corpus of Spanish tweets for SA at aspect level. The tweets were gathered from the 23rd to the 24th of April of 2015 during the Spanish political campaign of regional and local elections. Each tweet was manually annotated at aspect level by two annotators, and a third one in case of disagreement. The topics of the tweets are: economics, health system, education, political party, electoral system or environmental policy.

The corpus is composed of 1,284 tweets, and was also divided into training set (784 tweets) and test set (500 tweets). Further details in Villena-Román et al. (2015), García-Cumbreras et al. (2016).

3 Tasks

TASS 2017 proposed two tasks addressing the main challenges of SA in Twitter in Spanish.

3.1 Task 1. Sentiment Analysis at Tweet level

This main task focused on the evaluation of polarity classification systems at tweet level in Spanish. Systems were evaluated on three different datasets: two versions of the General Corpus (the complete test set and the 1k test set), and the new InterTASS corpus.

Participants had to identify the intensity of the opinion expressed in each tweet in any of the 4 intensity levels of polarity in which the datasets were annotated. For the two sets of the General Corpus, which were originally annotated in 6 polarity classes, a direct translation to 4 classes (P+ changed to P and N+ to N) was performed so that the evaluation was consistent with InterTASS.

The three datasets were divided into training, development and test datasets, which were provided to participants in order to train and evaluate their systems. Systems were allowed to use any set of data as training dataset, i.e., the training set of InterTASS, other training sets from the previous editions of TASS or other sets of tweets. However, using the test set of InterTASS and the test set of the datasets of previous editions as training data was obviously forbidden. Apart from that, participants could use any kind of linguistic resource for the development of their classification model.

Participants were expected to submit 3 experiments per each test set, so each participant team could submit a maximum of 9 files of results. Accuracy and the macro-averaged versions of Precision, Recall and F1 were used as evaluation measures. Systems were ranked by the Macro-F1 and Accuracy measures.

3.2 Task 2. Aspect-based Sentiment Analysis

This second task proposed the development of aspect-based polarity classification systems. Two datasets from previous editions were used to evaluate the systems: Social-TV and STOMPOL (see section 2). The aspect, the main category of the aspect, and the opinion in three intensity levels (P, N, textscneu) were annotated in the two datasets.

Participants were expected to submit up to 3 experiments for each corpus. For evaluation, exact match with a single label combining "aspect-polarity" was used. The evaluation measures were the same as in Task 1.

4 Analysis of Submissions

In TASS 2017, the following 11 different groups presented their runs in the tasks:

- ELiRF, Universidad Politécnica de Valencia (Spain)
- RETUYT, Universidad de la República, Montevideo (Uruguay)
- ITAINNOVA, Zaragoza (Spain)
- jacerong, Santiago de Cali (Colombia)
- INGEOTEC, Universidad Panamericana (Mexico)
- tecnolengua, Universidad de Málaga (Spain)
- SINAI, Universidad de Jaén (Spain)
- LexFAR, Universidad Autónoma Metropolitana (Mexico)
- OEG, Universidad Politécnica de Madrid (Spain)
- GSI, Universidad Politécnica de Madrid (Spain)
- C100T-PUCP, Universidad Católica del Perú (Peru)

It must be pointed out that five groups are from countries other than Spain, so the workshop is relevant in other American countries. Table 2 shows the participation of each group in the TASS 2017 tasks: 1I (Task 1, InterTASS corpus), 1G (Task 1, General corpus), 2SO (Task 2, Social-TV corpus) and 2ST (Task 2, STOMPOL corpus).

	1I	1G	2SO	2ST
ELiRF	X	X	X	X
RETUYT	X	X	X	X
ITAINNOVA	X			
jacerong	X	X		
INGEOTEC	X	X		
tecnolengua	X	X		
SINAI	X			
LexFAR	X			
OEG	X	X		
GSI	X	X		
C100T-PUCP				X
Total	10	7	2	2

Table 2: Groups and tasks

Most of the systems were based on Deep Learning techniques, but there were solutions based on traditional machine learning methods and meta-classifiers.

Hurtado, Pla, and González (2017) (ELiRF) created a set of domain-specific word embeddings following the approach of Tang (2015) for tasks 1 and 2. The former word embeddings set is jointly used with a general-domain set of embeddings to represent each token. They evaluated three different neural networks architectures: multilinear perceptron (MLP), convolutional recurrent neural network (CNN) and long-short term memory (LSTM) recurrent networks (RNN).

Cerón-Guzmán (2017) (jacerong) presented an ensemble classifier system for the first task. Their system generated quantitative features from the tweets (the number of words in upper case, the number of words with repeated letters, etc.), and then they used lists of opinion bearing words (iSOL (Molina-González et al., 2013)), as well as the inversion of the polarity of words following a window shifting approach for negation handling. The base classifiers of the ensemble system were Logistic Regression and SVM.

Montañés Salas et al. (2017) (ITAINNOVA) used the FastText classifier (Joulin et al., 2016) for the InterTASS dataset. After a traditional pre-processing to the input tweets, the system substituted the words with an emotional meaning by their synonyms from a list of words with an emotional

meaning (Bradley and Lang, 1999).

Rosá et al. (2017) (RETUYT) used three different approaches: an SVM classifier with word embeddings and quantitative linguistic properties as features; a deep neural network grounded on the use of a CNN for encoding the input tweets; and the combination of the two previous classifiers by the selection of the output class with a higher probability mean from the two previous classifiers.

García-Vega et al. (2017) (SINAI) used for InterTASS an SVM classifier that uses word-embeddings as features. They introduced the use of the language of each user in the classification. Other approaches are based on deep neural networks grounded on the use of LSTM RNN for the encoding of the meaning of the input tweets.

The approach by Moctezuma et al. (2017) (INGEOTEC) was an ensemble of SVM classifiers combined into a non-linear model created with genetic programming to tackle the task of global polarity classification at tweet level. They used B4MSA algorithm, a proposed entropy-based term-weighting scheme, which is a baseline supervised learning system based on the SVM classifier, an entropy-based term-weighting scheme. Additionally, they used EvoDAG, a GP system that combines all decision values predicted by B4MSA systems. They also used two external datasets to train the B4MSA algorithm.

Navas-Loro and Rodríguez-Doncel (2017) (OEG) used two classifiers: Multinominal Naïve Bayes and Sequential Minimal Optimization for SVM. Furthermore, they applied morphosyntactic analyses for negation detection, along with the use of lexicons and dedicated preprocessing techniques for detecting and correcting frequent errors and expressions.

Araque et al. (2017) (GSI) applied an RNN architecture composed of LSTM cells followed by a feed-forward network. The architecture makes use of two different types of features: word embeddings and sentiment lexicon values. The recurrent architecture allows to process text sequences of different lengths, while the lexicon inserts directly into the system sentiment information. Two variations of this architecture were used: an LSTM that iterates over the input word vectors, and a combination of the input word vectors and polarity values from a sentiment lexicon.

Tume Fiestas and Sobrevilla Cabezudo (2017) (C100T-PUCP) proposed for Task 2 an approach based on word embeddings for polarity classification at aspect-level. They used vectors of the words to measure their similarity and make a model to classify each polarity of each aspect for each tweet.

Reyes-Ortiz et al. (2017) (LexFAR) used, for Task 1, support vector machines algorithm and lexicons of semantic polarities at the level of lemma for Spanish. Features extracted from lexicons are represented by the bag-of-words model and they are weighted using Term Frequency measure at tweet level.

Moreno-Ortiz and Pérez Hernández (2017) (tecnolengua) proposed a classification model based on the Lingmotif Spanish lexicon with a number of formal text features, both general and CMC-specific, as well as single-word keywords and n-gram keywords. They used logistic regression classifier trained with the optimal set of features, SVM classifier on the same features set.

The fifteen best results reached by systems in Task 1, using the test sets of InterTASS and the General Corpus are showed in Tables 3, 4 and 5.

System	**M-F1**	**Acc.**
ELiRF-UPV-run1	0.493	0.607
RETUYT-svm_cnn	0.471	0.596
ELiRF-UPV-run3	0.466	0.597
ITAINNOVA-model4	0.461	0.576
jacerong-run-2	0.460	0.602
jacerong-run-1	0.459	0.608
INGEOTEC-evodag_001	0.457	0.507
RETUYT-svm	0.457	0.583
tecnolengua-sent_only	0.456	0.582
ELiRF-UPV-run2	0.450	0.436
ITAINNOVA-model3	0.445	0.561
RETUYT-cnn3	0.443	0.558
SINAI-w2v-nouser	0.442	0.575
tecnolengua-run3	0.441	0.576
tecnolengua-sent_only_fixed	0.441	0.595

Table 3: Task 1 InterTASS corpus, fifteen best results

Table 6 and Table 7 show the results reached by the submitted systems in Task 2, using the test sets of Social-TV corpus and STOMPOL corpus respectively.

System	M-F1	Acc.
INGEOTEC-evodag_003	0.577	0.645
jacerong-run-1	0.569	0.706
jacerong-tass_2016-run_3	0.568	0.705
ELiRF-UPV-run2	0.549	0.659
ELiRF-UPV-run3	0.548	0.725
RETUYT-svm_cnn	0.546	0.674
jacerong-run-2	0.545	0.701
ELiRF-UPV-run1	0.542	0.666
RETUYT-cnn	0.541	0.638
RETUYT-cnn3	0.539	0.654
tecnolengua-run3	0.528	0.657
tecnolengua-final	0.517	0.632
tecnolengua-531F1_no_ngrams	0.508	0.652
INGEOTEC-evodag_001	0.447	0.514
OEG-victor2	0.389	0.496

Table 4: Task 1 General Corpus (full test), fifteen best results

System	M-F1	Acc.
RETUYT-svm	0.562	0.700
RETUYT-cnn4	0.557	0.694
RETUYT-cnn2	0.555	0.694
INGEOTEC-evodag_003	0.526	0.595
tecnolengua-run3	0.521	0.638
ELiRF-UPV-run1	0.519	0.630
jacerong-tass_2016-run_3	0.518	0.625
jacerong-run-1	0.508	0.678
jacerong-run-2	0.506	0.673
ELiRF-UPV-run2	0.504	0.596
tecnolengua-final	0.488	0.618
tecnolengua-run4	0.483	0.612
ELiRF-UPV-run3	0.477	0.588
INGEOTEC-evodag_002	0.439	0.431
INGEOTEC-evodag_001	0.388	0.486

Table 5: Task 1 General Corpus (1k), fifteen best results

4.1 InterTASS Analysis

If the test set is grouped by the number of correct labels assigned by one or some of the submitted systems, the obtained results are shown in Figure 2. The test set is balanced according to complexity, and there are more than 10% of the tweets that are not correctly labelled by any system.

Figure 3 analyses the relation with the polarity label and the correct predictions. Most of the rightly predicted tweets are positive or negative, in contrast the submitted systems use to fail in the classification of NONE and NEU tweets.

System	M-F1	Acc.
ELiRF-UPV-run3	0.537	0.615
ELiRF-UPV-run2	0.513	0.600
ELiRF-UPV-run1	0.476	0.625
RETUYT-svm2	0.426	0.595
RETUYT-svm	0.413	0.493

Table 6: Task 2 Social-TV corpus results

System	M-F1	Acc.
ELiRF-UPV-run1	0.537	0.615
RETUYT-svm2	0.508	0.590
ELiRF-UPV-run3	0.486	0.578
ELiRF-UPV-run2	0.486	0.541
C100T-PUCP-run3	0.445	0.528
C100T-PUCP-run1	0.415	0.563
C100T-PUCP-run2	0.414	0.517
RETUYT-svm	0.377	0.514

Table 7: Task 2 STOMPOL corpus results

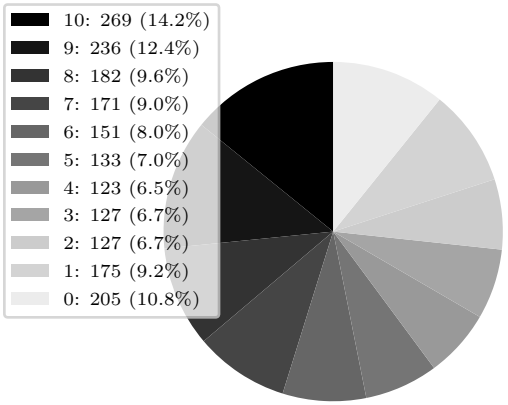

Figure 2: Systems that correctly classified a number of tweets (InterTASS corpus)

Last, we compared the statistics of the correct results with the number of words in tweets, as during the manual labelling, the annotators warned that tweets with a low number of words were noticeably more difficult to annotate. Table 8 shows the statistics. The first column shows different groups with the number of words of the tweets, and the other columns represent the number of systems that have hit the correct label. The percentage is calculated with the total number of tweets regarding the total value of the same column. Statistics are comparable, regardless of the number of words in the tweets, so, apparently, there is not a direct relation between the number of words of the tweets

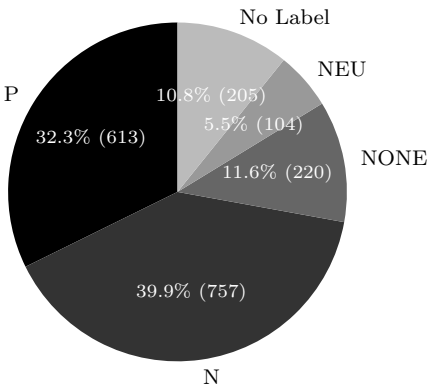

Figure 3: Polarity label and successful results (InterTASS corpus)

and the number of right predictions.

Words	0	1-5	6-9	10
0-4	20 (10%)	98 (14%)	101 (14%)	21 (8%)
5	14 (7%)	101 (15%)	87 (12%)	38 (14%)
6	32 (16%)	79 (12%)	92 (12%)	37 (14%)
7	29 (14%)	86 (13%)	81 (11%)	44 (16%)
8	24 (12%)	84 (12%)	88 (12%)	44 (16%)
9	32 (16%)	67 (10%)	107 (14%)	33 (12%)
11-18	25 (12%)	78 (11%)	97 (13%)	20 (7%)

Table 8: Correct tweet labels vs number of words (InterTASS corpus)

5 Conclusions and future work

The main objectives of TASS 2017 were: 1) to keep the interest of the research community in SA in Spanish; 2) the release of InterTASS, a new corpus for SA in Spanish; and 3) to forward the state-of-the-art through the debate of the features of the systems, most of them based on the use of Deep Learning methods and meta-classifiers.

We analyzed (see section 4) the performance of the submitted systems in the InterTASS corpus, and we conclude that there is room for improvement in the classification of the classes: NEU or NONE. Furthermore, no relation between the length of the tweets and the accuracy of the classification was found.

The work for further editions of TASS will be led by two goals. The first one is to broaden the number of tasks related to SA and semantic analysis with the aim of keep fostering the research in SA tasks in Spanish. The first milestone of the first goal was the update of the name of TASS to Workshop on Semantic Analysis at SEPLN in the edition of 2017. The second milestone will be the invitation to other research groups to organize and generate linguistic resources for SA tasks in Spanish. The second goal is to conclude the development of InterTASS with tweets written in the Spanish varieties of (as many as possible) Spanish speaking countries.

Acknowledgements

This research work is partially supported by REDES project (TIN2015-65136-C2-1-R) and SMART project (TIN2017-89517-P) from the Spanish Government, and a grant from the Fondo Europeo de Desarrollo Regional (FEDER). Eugenio Martínez Cámara was supported by the Juan de la Cierva Formación Programme (FJCI-2016-28353) from the Spanish Government.

References

Araque, O., R. Barbado, J. F. Sánchez-Rada, and C. A. Iglesias. 2017. Applying recurrent neural networks to sentiment analysis of spanish tweets. In *Proceedings of TASS 2017*, volume 1896 of *CEUR Workshop Proceedings*, Murcia, Spain, September. CEUR-WS.

Bradley, M. M. and P. J. Lang. 1999. Affective norms for english words (anew): Stimuli, instruction manual, and affective ratings. Technical report, Center for Research in Psychophysiology, University of Florida.

Cerón-Guzmán, J. A. 2017. Classier ensembles that push the state-of-the-art in sentiment analysis of spanish tweets. In *Proceedings of TASS 2017*.

García-Cumbreras, M. A., J. Villena-Román, E. Martínez-Cámara, M. C. Díaz-Galiano, M. T. Martín-Valdivia, and L. A. Ureña López. 2016. Overview of tass 2016. In *TASS 2016: Workshop on Sentiment Analysis at SEPLN*, pages 13–21.

García-Vega, M., A. Montejo-Ráez, M. C. Díaz-Galiano, and S. M. Jiménez-Zafra.

2017. Sinai en tass 2017: Clasificación de la polaridad de tweets integrando información de usuario. In *Proceedings of TASS 2017*.

Hurtado, L.-F., F. Pla, and J.-A. González. 2017. Elirf-upv en tass 2017: Análisis de sentimientos en twitter basado en aprendizaje profundo. In *Proceedings of TASS 2017*.

Joulin, A., E. Grave, P. Bojanowski, and T. Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Martínez Cámara, E., M. A. García Cumbreras, J. Villena Román, and J. García Morera. 2016. TASS 2015-The evolution of the spanish opinion mining systems. *Procesamiento del Lenguaje Natural*, 56(0):33–40.

Moctezuma, D., M. Graff, S. Miranda-Jiménez, E. S. Tellez, A. Coronado, C. N. Sánchez, and J. Ortiz-Bejar. 2017. A genetic programming approach to sentiment analysis for twitter: Tass'17. In *Proceedings of TASS 2017*, volume 1896 of *CEUR Workshop Proceedings*, Murcia, Spain, September. CEUR-WS.

Molina-González, M. D., E. Martínez-Cámara, M.-T. Martí-Valdivia, and J. M. Perea-Ortega. 2013. Semantic orientation for polarity classification in spanish reviews. *Expert Systems with Applications*, 40(18):7250 – 7257.

Montañés Salas, R. M., R. del Hoyo Alonso, J. Vea-Murguía Merck, R. Aznar Gimeno, and F. J. Lacueva-Pérez. 2017. FastText como alternativa a la utilización de deep learning en corpus pequeños. In *Proceedings of TASS 2017*.

Moreno-Ortiz, A. and C. Pérez Hernández. 2017. Tecnolengua lingmotif at tass 2017: Spanish twitter dataset classification combining wide-coverage lexical resources and text features. In *Proceedings of TASS 2017*.

Navas-Loro, M. and V. Rodríguez-Doncel. 2017. Oeg at tass 2017: Spanish sentiment analysis of tweets at document level. In *Proceedings of TASS 2017*, volume 1896 of *CEUR Workshop Proceedings*, Murcia, Spain, September. CEUR-WS.

Reyes-Ortiz, J. A., F. Paniagua-Reyes, B. Priego-Sánchez, and M. Tovar. 2017. Lexfar en la competencia tass 2017: Análisis de sentimientos en twitter basado en lexicones. In *Proceedings of TASS 2017*, volume 1896 of *CEUR Workshop Proceedings*, Murcia, Spain, September. CEUR-WS.

Rosá, A., L. Chiruzzo, M. Etcheverry, and S. Castro. 2017. Retuyt en tass 2017: Análisis de sentimientos de tweets en español utilizando svm y cnn. In *Proceedings of TASS 2017*.

Tang, D. 2015. Sentiment-specific representation learning for document-level sentiment analysis. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 447–452, New York, NY, USA. ACM.

Tume Fiestas, F. and M. A. Sobrevilla Cabezudo. 2017. C100tpucp at tass 2017: Word embedding experiments for aspect-based sentiment analysis in spanish tweets. In *Proceedings of TASS 2017*, volume 1896 of *CEUR Workshop Proceedings*, Murcia, Spain, September. CEUR-WS.

Villena-Román, J., J. García-Morera, M. A. García-Cumbreras, E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña López. 2015. Overview of tass 2015. In *TASS 2015: Workshop on Sentiment Analysis at SEPLN*, pages 13–21.

Villena-Román, J., J. García-Morera, S. Lana-Serrano, and J. C. González-Cristóbal. 2014. Tass 2013 - a second step in reputation analysis in spanish. *Procesamiento del Lenguaje Natural*, 52(0):37–44, March.

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

Villena Román, J., E. Martínez Cámara, J. García Morera, and S. M. Jiménez Zafra. 2015. Tass 2014 - the challenge of aspect-based sentiment analysis. *Procesamiento del Lenguaje Natural*, 54(0):61–68.

Procesamiento del Lenguaje Natural, Revista nº 60, marzo de 2018, pp. 45-52 recibido 20-11-2017 revisado 04-01-2018 aceptado 15-02-2018

Estudio preliminar de la anotación automática de códigos CIE-10 en informes de alta hospitalarios

Preliminary Study of the Automatic Annotation of Hospital Discharge Report with ICD-10 codes

Mario Almagro[1], **Raquel Martínez**[1], **Víctor Fresno**[1], **Soto Montalvo**[2]
[1]Universidad Nacional de Educación a Distancia (UNED), Juan del Rosal, 16, 28040 - Madrid
[2]Universidad Rey Juan Carlos (URJC), Tulipán, S/N, 28933 - Móstoles
{malmagro, raquel, vfresno}@lsi.uned.es, soto.montalvo@urjc.es

Resumen: En la actualidad, la cantidad de recursos que se destinan a la codificación de informes médicos es enorme. Con la reciente implantación del estándar CIE-10 en el sistema sanitario español se acrecientan las dificultades, ya que se incrementa el número de posibles códigos CIE por cada informe, disponiendo de una escasa cantidad de datos generados y codificados. En este artículo se describen los retos que plantea esta tarea y se propone una primera aproximación de combinación de técnicas para implantar un sistema capaz de, dado un informe, recomendar automáticamente códigos CIE-10 a los codificadores.
Palabras clave: Clasificación automática multietiqueta, códigos CIE-10, recuperación de información, divergencia de Kullback-Leibler

Abstract: Nowadays, the amount of resources dedicated to encode medical reports is huge. In spite of this, an automatic solution for annotation does not appear to be consolidated. With the recent implantation of the CIE-10 standard, encoding becomes increasingly complex, since the number of possible CIE codes for each report is increased and at the same time a minimal quantity of generated and encoded data is avaliable. In this paper, it is described the challenges posed by this task and proposed a first combination of techniques for implementing a system capable of automatically assisting coding specialists by recommending CIE-10 codes regarding a medical report.
Keywords: Automatic multilabel classification, ICD-10 codes, information retrieval, Kullback-Leibler divergency

1 Introducción

La codificación médica según la Clasificación Internacional de Enfermedades (CIE) consiste en la asignación de códigos estandarizados a informes médicos en representación de diagnósticos y procedimientos, cuya finalidad es la generación de estadísticas de morbilidad y mortalidad. En 2015 se publicó el Real Decreto 69/2015, por el que se establece la obligatoriedad de utilizar la nueva versión CIE-10-ES como sistema de codificación clínica a partir del 1 de enero de 2016 en España.

CIE-10 es una clasificación alfanumérica jerárquica que contiene entre 3 y 7 dígitos, y aporta una información clínica más detallada que la anterior codificación CIE-9-MC. El número de diagnósticos aumenta de 14.315 a 69.099, y el número de procedimientos de 3.838 a 72.000. Al mismo tiempo, su recopilación en la historia clínica digital requiere un mayor esfuerzo para los profesionales, aún no familiarizados con el nuevo estándar. En estos dos años, el enorme coste de los recursos que los centros hospitalarios dedican a la anotación manual (o mínimamente asistida) de códigos CIE-10 ha hecho evidente la necesidad de desarrollar herramientas automáticas que asistan en dicha tarea.

Sin embargo, el proceso de codificación automática no es trivial. Además del gran número de códigos existentes, los informes médicos están escritos en lenguaje natural, sujetos a la variabilidad inherente al lenguaje libre. Es común que los facultativos utilicen terminología distinta a la reflejada en las descripciones de los códigos CIE-10 para expresar información relacionada al desorden o procedimiento de forma más detallada. Asimismo, el texto presente en los informes puede incluir errores ortográficos, estructuras sintácticas incorrectas, uso de jerga, sinóni-

ISSN 1135-5948. DOI 10.26342/2018-60-5 © 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

mos, acrónimos y abreviaturas, generando ambigüedad, lo que dificulta aún más su procesamiento automático.

Un posible enfoque del problema es el establecimiento de correspondencias entre los códigos CIE-9 y CIE-10, aprovechando así las aproximaciones ya existentes a la anotación con códigos CIE-9. La administración sanitaria americana ha descrito equivalencias entre ambas versiones a través de los Mapeos GEMs (General Equivalence Mappings)[1]. Aunque éstos están disponibles en varias lenguas, la mayor exhaustividad y especificidad de los CIE-10 imposibilita una correspondencia directa entre ambas codificaciones para una parte considerable de los casos. Consecuentemente, los sistemas de anotación de códigos CIE-9 no son directamente adaptables al nuevo estándar, especialmente aquellos basados en reglas.

Otras aproximaciones abordan el problema por medio de enfoques basados en aprendizaje automático supervisado, por lo que requieren un entrenamiento con ejemplos manualmente anotados. Sin embargo, las anotaciones manuales disponibles indican que solo una pequeña fracción del total de códigos es utilizada con frecuencia, mientras que la inmensa mayoría de códigos o son poco frecuentes o no aparecen en ningún informe; dicha distribución limitará irremediablemente las aproximaciones basadas en enfoques supervisados. Aunque este problema se irá atenuando con el paso del tiempo a la par que aumenta la disponibilidad de documentos anotados, los centros hospitalarios necesitan soluciones tecnológicas efectivas a corto y medio plazo para asistir a los anotadores. Por todas estas razones, consideramos que es necesario explorar además tanto las aproximaciones basadas en aprendizaje semi-supervisado como los enfoques no supervisados.

El objetivo principal de esta propuesta es analizar la eficacia de los enfoques básicos; en concreto, se han comparado los enfoques supervisado y basado en Recuperación de Información (RI) aplicados al contexto de la codificación CIE-10 en español. Una vez analizados los resultados, planteamos una mejora del enfoque de RI a través del enriquecimiento de los índices, la selección de diferentes elementos estructurales en los informes médicos y la combinación de distintos tipos de consultas.

En el segundo apartado del artículo se revisa el estado del arte, mientras que el tercero describe las características de la colección utilizada en la experimentación. En el apartado cuarto se presentan los *baselines* para el problema, y en el quinto nuestra propuesta de mejora del enfoque basado en RI. Finalmente, el apartado sexto resume las conclusiones y el trabajo futuro.

2 Trabajos relacionados

En general, los enfoques utilizados en el estado del arte para la recomendación y asignación de códigos CIE se pueden dividir en dos tipos: los basados en Procesamiento de Lenguaje Médico (PLM) y los basados en técnicas de clasificación.

En cuanto a los primeros, emplean bases de conocimiento y ontologías médicas para identificar los conceptos médicos en informes, y posteriormente asociarlos a los conceptos del esquema de clasificación. Por ejemplo, Ning et al. (2016) utilizan un modelo basado en ejemplos generado a partir de una base terminológica en chino con correspondencias con los códigos CIE-10 de 4 dígitos, por lo que aprovechan la estructura jerárquica del esquema de codificación; Chen et al. (2017) exploran la similitud semántica mediante el concepto de Longest Common Subsequence (LCS) entre los diagnósticos y los nombres dados por los CIE-10. Ambas propuestas basadas en aprendizaje no supervisado.

Por otro lado, en el segundo enfoque utilizado se generan clasificadores mediante aprendizaje automático supervisado para relacionar los informes con los códigos CIE. Por ejemplo, Subotin y Davis (2014) proponen un clasificador para anotar los procedimientos CIE-10, complementando el reducido tamaño de su corpus de entrenamiento mediante informes anotados con CIE-9 y las correspondencias del GEMs; Jatunarapit et al. (2016) emplean clasificadores basados en corpus ingleses y una serie de técnicas de RI para establecer similitudes con los términos tailandeses; Miftakhutdinov y Tutubalina (2017) emplean word embeddings entrenados a partir de un corpus de opiniones médicas de usuarios, junto con redes neuronales recurrentes para asignar los códigos.

A su vez, pueden encontrarse propuestas mixtas que combinan ambos enfoques. Por ejemplo, Boytcheva (2011) propone un clasificador supervisado multiclase, donde cada clase se corres-

[1] https://www.asco.org/practice-guidelines/billing-coding-reporting/icd-10/general-equivalence-mappings-gems

ponde con un código CIE-10 de 4 dígitos, y utilizan repositorios de terminología médica para enriquecer la representación de los documentos; Zweigenbaum y Lavergne (2016) combinan dos clasificadores: uno entrenado con un conjunto de informes, y otro a partir de distintos diccionarios médicos. Seva et al. (2017), en cambio, emplean un enfoque de RI para buscar posibles códigos CIE-10 candidatos en distintos diccionarios, junto con una serie de clasificadores para filtrarlos.

En lo que respecta al castellano, los trabajos existentes se han centrado en la anotación del esquema CIE-9 (Goicoechea et al., 2013; Perez et al., 2015), no habiendo hasta donde conocemos ninguna propuesta que no se base en los mapeos GEMs para recomendar o clasificar códigos CIE-10 a partir del texto libre de los informes médicos. Sí hay propuestas para otras lenguas, como el inglés (Subotin y Davis, 2014; Miftakhutdinov y Tutubalina, 2017), el búlgaro (Boytcheva, 2011), turco (Arifoglu et al., 2014), el chino (Ning et al., 2016; Chen et al., 2017) o el tailandés (Jatunarapit et al., 2016). Desgraciadamente, los resultados entre todos estos enfoques no son comparables entre sí dado que utilizan diferentes corpus y lenguas. A ello se suman las limitaciones de acceso a las colecciones de datos que impiden su distribución.

Aun siendo éste un problema típico de clasificación supervisada multiclase y multietiqueta, también se ha abordado desde enfoques no supervisados basados en técnicas de RI, fundamentalmente debido a los problemas que conlleva tener un número de clases tan elevado. Algunas aproximaciones plantean el análisis lingüístico del documento para la generación de diferentes consultas, así como el uso de recursos externos, como Wikipedia, manuales de CIE, PubMed, SNOMED-CT u otras ontologías médicas, lo que permite aplicar distintos métodos de expansión de consulta (Rizzo et al., 2015; Zhang et al., 2017). Otra posible aproximación consistiría en acceder a bases de conocimiento con los códigos CIE y representadas por grafos RDF mediante consultas SPARQL (Chiaravalloti et al., 2014).

En relación a la anotación de informes con códigos CIE-10 y los sistemas de RI, se han desarrollado proyectos para informes en turco (Arifoglu et al., 2014), japonés (Chen et al., 2014), francés (VanMulligen et al., 2016) y alemán (Schmidt et al., 2017; Ho-Dac et al., 2017). Los dos primeros casos siguen la tendencia marcada por los trabajos anteriores, explotando especialmente la estructura jerárquica de la codificación CIE-10, mucho más precisa que su versión anterior. En general, todas estas aproximaciones recurren al enriquecimiento de las descripciones de los códigos CIE con recursos externos, como diccionarios, otros corpus o sinónimos.

Actualmente existen aplicaciones comerciales que asisten a los anotadores en el proceso de asignación de códigos CIE-10-ES. Aunque éstas emplean los mapeos GEMs y las descripciones asociadas a los códigos, en su mayoría requieren palabras clave propuestas por los codificadores. Además, tras recibir la recomendación de los códigos por parte de la herramienta, el anotador tendrá que filtrar y seleccionar aquellos relativos al informe. Nuestra propuesta consiste en otras técnicas, basadas en clasificadores automáticos y RI; por tanto, una futura aplicación con codificadores reales requeriría modelos pre-entrenados y una aproximación de aprendizaje activo para ir mejorándolos a la vez que los codificadores van anotando nuevos informes.

3 Colección

Disponemos de los informes de alta del Hospital Universitario Fundación de Alcorcón (HUFA) del año 2016 anotados con códigos CIE-10; en concreto son 13.177 informes. El corpus generado sobre estos informes contiene un total de 106.304 códigos asociados, de los cuales 82.020 representan diagnósticos y 24.284 procedimientos. A su vez, estos 106.304 códigos asignados se corresponden con 8.445 códigos CIE-10 diferentes, lo que supone un 6 % de todas las posibles codificaciones de esta clasificación. Las ocurrencias de los códigos responden a una ley de potencias; es decir, la mayoría de ellos aparecen escasas veces, al contrario que una minoría, que presenta un amplio número de ocurrencias. En concreto, el 90 % de las codificaciones existentes en el corpus comprende el 20 % de todos los códigos anotados, y de ese 90 % la mitad se han anotado una única vez.

Aunque el número de códigos CIE-10 en cada informe varía considerablemente, en promedio alcanza una media de 8 diagnósticos y 2 procedimientos. Además, en el proceso de anotación se establece un orden a la hora de asignar códigos, de forma que los códigos CIE-10 que ocupan la primera posición en los diagnósticos anotados en cada informe suelen representar el factor de riesgo principal, mientras que el segundo código suele ser menos relevante, y así sucesiva-

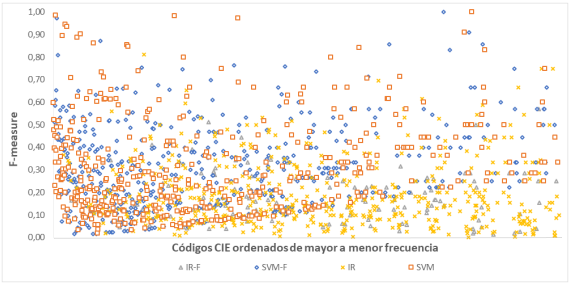

Figura 1: Baselines: distribución de las mejores aproximaciones (*F-measure*) por cada código CIE

mente.

Los informes médicos de alta de nuestra colección suelen seguir una estructura basada en secciones, en las que se agrupan diferentes tipos de información. A pesar de que existen varias plantillas para organizar estas secciones, los facultativos no siempre las utilizan; en ocasiones no definen las secciones o no marcan los límites de éstas. Según nos indicó el personal codificador del HUFA, considerando esa estructura, las secciones *Juicio Clínico*, *Procedimientos* y *Antecedentes* contienen la información más relevante para la codificación CIE-10. Se ha observado que, en general, *Procedimientos* y *Antecedentes* agrupan la información relativa a los procedimientos empleados, mientras que en la sección *Juicio Clínico* se resumen los principales diagnósticos de manera breve.

4 *Baselines basados en clasificación supervisada y en RI*

Con el fin de establecer un punto de partida y reflejar la complejidad que encierra la asignación automática de códigos CIE-10 en español, se han establecido dos *baselines*: uno basado en clasificación automática supervisada multietiqueta y otro basado en RI.

Hemos tenido en cuenta las secciones *Juicio Clínico*, *Procedimientos* y *Antecedentes* para representar los informes de alta previamente anonimizados. Por otra parte, hemos utilizado una versión de FreeLing enriquecida con 80.000 conceptos médicos (Oronoz et al., 2013), de la que se han eliminado los términos de la base de datos *BOT Plus* (en su mayoría medicamentos), debido a restricciones en la licencia de uso.

El *baseline* basado en aprendizaje supervisado utiliza un clasificador *SVM* (Máquinas de Vectores Soporte) mediante una validación cruzada sobre 10 iteraciones. En cada iteración se utiliza un 90 % del corpus para entrenamiento y el restante 10 % para test. En cuanto al preprocesamiento, se han aplicado dos configuraciones distintas. Una primera aproximación en la que se ha empleado un proceso de *stemming* y una lista de *stop words*, y otra en la que se han utilizado solo los conceptos médicos detectados por FreeLing. En ambas, la representación se realiza dentro del modelo de espacio vectorial, utilizando *TF-IDF* como medida de pesado.

Por otro lado, el *baseline* de RI se ha diseñado utilizando la biblioteca Apache Lucene, indexando los textos de las descripciones de los códigos CIE-10. Durante la indexación se ha utilizado el mismo proceso de *stemming* y la lista de *stop words*. Las consultas se construyen a partir de cada oración de las secciones consideradas del informe, combinado todas ellas mediante funciones lógicas OR en una única *query*; de esta forma, será el propio motor de búsqueda el que calcule la relevancia de cada código respecto a la consulta por medio de su función de *ranking* estándar. Para generar distintas configuraciones del sistema se han considerado diferentes aspectos en la consulta. Por un lado, se han utilizado dos tipos de *matching*: exacto, en el que la unidad de búsqueda en el campo descripción es la oración, o por palabras, que incrementa las probabilidades de asignación con cada coincidencia entre los términos y la descripción del código. Adicionalmente, se ha empleado FreeLing para

Aproximación	Precision	Recall	F-measure
SVM (SVM con Stemming)	**0.5370**	**0.2115**	**0.3034**
SVM-F (SVM con Freeling)	0.4667	0.1980	0.2780
RI (RI sin FreeLing, matching por palabras y radio de búsqueda 15)	0.0551	0.0799	0.0653
RI-F (RI con FreeLing, matching exacto y radio de búsqueda 15)	0.0415	0.0310	0.0355

Tabla 1: Resultados de los *baselines*

extraer los conceptos médicos. Por último, se han establecido diferentes radios de búsqueda para limitar el número de códigos CIE-10 devueltos por el motor de búsqueda con cada informe.

Aproximación	N. CIEs mejor modelados	$\overline{\text{F-measure}}$	$\overline{\Delta \text{ F-measure}}$	$\overline{\text{N. ocurrencias}}$
SVM-F	563	**0.3587**	0.1017	70
SVM	493	0.3427	0.1458	92
RI-F	278	0.2576	0.1488	10
RI	**1198**	0.2406	**0.2037**	11
Sin Mod.	5960	-	-	3

Tabla 2: Mejores *baselines* por cada CIE-10. $\overline{\Delta \text{ F-measure}}$ es el incremento medio de *F-measure* del mejor *baseline* respecto al resto

Las mejores aproximaciones y sus resultados pueden verse en la Tabla 1. Se han empleado las medidas estándar de evaluación *Precision*, *Recall* y *F-measure* para comparar los códigos CIE correctamente asociados en cada enfoque. Por otra parte, la Figura 1 representa la distribución de los mejores *baselines* por cada código CIE de la colección en función de *F-measure*, ordenándolos de mayor a menor frecuencia dentro del corpus. En la la Tabla 1 se observa que los mejores resultados se alcanzan con el enfoque de clasificación supervisada, mientras que la aproximación de RI ofrece unos resultados bastante distanciados de éstos. Sin embargo, observando la distribución de los resultados de *F-measure* en la Figura 1, se llega a la conclusión de que el enfoque supervisado principalmente consigue clasificar codificaciones CIE-10 con una alta frecuencia en el corpus, las cuales suponen un reducido número; en concreto, consigue modelar adecuadamente un 10 % de todas las codificaciones CIE-10 del corpus, que se corresponden con el 80 % de todos los códigos anotados. Cómo ya se indicó, la mayoría de los códigos CIE-10 del corpus aparecen con poca frecuencia, lo que dificulta su modelado. El enfoque de RI, sin embargo, aunque arroja peores valores, no tiene la misma dependencia de la frecuencia, lo que podría suponer una posible vía para complementar el enfoque supervisado.

En la Tabla 2 se muestra el número de códigos CIE-10 en los que una aproximación representa el mejor sistema de anotación para ese código. El enfoque de RI sin aplicar FreeLing duplica a los sistemas *SVM* en el número de mejores codificaciones recuperadas en base a *F-measure*. Como era de esperar, observando la media de ocurrencias de esas codificaciones se aprecia que los sistemas *SVM* modelan mejor aquellos códigos con mayor número de apariciones en la colección de entrenamiento, mientras que el enfoque de RI no refleja esa dependencia presentando además el mayor incremento medio de *F-measure* con respecto al resto de *baselines*.

El objetivo de este estudio preliminar es establecer mediante qué técnicas y con qué configuraciones sería posible asistir al proceso de anotación manual de informes médicos, por lo que se persigue la búsqueda de valores altos de *Precision* y *F-measure*. Por ello, y a la vista de los resultados anteriores, parece claro que las aproximaciones de clasificación pueden proporcionar mejores resultados en los códigos CIE-10 más frecuentes, mientras que el enfoque de RI puede funcionar mejor en los casos de códigos más inusuales. Centrándonos ahora en el sistema de RI, en el siguiente apartado se van a presentar distintos enfoques para mejorar sus resultados.

5 Mejora de la propuesta basada en RI

El *baseline* de RI se basa únicamente en la indexación de la descripción de los códigos CIE-10. Para mejorar sus resultados se han estudiado: el enriquecimiento de los índices con terminología asociada a cada código, y la combinación de consultas obtenidas a partir de distintas secciones del informe.

5.1 Enriquecimiento de los índices con KLD

Se propone emplear la *Divergencia de Kullback-Leibler* (KLD) como método para extraer terminología relacionada con cada código y añadirla al índice. Con esta técnica se pretende encontrar qué términos aparecen con mayor probabilidad en los informes anotados con un determi-

Aproximación	Precision	Recall	F-measure
RI sobre la Descripción, sin FreeLing, matching por palabras y radio de búsqueda 15	0.0551	0.0799	0.0653
RI sobre KLD con 30 términos, sin FreeLing, matching por palabras y radio de búsqueda 15	**0.2093**	0.0189	0.0346
RI sobre Descripción y KLD con 30 términos (pesado en 2), sin FreeLing, matching por palabras y radio de búsqueda 10	0.1327	**0.0924**	**0.1090**

Tabla 3: Resultados mediante el enriquecimiento de los índices con KLD

nado código, y que a su vez tengan una baja probabilidad en el resto. Para ello, se ha empleado el mismo 90 % de los informes usados en el entrenamiento del *SVM* para calcular cuáles serían los términos que mejor caracterizan los documentos en los que aparece cada código CIE-10 del corpus. Con el objetivo de estudiar distintas configuraciones en esta aproximación, se ha aplicado KLD sobre el conjunto de términos presente en las secciones *Juicio Clínico*, *Procedimientos* y *Antecedentes*, así como únicamente sobre los términos del *Juicio Clínico*. Posteriormente, por cada código CIE-10, además de la descripción del código, se indexan en campos separados los *N* términos más representativos de cada código, variando *N* de 1 a 80. Por último, se ha considerado asignar diferentes factores de empuje a los campos del índice para estudiar su influencia en el proceso de anotación.

La Tabla 3 presenta las configuraciones que han obtenido los mejores valores de *F-measure*. Aunque se observa una mejora de la *Precision* al realizar las consultas solo contra el campo KLD que contiene los 30 términos más representativos de cada código CIE-10, la mejora más importante de los resultados se produce al combinar este campo con la descripción del código en una única consulta.

5.2 Consultas considerando la estructura del informe

Como no se dispone de evidencias para averiguar qué parte específica del informe determina cada código, se ha querido analizar el efecto de realizar las consultas con las secciones *Juicio Clínico*, *Procedimientos* y *Antecedentes* de forma individual y combinadas. Además, se han estudiado los resultados para diagnósticos y procedimientos de forma independiente.

Tras el análisis de múltiples configuraciones, en la Tabla 4 se muestran los mejores valores de *Precision* y *F-measure* por cada tipo de código. Se puede ver que el enfoque de RI consigue anotar más diagnósticos que procedimientos, lo que puede indicar que los diagnósticos están mejor caracterizados en los informes. Además, se muestran también la *Precision* para los primeros (P_1), segundos (P_2) y terceros códigos (P_3), cuantificando así los códigos correctos recuperados en los informes en dichas posiciones.

Si nos fijamos en los valores de *Precision* según la relevancia de los códigos (P_1, P_2 y P_3), se puede observar la existencia de diferentes configuraciones que favorecen la obtención de códigos en distintas posiciones y, por lo tanto, sería de esperar que estableciendo una combinación de ellas se pudiera incrementar la precisión global del sistema. Por otro lado, observando de forma separada los altos valores de *Precision* obtenidos para diagnósticos y procedimientos, sería conveniente aplicar diferentes técnicas de RI en ambos tipos de código CIE-10. En definitiva, se puede pensar en diseñar un método basado en la combinación de algunas de las configuraciones mostradas para potenciar este enfoque de anotación basado en RI.

En la Tabla 4 también se muestra esa posible combinación (Q7), alcanzando unos valores de

Tipo	Query	Secciones	Campos de búsqueda	Aproximación	P_1	P_2	P_3	P_R	Precision	Recall	F-measure
Diagnósticos	Q_1	Juicio clínico y Antecedentes	KLD con 30 términos (pesado en 2)	RI sin FreeLing, matching por palabras y radio de búsqueda 10	0.0013	**0.5449**	**0.3015**	**0.2636**	0.1608	**0.1796**	**0.1697**
	Q_2	Juicio clínico	Descripción y KLD con 10 términos del juicio clínico	RI sin FreeLing, matching por palabras y radio de búsqueda 1	**0.1740**	0.0634	0.0028	0.0000	0.3655	0.0405	0.0729
	Q_3	Juicio clínico y procedimientos	Descripción y KLD con 1 término del juicio clínico (pesado en 2)	RI sin FreeLing, matching exacto y radio de búsqueda 1	0.0486	0.0028	0.0000	0.0000	**0.4926**	0.0102	0.0200
Procedimientos	Q_4	Juicio clínico y Antecedentes	KLD con 30 términos (pesado en 2)	RI sin FreeLing, matching por palabras y radio de búsqueda 10	0.0150	**0.1926**	**0.0685**	**0.1576**	0.1139	**0.0708**	**0.0874**
	Q_5	Procedimientos	KLD con 10 términos	RI sin FreeLing, matching por palabras y radio de búsqueda 2	0.0039	0.0043	0.0005	0.0000	**0.2259**	0.0028	0.0055
	Q_6	Juicio clínico, procedimientos y antecedentes	KLD con 3 términos	RI sin FreeLing, matching por palabras y radio de búsqueda 2	**0.0553**	0.0985	0.0016	0.0000	0.2012	0.0482	0.0777
$Q_7 = Q_1 + Q_2 + Q_3 + Q_4 + Q_5$					**0.0077**	**0.4298**	**0.2578**	**0.2870**	**0.1465**	**0.1165**	**0.1298**

Tabla 4: Resultados con otras secciones del informe desglosados por tipo de código. P_1, P_2 y P_3 representan la fracción de códigos correctos recuperados en las posiciones de anotación 1, 2 y 3 respectivamente. P_R indica esa fracción en el resto de posiciones

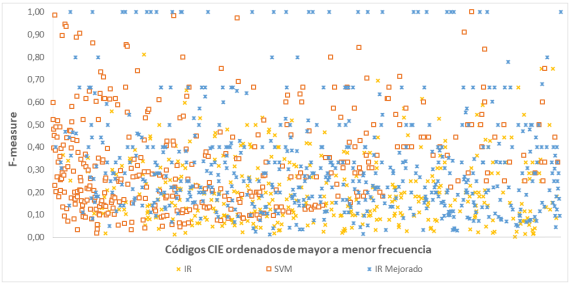

Figura 2: Mejora de RI frente a los *baselines*: distribución de las mejores aproximaciones (*F-measure*) por cada código CIE

F-measure de 0.13, duplicando los valores obtenidos por el *baseline*. Se ha representado su distribución (Figura 2) frente a los resultados de los *baselines*. Como se aprecia, esta combinación de consultas logra mejorar a las otras aproximaciones en una de cada dos codificaciones CIE. Por último, en la Tabla 5 se observa que la nueva aproximación del sistema de RI duplica el número de códigos CIE mejor caracterizados con respecto al sistema que genera las consultas sobre la descripción de los códigos. Al mismo tiempo, el número de códigos CIE sin modelar desciende de 5960 a 4334.

Aproximación	N. CIEs mejor modelados	$\overline{\text{F-measure}}$	$\overline{\Delta \text{ F-measure}}$	$\overline{\text{N. ocurrencias}}$
SVM	427	0.3649	0.1487	97
RI	914	0.2751	0.1948	10
RI Mejorado	**2054**	0.2980	**0.2636**	8
Sin Mod.	4334	-	-	3

Tabla 5: Mejores aproximaciones por cada CIE-10: comparación con RI mejorado. $\overline{\Delta \text{ F-measure}}$ es el incremento medio de *F-measure* del mejor *baseline* respecto al resto

6 *Conclusiones y trabajo futuro*

Los métodos de clasificación basados en un aprendizaje supervisado demuestran su eficacia en problemas con un gran número de datos disponibles. En nuestro caso concreto, resultan útiles para anotar los códigos CIE presentes en un número considerable de informes del corpus. Sin embargo, el comportamiento obtenido para los códigos CIE-10 con pocas ocurrencias es opuesto, ya que el sistema en la mayoría de los casos no es capaz de generar modelos de clase a partir de los pocos informes disponibles.

El problema de la asignación de códigos CIE-10 viene condicionado por el enorme número de códigos y por la distribución real de diagnósticos y procedimientos. Llevado a la práctica, los desórdenes que padecen las personas, así como los procedimientos que llevan a cabo los facultativos, suelen seguir una tendencia general que va a repercutir sobre los datos finalmente disponibles. Si bien es cierto que un método que se centre en los códigos más frecuentes puede responder con unos niveles de *Precision* considerables, supone un sesgo nada despreciable, ya que cualquier código menos usual prácticamente no se contempla. Siendo éste un problema en el que las soluciones se basan en códigos bien definidos y estructurados, no parecería lógico desarrollar un sistema incapaz de asignar una gran parte de los códigos, aún cuando las condiciones lo demandasen; por ello, en este artículo se estudian técnicas complementarias al aprendizaje automático, específicamente desde un enfoque de RI.

Como muestra la experimentación realizada, la generación de consultas a partir del informe y su aplicación únicamente sobre la descripción de los códigos no arroja buenos resultados. No obstante, dicha consulta en combinación con la terminología extraída a partir de la *Divergencia de Kullback-Leibler* logra recuperar más códigos CIE-10 afines a los informes. Un análisis más detallado de esos resultados desvela la distinta naturaleza de diagnósticos y procedimientos, así como los códigos más relevantes para el anotador y aquellos asignados en una segunda o tercera posición. Todas estas pautas parecen indicar la necesidad de seguir explorando la combinación de aproximaciones complementarias, así como de combinar distintas consultas, cada una orientada a esa fracción de códigos CIE que persigue recuperar.

Agradecimientos

Este trabajo ha sido financiado por el Ministerio de Ciencia e Innovación a través del proyecto

PROSA-MED: TIN2016-77820-C3.

Bibliografía

Arifoglu, D., Deniz, O., Aleçakir, K., Yöndem, M. 2014. CodeMagic: Semi-Automatic Assignment of ICD-10-AM Codes to Patient Records. En *Information Sciences and Systems 2014*, páginas 259–268.

Boytcheva, S. 2011. Automatic Matching of ICD-10 codes to Diagnoses in Discharge Letters. En *Proceedings of BioNLP*, páginas 11–18.

Chen, S., Lai, P., Tsai, Y., Chung, J., Hsiao, S., Tsai, R. 2014. NCU IISR System for NTCIR-11 MedNLP-2 Task. En *Proceedings of the 11th NTCIR Conference*, páginas 9–12.

Chen, Y., Lu, H., Li, L. 2017. Automatic ICD-10 coding algorithm using an improved longest common subsequence based on semantic similarity. En *PloS one*, vol. 12(3).

Chiaravalloti, M.T., Guarasci, R., Lagani, V., Pasceri, E., Trunfio, R. 2014. A Coding Support System for the ICD-9-CM standard. En *Proceedings of ICHI 2014*, páginas 71–78.

Goicoechea, J.A., Nieto, M.A., Laguna, A., Canto V.D., Rodríguez, J., Murillo, F. 2013. Desarrollo de un sistema de codificación automática para recuperar y analizar textos diagnósticos de los registros de servicios de urgencias hospitalarios. En *Emergencias*, vol 25:430–436.

Ho-Dac, L. M., Fabre, C., Birski, A., Boudraa, I., Bourriot, A., Cassier, M., Delvenne, L., Garcia-Gonzalez, C., Kang, E. B., Piccinini, E., Rohrbacher, C., Séguier, A. 2017. LITL at CLEF eHealth2017: automatic classification of death reports. En *CLEF*.

Jatunarapit, P., Piromsopa, K., Charoeanlap, C. 2016. Development of thai text-mining model for classifying ICD-10 TM. En *Proceedings of ECAI 2016*, páginas 1–6.

Miftakhutdinov, Z., Tutubalina, E. 2017. KFU at CLEF eHealth 2017 Task 1: ICD-10 Coding of English Death Certificates with Recurrent Neural Networks. En *CLEF*.

van Mulligen, E. M., Afzal, Z., Akhondi, S. A., Vo, D., Kors, J. A. 2016. Erasmus MC at CLEF eHealth 2016: Concept Recognition and Coding in French Texts. En *CLEF Working Notes*, páginas 171–178.

Ning, W., Yu, M., Zhang, R. 2016. A hierarchical method to automatically encode Chinese diagnoses through semantic similarity estimation. En *BMC Medical Informatics and Decision Making*, vol. 1:16–30.

Pérez, A., Gojenola, K., Casillas, A., Oronoz, M, Díaz de Ilarraza, A. 2015. Computer aided classification of diagnostic terms in Spanish. En *Expert Systems with Applications*, vol. 42(6):2949–2958.

Rizzo, S.G., Montesi, D., Fabbri, A, and Marchesini, G. 2015. ICD Code Retrieval: Novel Approach for Assisted Disease Classification. En *Data Integration in the Life Sciences*, LNCS, vol. 9162:147–161.

Schmidt, D., Budde, K., Sonntag, D., Profitlich, H. J., Ihle, M., Staeck, O. 2017. A novel tool for the identification of correlations in medical data by faceted search. En *Computers in Biology and Medicine*, vol. 85:98–105.

Seva, J., Kittner, M., Roller, R., Leser, U. 2017. Multi-lingual ICD-10 coding using a hybrid rule-based and supervised classification approach at CLEF eHealth 2017. En *CLEF*.

Subotin, M., Davis, A. 2014. A System for Predicting ICD-10-PCS Codes from Electronic Health Records. En *Proceedings of BioNLP*, páginas 59–67.

Zhang, D., He, D., Zhao, S., Li, L. 2017. Enhancing Automatic ICD-9-CM Code Assignment for Medical Texts with PubMed. En *Proceedings of BioNLP*.

Zhao, S., He, D., Zhang, D., Li, L., Meng, R. 2017. Automatic ICD Code Assignment to Medical Text with Semantic Relational Tuples. En *Proceedings of iConference 2017*, vol. 2:156–158.

Zweigenbaum, P., Lavergne, T. 2016. Hybrid methods for ICD-10 coding of death certificates. En *Proceedings of LOUHI*, páginas 96–105.

Oronoz, M., Casillas, A., Gojenola, K., Perez, A. 2013. Automatic annotation of medical records in Spanish with disease, drug and substance names. En *Iberoamerican Congress on Pattern Recognition*, páginas 536–543.

Procesamiento del Lenguaje Natural, Revista nº 60, marzo de 2018, pp. 53-60 recibido 22-11-2017 revisado 16-12-2017 aceptado 20-02-2018

Sequential dialogue act recognition for Arabic argumentative debates

Reconocimiento de acto de diálogo secuencial para debates argumentativos árabes

Samira Ben Dbabis[1], Hatem Ghorbel[2], Lamia Hadrich Belguith[3]

[1,3] ANLP Research Group, MIRACL Laboratory, University of Sfax, Tunisia
[2] University of Applied Science of West Switzerland HE-Arc Ingénierie, Switzerland

[1] samira.benedbabis@fsegs.rnu.tn
[2] hatem.ghorbel@he-arc.ch
[3] l.belguith@fsegs.rnu.tn

Abstract: Dialogue act recognition remains a primordial task that helps user to automatically identify participants' intentions. In this paper, we propose a sequential approach consisting of segmentation followed by annotation process to identify dialogue acts within Arabic politic debates.To perform DA recognition, we used the CARD corpus labeled using the SADA annotation schema. Segmentation and annotation tasks were then carried out using Conditional Random Fields probabilistic models as they prove high performance in segmenting and labeling sequential data. Learning results are notably important for the segmentation task (F-score=97.9%) and relatively reliable within the annotation process (f-score=63.4%) given the complexity of identifying argumentative tags and the presence of disfluencies in spoken conversations.
Keywords: DA recognition, annotation scheme, Arabic debates, CRF classifier.

Resumen: El reconocimiento del acto de diálogo sigue siendo una tarea primordial que ayuda al usuario a identificar automáticamente las intenciones de los participantes. En este documento, proponemos un enfoque secuencial que consiste en la segmentación seguida de un proceso de anotación para identificar actos de diálogo dentro de los debates políticos árabes. Para realizar el reconocimiento DA, utilizamos el corpus CARD etiquetado utilizando el esquema de anotación SADA. Las tareas de segmentación y anotación se llevaron a cabo utilizando modelos probabilísticos de Campos aleatorios condicionales, ya que demuestran un alto rendimiento en la segmentación y el etiquetado de datos secuenciales. Los resultados de aprendizaje son especialmente importantes para la tarea de segmentación (F-score = 97.9%) y relativamente confiables dentro del proceso de anotación (f-score = 63.4%) dada la complejidad de identificar etiquetas argumentativas y la presencia de disfluencias en las conversaciones habladas.
Palabras clave: Reconocimiento DA, esquema de anotación, debates árabes, clasificador CRF.

1 Introduction

Dialogue acts (DA) are considered as the minimal units of linguistic communication that reveal speaker's intention (Grosz and Sidner, 1986). Automatic dialogue act detection is an important clue for various applications like dialogue systems, human conversations understanding, machine translation, topic detection and summarization.

In our work, the aim of dialogue acts recognition is to better understand human conversations based mainly on argumentative tags in order to extract participants' conflicts in terms of opinions' reject or accept and arguments presented to defund their ideas.

ISSN 1135-5948. DOI 10.26342/2018-60-6 © 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

To perform this task, we propose in a first step a complete annotation scheme consisting of 40 DAs. In a second step, we reduced the initial scheme to 19 acts as we decided to focus mainly on argumentative tags and merge others for instance social obligation management and turn management categories.

The proposed DAs are automatically identified using machine learning techniques applied on a large corpus collected from politic debates found to have explicit argumentative taxonomy and forms like opinions, arguments, acceptations, rejects, explanations, justifications, etc.

This paper is organized as follows. The first section describes the major DA recognition approaches. In section 2 we detail the implications of DA recognition explored mainly in building argumentative discourse structure. In sections 3 and 4, we present the proposed annotation scheme and the corpus used to perform learning machine experiments. Section 5 details the proposed recognition sequential approach consisting of two main tasks: segmentation followed by annotation of dialogue acts. For each task, we focus on the used learning technique, the experimental data, the adopted features and the evaluation results.

2 Building Argumentative Structure

Dialogue acts play a vital role in the identification of discourse structure. In this context, Grosz and Sidner (1986) claim about task structure influencing dialogue structure. It seems likely that there are structures higher than a single utterance, yet more fine grained than a complete dialogue. Several researchers identify structures within dialogue at levels higher than individual utterances or speaker turns, but below the level of complete discourse description. There has been some significant exploration of the use of sequences of Dialogue Acts, at a number of levels of granularity.

The simplest dialogue sequence model is the use of adjacency pairs (Schegloff et al., 1973) which are functional links between pairs of utterances such as question/answer, opinion request/opinion, etc.

Within the adjacency pairs model, the importance of tracking a deeper structured representation has been recognized in Ezen-Can and Boyer (2015), Swapna and Wiebe (2010) and Galley et al. (2004).

In fact, Ezen-Can and Boyer (2015) investigate sequences of acts to automatically detect the interaction mode between students and teachers (tutor lecture, tutor evaluator, Extra-domain and student). Swapna and Wiebe (2010) use the AMI corpus (Carletta et al., 2005) to detect opinions' categories such argument and sentiment in meetings. Galley et al. (2004) also explored adjacent act chains to extract the agreement and disagreement pairs within meetings of the ICSI corpus (Shriberg et al., 2004).

In our work, the main implications of recognizing dialogue acts are to build argumentative chains consisting of pairs or more than two acts to highlight argumentative interaction between participants. For instance an opinion request asked by the animator is generally followed by an opinion tag which can be rejected or accepted by other participants. The opinion holder can reinforce his point of view by exposing arguments, explanations or justifications.

Thus, dialogue act sequences can help in capturing the essential argumentative information in terms of what topics have been discussed and what alternatives have been proposed and accepted by the participants. They can be also useful in opinion question/answering systems to answer complex real user queries like "who rejected the opinion of X?" which is not evident to reply using traditional information retrieval engines.

3 Annotation scheme

Over the years a number of dialogue act annotation schemas has been developed, such as those of the MapTask studies outlining road mapping task-oriented dialogues (Carletta, 1996) and the Verbmobil project (Alexandersson et al., 1998) focusing on meeting scheduling and travel planning domains. Later, DAMSL (Core, and Allen, 1997) annotation schema was developed for multidimensional dialogue act annotation. As an extension of DAMSL, The DIT++ schema (Bunt, 2009) combines the multidimensional DIT schema, developed earlier (Bunt, 1994) with concepts from these various alternative schemas, and provides precise and mutually consistent definitions for its communicative functions and dimensions.

These annotation schemes have been used to mark-up several dialogue corpora in non Arabic

languages. To the best of our knowledge, few works were developed in Arabic language. We mention the taxonomy proposed by Shala et al. (2010) that proposed speech acts taxonomy including the following set of 10 categories dealing with general information requests followed by answers.

Recently, Elmadany et al. (2014) reported a schema for inquiry-answer instant messages in Egyptian dialect such as flights, mobile service operators, and banks; this schema contains 25 DAs based on request and response dimensions.

Given that the main purpose of identifying dialogue acts is to build argumentative discourse structure, we cannot profit from previous annotation schemes and we need to develop a specific-purpose taxonomy based mainly on argumentative acts called SADA: **S**cheme **A**nnotation for **D**ebates in **A**rabic.

The first release of SADA (BenDbabis et al., 2012) is a complete tagset consisting of 40 dialogue acts related to the following categories: *social obligation management*, *turn management*, *Request*, *Argumentative*, *Answer, statement and others.*

In a second step, we reduced the initial tagset to 19 acts (BenDbabis et al., 2015). We merge acts expressing social obligation management into a single dialogue act named *SOM*. We also combine acts expressing Turn Management in one act labeled *TM*. We eliminate acts having very few occurrences in the corpus like *statement*, *propose, hope, wish, invoke, warn* and *order.*

We also eliminate the following tags expressing Appreciation (*app*), disapproval (*disap*), partial accept (*part_acc*) and partial reject (*part_rej*). In fact, we considered appreciation and partial accept acts as acceptation tags while disapproval and partial reject was considered as forms of reject. We add the tag *Thesis* in the argumentative category referring to a new topic or idea introduced by the presenter that can be retained or rejected by the audience.

4 CARD corpus

Corpora annotated for Dialogue Acts play a key role in the validation and evaluation of the proposed annotation taxonomies. In our context of work, our main purpose is to track argumentative information from human conversations. Thus, we collected a set of politic debates from Aljazeera TV broadcasts discussing hot topics (Tunisian and Egyptian revolutions, Syrian war, Tunisian elections, etc); named CARD: **C**orpus of **AR**abic **D**ebates. The choice of this corpus is argued by the important argumentation hold in its content mainly conveyed by exchanging opinions, agreements, disagreements, etc.

The CARD corpus was manually annotated using the ActAAr annotation tool: **A**ct **A**nnotation in **Ar**abic (BenDbabis et al., 2012) in three steps reaching 50 conversations in the latest release. Basic information of the different versions of the CARD corpus is detailed in Table 1.

	CARD 1.0	**CARD 1.1**	**CARD 1.2**
Total number of conversations	8	22	50
Total number of turns	773	1805	5085
Total number of utterances	2367	6050	14062
Total number of words	37075	101169	260212
Average number of turns/conversation	97	82	102
Average number of utterances/conversation	296	275	281
Average number of words/conversation	4635	4599	5204

Table 1: CARD corpus statistics

5 DA recognition

Dialogue act recognition consists mainly of two subtasks as segmentation and annotation. These two steps may be carried separately; segmentation followed by annotation or simultaneously at one joint step. In our work, we typically assumed that the true segmentation boundaries lead to better annotation results. As a consequence, a degradation of the performance due to imperfect segmentation boundaries is to be expected. Thus, we decided to carry out a sequential approach that separate the two subtasks of dialogue acts recognition framework.

5.1 Segmentation task

The Segmentation task consists of dividing the conversation into turns; each turn is then segmented into meaningful units named utterances. For each utterance, a dialogue act unit is assigned. The problem of identifying utterance boundaries has been addressed with

machine learning approaches. Most researchers applied generative models Hidden Markov Models (HMM) experimented by Ivanovic (2005) to find the most likely segment boundaries in online instant messages based services and Naïve Bayes generative classifier (Geertzen et al., 2007) for assistance-seeking Dutch dialogues within the DIAMOND corpus.

Discriminative models have been experimented to perform better than HMMs and maximum entropy approaches for utterance segmentation. The most common discriminative models are Conditional Random Fields (CRF) introduced by Lafferty et al. (2001). It was applied by Silvia et al. (2011) using two corpora namely SWITCHBOARD and LUNA corpus.

Semi-supervised learning approaches were also implemented in the purpose to reduce the amount of labeled data needed to train statistical models.

In this context, Guz et al. (2010) applied self-training and co-training approaches using the ICSI meeting corpus (Schrieberg et al., 2004) of multichannel conversational speech data.

Most of utterance segmentation researches were applied on various languages corpora like English, German and Italian. Few works focus on utterance segmentation of Arabic conversations. We cite the work of Elmadany et al. (2015) who proposed an automatic segmentation utterance approach using SVM classifier for Egyptian instant messages.

In our work, we applied the probabilistic CRF learner to automatically define utterances boundaries. The choice of this model is justified by its efficiency for labeling and segmenting sequential data.

To perform training and test tasks, we used the CARD corpus enhanced in three steps ranging from 8 to 50 conversations. BenDbabis et al. (2016) expose utterance segmentation experiments using CARD 1.1 corpus.

5.1.1 Features selection

Selecting most pertinent features has a great effect on learning machine process mainly on resulting labeled data. In our work, we explored lexical features namely punctuation marks and cue words as important indicators of segment boundaries. We also use morpho-syntactic features as the Part Of Speech (POS) of words. For each word, we take into account a context window of +2/-2; that means we consider dependencies between the current word and the two previous and next words.

As a lexical characteristic, we focus on punctuation as a determinant clue that occurs frequently and the end of an utterance. For example question marks mostly delimit the end of a question.

Question words are also considered as pertinent cue words that express a request or a general question in the beginning of conversation segments.

Lexical cues are frequently used to identify the beginning of a segment. For in
stance the words "مرحبا"/"welcome", " أهلا بكم"/"Hello", "طيب"/"ok", "نعم"/"yes", "أوافق"/"I agree" occur generally at first of utterances.

The POS of each word can also help to recognize utterance delimiters given that utterances often start with prepositions ("في"/ "in", "من"/"from"), adverbs ("طيب"/"ok", " أولا"/"first") or verbs ("أرى"/"I see", "أعتقد"/ "I think").

5.1.2 Results

We experiment the CRF classifier using the different CARD versions. For each release of the corpus, we assess precision, recall and f-measure traditional evaluation metrics. To better evaluate CRF efficiency, learning results were compared to SVM, Naïve Bayes (NB) and Decision trees (J48) classifiers. Comparison results of the used classifiers are shown in Figure 1.

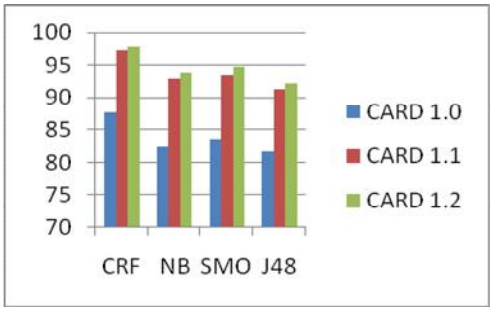

Figure 1: CARD f-measure results

Evaluation results prove the high performance of CRF models in segmenting conversations into meaningful utterances. We obtained a recall rate of 98%, a precision value of 97,8% and an f-measure score of 97,9% using the CARD 1.2 corpus. We also confirm the importance of amount of data in learning machine experiments. Best results are acquired when using larger corpus (CARD 1.2) including 260383 words.

Segmentation errors are mainly due to the fact that punctuation marks can be used inside segments. Mis-segmented utterances can be also explained by the presence of cue words inside utterances which can lead to wrong segment boundaries.

5.2 Annotation task

The annotation task is fundamental in dialogue acts recognition framework. For each segmented utterance, we assign a label expressing the user's intention throughout the conversation. Research has continued to experiment machine learning techniques to automatically identify DAs. Supervised modeling approaches are frequently used including sequential approaches and vector-based models.

Sequential approaches typically formulate dialogue as a Markov chain in which an observation depends on a finite number of preceding observations. HMM-based approaches generate optimal dialogue act sequences using the Viterbi (Stolcke et al., 2000; Bangalore et al., 2008; Ondáš et al., 2016). Research using sequential approaches usually involves combinations of N-grams and Hidden Markov Models.

Vector-based approaches such as maximum entropy (Sridhar et al., 2007) and SVM models (Zhou et al., 2015) frequently take into account lexical, syntactic and structural features. Lexical and syntactic cues are extracted from local utterance context, while structural features involve longer dialogue act sequences in task-oriented domains.

Neuronal networks (Shen et al., 2016) were also investigated to automatically classify dialogue acts. Zhou et al. (2015) applied a combination of heterogeneous deep neural networks with conditional random fields for Chinese corpus.

More interestingly, researchers focused on features enhanced dialogue context (Webb et al., 2005; Hoque et al., 2007; Coria el al., 2007) ;Di Eugenio et al., 2010a; Samei et al., 2014; Ribeiro et al., 2015) that shows a predictive power on Dialogue Act classification. Recently semantic information was explored in the annotation of Czech dialogue corpus (Pavel et al., 2015). Yeh (2016) also involve using semantic dependency graphs with probabilistic context-free grammars (PCFGs).

Most DA annotation classifiers were experimented using several dialogue corpora in different languages such as English, German and Spanish.

However, very few works were developed for Arabic language. Shala et al. (2010) propose speech acts classification model using SVM for the labeling of a tagset of 10 acts. This tagset includes general-purpose actions that can be applied to independent domain corpora.

Elmadany et al. (2015) also experiment SVM model for question-inquiry dialogue acts recognition with a reduced labeling schema of 25 acts for Egyptian spontaneous dialogues and instant messages.

Nevertheless, the proposed annotation works mainly label short utterances expressing requests, questions and answers that are not complex to identify especially with the presence of a predefined list of cue words.

In our work, we implement the CRF model to label utterances segmented in the previous task. The proposed model takes advantage of dependencies between interconnected annotations compared to conventional classification models.

To perform the annotation process, we used the CARD corpus annotated with the SADA annotation schema.

5.2.1 Features selection

DA classification involves linguistic, prosodic and multimodal features. Most of researches explore linguistic features that include lexical, syntactic, semantic and context-based features (Sridhar et al., 2009; Kim et al., 2012). In our context, we choose the most relevant characteristics to our task namely lexical, morpho-syntactic, utterance and structural learning features. We detail below the selected features.

- *Question words:* expressing requests and general questions; for example the word " لماذا"/ " why" indicates a justification request.
- *Cue words :* are most common words frequently used along the conversation. For instance " أهلا بكم" /"welcome", and " شكرا"/"thanks" are used for introducing social obligation management acts.
- *Opinion words:* used when presenting argumentative information like opinions, arguments, acceptations and rejects.
- *Part Of Speech:* grammatical categories of words can reflect the act expressed; for

instance, verbs are frequently used for argumentative tags.

- *Utterance speaker:* the actor of the current utterance.
- *Speaker role:* whether the speaker is the animator of the discussion or just a participant. *Mostly, the animator introduces and ends the discussion and manages the participants' turn taking.*
- *Previous act*: can help to anticipate the next DA label. *For instance, a confirmation request is generally followed by a confirmation.*
- *Previous utterance speaker:* it is important to identify whether the previous utterance has the same actor as the current one.

5.2.2 Results

To train the CARD data, we defined a template that includes unigrams and bigrams of features to focus on the dependencies between features. For each feature, we take into account the two previous and next words (context window=2). We used the CRF++ platform to train and test the CRF model. This tool is an implementation of CRF for labeling sequential data.

Annotation relevance is evaluated using the known metrics as recall, precision and f-measure. All evaluation results shown below were carried out using 10 folds cross validation. To evaluate CRF performance, we compared the obtained results to Naïve Bayes (NB), Decision tree (J48) and SMO classifiers in Figures 2, 3 and 4. For all classifiers, we notice that the increase of the corpus size improves notably the annotation results. For instance, CRF achieves an f-score of 32,4% with CARD 1.0 while the latest release CARD 1.2 reaches an f-measure rate of 63,4% using the same classifier.

Annotation results show that CRF model outperform other classifiers with all releases of CARD corpus with a recall value of 62,2%, a precision rate equal to 64,7% and an f-measure of 63,4%. Thus, CRF results reinforce the high performance of this classifier in labeling sequential data.

Main annotation errors are due classification ambiguities for identifying argumentative tags. There are confusions between arguments, explanations and opinion acts especially when specific lexical cues are absent. Turn management utterances are generally predicted by the enunciation context. So it is difficult to identify these tags that don't obey to the use of general rules or particular cue words. In addition, some lexical cues can have different meaning depending on context. For example, the word "ok" can be used as a form of acceptation or as an acknowledgement act to manage the conversation turn takings.

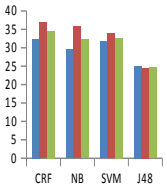

Figure 2: CARD 1.0

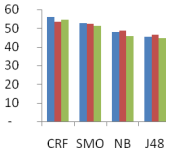

Figure 3: CARD 1.1

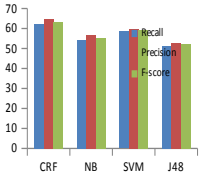

Figure 4: CARD 1.2

6 Conclusion and perspectives

To the best of our knowledge there is no similar work that identifies argumentative dialogue acts within politic debates. In this paper, we proposed a novel sequential dialogue act recognition approach carrying out separately segmentation and annotation tasks. To automatically perform dialogue act identification process, we applied the probabilistic model CRF in both segmentation and labeling subtasks. Results confirm the effectiveness of CRF compared to naïve bayes, SVM and decision trees learning algorithms. Annotation experiments are very encouraging with an average F-score of 63,4%. These results are due to the complexity of labeling argumentative information and difficulties to differentiate between corresponding acts which can need a pragmatic level to enhance the recognition process.

As future work, we intend to integrate context-based and semantic features to improve the annotation results. We also project to investigate the annotated dataset in an extrinsic task such opinion question answering, argumentative discourse structure building and conversations summarization.

7 References

Alexandersson J., B. Buschbeck-Wolf, T. Fujinami, M. Kipp, S. Koch, E. Maier, N. Reithinger, B. Schmitz, and M. Siegel. 1998. Dialogue Acts in VERBMOBIL-2 (second edition). *Vm report 226*, DFKI GmbH, Universities of Berlin, Saarbrcken and Stuttgart.

Bangalore S., G. Di Fabbrizio and A. Stent. 2008. Learning the structure of task-driven human–human dialogs. *IEEE Transactions on Audio, Speech, and Language Processing*. 16(7):1249–1259.

BenDbabis S., H. Ghorbel, L. Belguith and M. Kallel. 2015. Automatic dialogue acts Annotation within Arabic debates. *16th International Conference on Intelligent Text Processing and Computational Linguistics* , April 14-20, Cairo, Egypt.

BenDbabis S., B. Rguii, H. Ghorbel and L. Belguith. 2016. Utterance Segmentation Using Conditional Random Fields. *27th International Business on Information Management Association*, May 3-5, 2016, Milano, Italy.

BenDbabis S., F. Mallek, H. Ghorbel and L. Belguith. 2012. Dialogue Acts Annotation Scheme within Arabic discussions. *Sixteenth Workshop on the Semantics and Pragmatics of Dialogue,* September 19-21, Paris, France.

Bunt H. 2009. The DIT++ taxonomy for functional dialogue markup. *In Proceedings of the AAMAS Workshop*, Budapest, May 12, 2009.

Bunt H. 1994. Context and Dialogue Control. THINK, 3:19-31.

Carletta J., S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. 2005. The AMI Meetings Corpus. *In Proceedings of the Measuring Behavior Symposium on "Annotating and measuring Meeting Behavior"*.

Carletta J. C. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2): 249-254.

Core M. and J. Allen. 1997. Coding Dialogs with the DAMSL annotation scheme. *In AAAI Fall Symposium on Communicative Action in Humans and Machines*, MIT, Cambridge, MA.

Coria S. R. and L. A. y Pineda. 2007. Prediction of Dialogue Acts on the Basis of Previous Act. *Procesamiento de Lenguaje Natural,* 39: 223 – 230.

Di Eugenio B., Z. Xie and R.o Serafin. 2014. Dialogue act classification, higher order dialogue structure, and instance-based learning. Dialogue and Discourse,1(2):1–24.

Elmadany A., S. M. Abdou and M. Gheith. 2015. Turn Segmentation Into Utterances For Arabic Spontaneous Dialogues And Instant Messages. *International Journal on Natural Language Computing,* 4(2). April 2015.

Elmadany A., S. M. Abdou, and M. Gheith. 2014. Arabic Inquiry-Answer Dialogue Acts Annotation Schema. *IOSR Journal of Engineering*, 4(12-V2):32-36.

Ezen-Can A. and K. E. Boyer. 2015. A tutorial dialogue system for real-time evaluation of unsupervised dialogue act classifiers: exploring system outcomes. *In International Conference on Artificial Intelligence in Education*. Springer International Publishing. Pages 105-114.

Galley M., K. McKeown, J. Hirschberg, and E. Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. *In ACL 2004*, Barcelona.

Geertzen J., V. Petukhova and H. Bunt. 2007. A Multidimensional Approach to Utterance Segmentation and Dialogue Act Classification. *In Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp.

Grosz B. J. and C. L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*. 12(3): 175-204.

Guz U., S. Cuendet, D. Hakkani-Tür and G. Tur. 2010. Multi-View Semi-Supervised Learning for Dialog Act Segmentation of Speech. *IEEE Transactions on Audio, Speech and Language Processing*: pages 320-329.

Hoque M. E., M. S. Sorower, M. Yeasin and M. M. Louwerse. 2007. What Speech Tells us about Discourse: The Role of Prosodic and Discourse Features in Dialogue Act Classification. *In IEEE International Joint Conference on Neural Networks (IJCNN)*, Orlando, FL.

Ivanovic E. 2005. Automatic Utterance Segmentation in Instant Messaging Dialogue. *Proceedings of the Australasian Language Technology Workshop*. Pages 241–249, Sydney, Australia.

Lafferty J., A. McCallum and F. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *In Proceedings of the 18th International Conference on Machine Learning*, pages 282-289.

Ondáš S. and J. Juhár. 2016. Towards human-machine dialog in Slovak . *International Conference on Systems, Signals and Image Processing (IWSSIP)*, Bratislava, pages 1-4.

Pavel K., L. Lenc and C. Cerisara. 2015. Semantic Features for Dialogue Act Recognition. *Statistical Language and Speech Processing*. Springer International Publishing, pages 153-163.

Kim S., L. Cavedon and T. Baldwin. 2012. Classifying dialogue acts in multi-party live chats. *In Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 463–472, Bali, Indonesia, November.

Ribeiro E., R. Ribeiro and D. M. de Matos. 2015. The Influence of Context on Dialogue Act Recognition. arXiv preprint arXiv:1506.00839.

Samei B., H. Li, F. Keshtkar, V. Rus, and A. C. Graesser. 2014. Context-based speech act classification in intelligent tutoring systems. *In International Conference on Intelligent Tutoring Systems*, pages 236-241. Springer International Publishing.

Schegloff E. A. and H. Sacks. 1973. Opening Up Closings. Semiotica, 7:289-327.

Shala, V. Rus and A. C. Graesser. 2010. L. Automated speech act classification in Arabic. *Subjetividad y Procesos Cognitivos*, 14: 284-292.

Shen S. S. and H. Y. Lee. 2016. Neural Attention Models for Sequence Classification: Analysis and Application to Key Term Extraction and Dialogue Act Detection. arXiv preprint arXiv:1604.00077.

Shriberg E., R. Dhillon, S. Bhagat, J. Ang and H. Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. *In Proceedings of the 5th SIGDIAL Workshop on Discourse and Dialogue*.

Silvia Q., V. Alexei and R. Giuseppe. 2011. Simultaneous dialog act segmentation and classification from human-human spoken conversations. *ICASSP 2011*: pages 5596-5599.

Sridhar V.K.R., S. Bangalore and S.S. Narayanan. 2007. Exploiting acoustic and syntactic features for prosody labeling in a maximum entropy framework. *NAACL-HLT*.

Sridhar V.K.R, S. Bangalore, and S.S. Narayanan. 2009. Combining lexical, syntactic and prosodic cues for improved online dialog act tagging. *Computer Speech and Language*, 23(4): 407-422. Elsevier Ltd.

Stolcke A., K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *In Computational Linguistics 2000*. 26(3): 339-373.

Swapna S. and J. Wiebe. 2010. Recognizing Stances in Ideological On-line Debates. *In Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116-124, Los Angeles, CA.

Webb N., M. Hepple and Y. Wilks. 2005. Dialogue act classification based on intra-utterance features. *In Proceedings of the AAAI Workshop on Spoken Language Understanding*, Pittsburgh, PA.

Yeh J.. 2016. Speech Act Identification Using Semantic Dependency Graphs with Probabilistic Context-Free Grammars. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 15(1).

Zhou Y., Q. Hu, J. Liu, and Y. Jia. 2015. Combining heterogeneous deep neural networks with conditional random fields for Chinese dialogue act recognition. *Neurocomputing*, 168: 408-417.

Procesamiento del Lenguaje Natural, Revista nº 60, marzo de 2018, pp. 61-68 recibido 22-11-2017 revisado 03-01-2018 aceptado 05-02-2018

From Sentences to Documents: Extending Abstract Meaning Representation for Understanding Documents

De Oraciones a Documentos: extendiendo Abstract Meaning Representation para la comprensión de textos

Paloma Moreda, Armando Suárez, Elena Lloret, Estela Saquete, Isabel Moreno
Department of Software and Computing Systems, University of Alicante
Apdo. de Correos 99 E-03080, Alicante, Spain
{moreda,armando,elloret,stela,imoreno}@dlsi.ua.es

Abstract: The overabundance of information and its heterogeneity requires new ways to access, process and generate knowledge according to the user's needs. To define an appropriate formalism to represent textual information capable to allow machines to perform language understanding and generation will be crucial for achieving these tasks. Abstract Meaning Representation (AMR) is foreseen as a standard knowledge representation that can capture the information encoded in a sentence at various linguistic levels. However, its scope only limits to a single sentence, and it does not benefit from additional semantic information that could help the generation of different types of texts. Therefore, the aim of this paper is to address this limitation by proposing and outlining a method that can extend the information provided by AMR and use it to represent entire documents. Based on our proposal, we will determine a unique, invariant and independent standard text representation, called canonical representation. From it and through a transformational process, we will obtain different text variants that will be appropriate to the users' needs.
Keywords: AMR, documents, canonical representation, user

Resumen: La sobreabundancia de información y su heterogeneidad requieren nuevas formas de acceder, procesar y generar conocimiento de acuerdo con las necesidades del usuario. Por ello, definir un formalismo adecuado para representar la información textual capaz de permitir a los ordenadores comprender y generar el lenguaje, es crucial para lograr esta tarea. Abstract Meaning Representation (AMR) es una representación del conocimiento estándar que puede capturar la información codificada en una oración en varios niveles lingüísticos. Sin embargo, su alcance se limita a una sola oración, y no se beneficia de la información semántica adicional que podría ayudar a la generación de diferentes tipos de textos. En este artículo propondremos un método que amplia la información proporcionada por AMR y la utiliza para representar documentos completos. En base a nuestra propuesta, definiremos una representación de texto estándar única, invariable e independiente, llamada representación canónica. A partir de la cual, y mediante un proceso de transformación, obtendremos diferentes variantes de texto que serán apropiadas para las necesidades de los usuarios.
Palabras clave: AMR, documentos, representación canónica, usuario

1 Introduction

In the context of the Digital Society, the over-abundance of information and its heterogeneity requires new ways to access, process and generate knowledge according to the user needs. In this regard, Human Language Technologies (HLT) play a key role in the analysis, processing and understanding of information. However, the progress made in HLT applications focuses on solving only specific tasks in specific domains, offering solutions from a partial and isolated perspective, without keeping a common model for knowledge extraction, and without considering the user needs as a cross-cutting and intrinsic as-

ISSN 1135-5948. DOI 10.26342/2018-60-7
© 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

pect in the process.

Therefore, the main goal of this paper is based on the need for conducting research into a new paradigm for text understanding that will allow us to determine a unique, invariant and independent standard text representation, called canonical representation. From this representation and through a transformational process, we will obtain different text variants that will be appropriate to the users' needs, so that these transformations can be applied to other HLT tasks, such as simplification, enrichment or summarization.

To that end, this paper defines the canonical representation of texts, and how it could be used to generate variations of texts is shown. Such representations are defined using as a basis the Abstract Meaning Representation (AMR) formalism (Banarescu et al., 2013) with improvements: extension of graph at document level and inclusion of additional information and annotation with the VerbNet set of roles (Schuler, 2006).

The remainder of this paper is organized as follows. Section 2 reviews previous work using AMR formalism. Next, Section 3 introduces the canonical representation of texts. Latter, Section 4 shows an example text together with its canonical representation and possible variations. Last, conclusions and future work are outlined.

2 Related Work

Among the specific formalisms for representing natural language at different linguistic levels (lexical, syntactical, semantic, etc.), AMR has gained popularity in the last years since this type of representation can capture semantic aspects of sentences, thus helping Natural Language Understanding and Generation. Although we can find research focused on developing visual tools to better understand AMR annotations (Saphra and Lopez, 2015) or the creation of AMR-annotated corpora (Banarescu et al., 2013; Vanderwende, Menezes, and Quirk, 2015) to be able to train parsers, previous literature has been mostly devoted to automatically address AMR semantic parsing in order to obtain the appropriate representation of a sentence following the AMR guidelines (Vanderwende, Menezes, and Quirk, 2015; Zhou et al., 2016; Goodman, Vlachos, and Naradowsky, 2016; Damonte, Cohen, and Satta, 2017). However, AMR has a great potential for HLT tasks, especially the ones related information generation (e.g., text summarization or natural language generation).

For instance, the use of AMR for text summarization is beneficial for producing abstractive summaries. In this manner, the approach proposed in Liu et al. (2015) partly address this task by building a summary graph from an AMR graph by a concept merging step. In their approach, the coreferent nodes of the graph were merged together. These nodes were either name entities or dates. The authors tested their method in newswire documents and compared the summaries generated from AMR gold-standard annotations with respect to use the output of an AMR parser (in particular JAMR (Flanigan et al., 2014)), and despite being differences in the results obtained, in both cases the results were state-of-the-art according to the summarization task, thus being a very promising method to integrate in abstractive summarization approaches. A similar idea is addressed in Dohare and Karnick (2017), where the authors try to overcome some of the limitations of the approach previously described. Whereas in Liu et al. (2015) a single summary graph from the story graph was extracted, assuming that all the important information from the graph could be extracted from a single subgraph, in Dohare and Karnick (2017), multiple subgraphs are extracted each focusing on information in a different part of the story. In this manner, a few important sentences are first selected and then a summary graph is built from the AMR representation. Finally, an existing text generation from AMR is used to finally produced the resulting summaries. The summarization approach was evaluated and compared to other baselines and approaches, improving the results of Liu et al. (2015) (51.3 vs. 44.3) for the ROUGE-1 F-measure metric.

Apart from text summarization, AMR has also been used to directly generate text. In Flanigan et al. (2016) a statistical method relying on discriminative learning is studied. First, a spanning tree is generated from the AMR and then tree-to-string decoder is applied to generate English, based on the probabilities given by a language model. On the other hand, the approach proposed in Pourdamghani, Knight, and Hermjakob (2016) addressed the problem of AMR-to-text as a

phrased-based machine translation problem. The proposed method learned to linearize AMR tokens into an English-like order. The aim is to induce an ordering function that takes any set of edge labels from AMR as input and produces a permutation of those labels. Several linearization methods were analyzed (e.g., taking into account the most common order for each role in the data, or using different binary classifiers to learn the order for each type of feature). The results achieved overperformed the previous results obtained in Flanigan et al. (2016).

Other formalisms for representing text have been proposed. In Martínez-Barco et al. (2013), a conceptual representation schema was proposed for decomposing natural language into smaller units that could be later combined to generate different types of text (such as summary, an enriched text, or a simplified text), taking into account users' needs. However, it was only a theoretical approach without any implementation, so despite being interesting, it could not be materialized and tested. Taking as a basis this conceptual model, and having analyzed that AMR may be an appropriate specific formalism to represent and generate language, we would like to combine the potentials of both of them by first extending the AMR representation to entire documents and enriching it with additional information, and then being able to generate different types of texts depending on users' needs (e.g. a summary, a simplified text, a schema), thus improving the accessibility of information for any type of user.

3 Canonical representation of texts

Our target is to define a standard representation of a text that allows us to generate different versions of it (summaries, simplifications, enrichments, etc.). In order to do this we are going to use the AMR language. In AMR, each sentence in a text is a single rooted and directed graph, that implies a semantic limited canonical representation of the sentence. Hence, taking AMR as a basis, we propose to enrich each sentence's AMR and use this new information to link as much as possible the sentence level graphs. In this manner, all the semantics of a document is added to express completely the meaning of a text. This would be particularly useful for expressing the meaning of text not only in a different manner, but also to be widely understood by any audience.

Our proposal is performed in two steps. Figure 1 shows the architecture of the system:

1. **Sentence level representation**: Generating AMR per sentence and enriching them with extra information.
   (a) Representing each sentence with AMR formalism using existing parsers, such as JAMR annotator (Flanigan et al., 2014).
   (b) Integrating VerbNet (Schuler, 2006) set of semantic roles, instead of PropBank owing to the fact that the latter is not able to generalize the meaning of the numeric roles, and thus finding cases in which, for instance, a location role for the verb "go" can be represented by ARG2 or ARG4. This problem disappears when the VerbNet set of roles is employed, as AMR originally defined in the PENMAN project (Langkilde and Knight, 1998).
   (c) **Temporal information resolution**. All the temporal expressions in the text will be detected and resolved using TIPSem system (Llorens, Saquete, and Navarro-Colorado, 2013). This system is based on morphosyntactic knowledge plus semantic knowledge, specifically, semantic networks and semantic roles. TIPSem is able to automatically annotate all the temporal information according to TimeML standard annotation scheme, that means annotating all the temporal expressions (TIMEX3), events (EVENT) and links between them.
   (d) Resolving **concepts and entities** in the text is tackled using Babelfy (Moro, Cecconi, and Navigli, 2014). This system addresses entity linking and word sense disambiguation in an unified-manner. Babelfly is a graph-based approach based on a loose identification that selects high-coherence semantic interpretations. It allows the extraction of information related to them, such as synonyms.

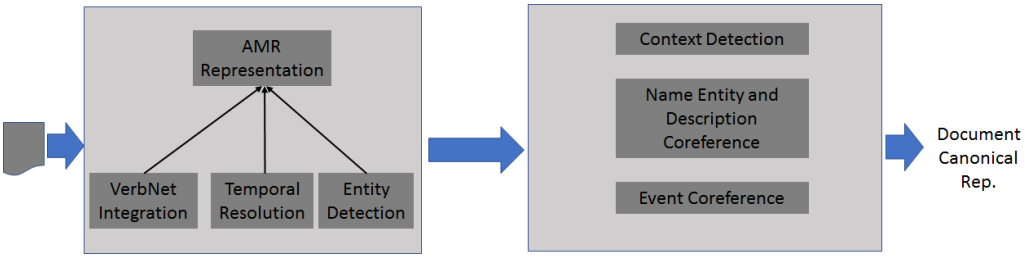

Figure 1: Architecture

2. **Document level representation**: Merging the different extended AMR's of the document by means of: context, name-entity coreference and event coreference.

   (a) Obtaining **context information** such as domain of the text and document creation time. In order to obtain the domain of the text our approach relies on calculating the more frequent domains in a text by agglomerating all the domains linked to all the words with a domain extracted from Babelnet in the previous step, and finally the list of labels is sorted according to the overall frequency.

   (b) **Coreference resolution**. All mentions referring to the same entity are extracted using an state-of-the-art tool, the Standford Coreference Resolution System (Clark and Manning, 2015). It tackles both pronominal and nominal coreference. The latter refers to definite descriptions, which are noun phrases introduced by a definite article and denoting a particular entity. This system implements an statistical mention-ranking model to iterate through each mention in the document to establish a coreference link with a preceding mention.

   (c) **Event Coreference resolution.** All those events mentions referring to the same real fact will be merged in the final graph in one single node in order to relate the different AMR's and simplify the graph of the whole document. Event coreference will be determined in a two clustering process. First, a temporal clustering will be performed, so all the events happening at the same time will be clustered together. After this, a semantic clustering is performed so the events are clustered using lexical semantics (lemmas and synonyms) and distributional semantic knowledge (word2vec) in order to resolve event coreference (Navarro-Colorado and Saquete, 2016). The temporal clustering is not a trivial task, but we are using the temporal information annotated by TipSEM system for this purpose. According to Tempeval-3 (UzZaman et al., 2013) evaluation, TipSEM system is obtaining an F1-score of 65.31% at temporal expression performance and an F1-score of 42.39% at temporal awareness regarding temporal relations. Regarding event coreference, we are using the system described at Navarro-Colorado and Saquete (2016), that combines both tasks with a final F1-score of 26.5% for the experiments involving temporal, lexical and distributional clustering, which improves the current state-of-the-art systems and shows a significant advance in the Cross-Document Event Ordering task.

3.1 The canonical formalism

In order to build a complete text graph from the AMR representation of each individual sentence, and considering the fact that all the nodes in AMR representation has a variable, once the coreference is resolved, all the nodes referring to the same thing (entities or events) will be identify with the same variable. Apart from this, our extended AMR representation will introduced these *new* relations or edges to the nodes (when necessary):

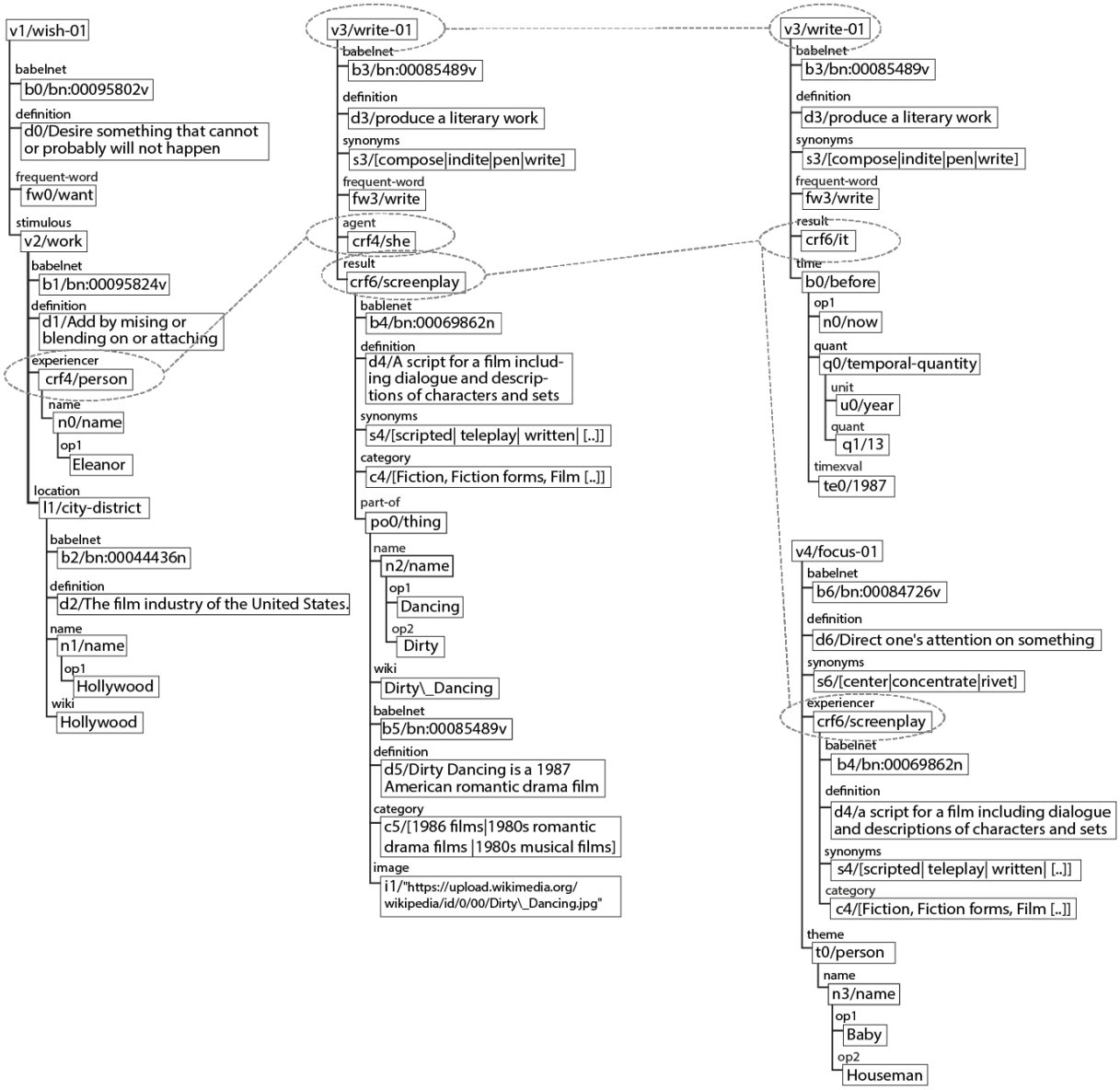

Figure 2: Four sentences AMR graphs with merging points

(a) *:dct* – document creation time;

(b) *:topic* – document topic;

(c) *:timexval* – ISO temporal value of temporal expressions;

(d) *:category_related* – category of the concept;

(e) *:domain* – ID of BabelDomain;

(f) *:babelnet* – ID from BabelNet for each concept or entity;

(g) *:definition* – explanation for a term;

(h) *:category* – the semantic class associated to the term;

(i) *:image* – the image associated to a concept or entity;

(j) *:synonyms* – alternative terms for the same concept or entity;

(k) *:frequent-word* – synonym that is most commonly used.

4 Case Example

This section shows a possible output from a text example formed by four sentences, as well as possible transformations derived from the original text.

Typically, an AMR representation in PENMAN notation would give us something similar to the next output for the sentence "*Eleanor wished to work in Hollywood.*", including some of the enrichments mentioned before:

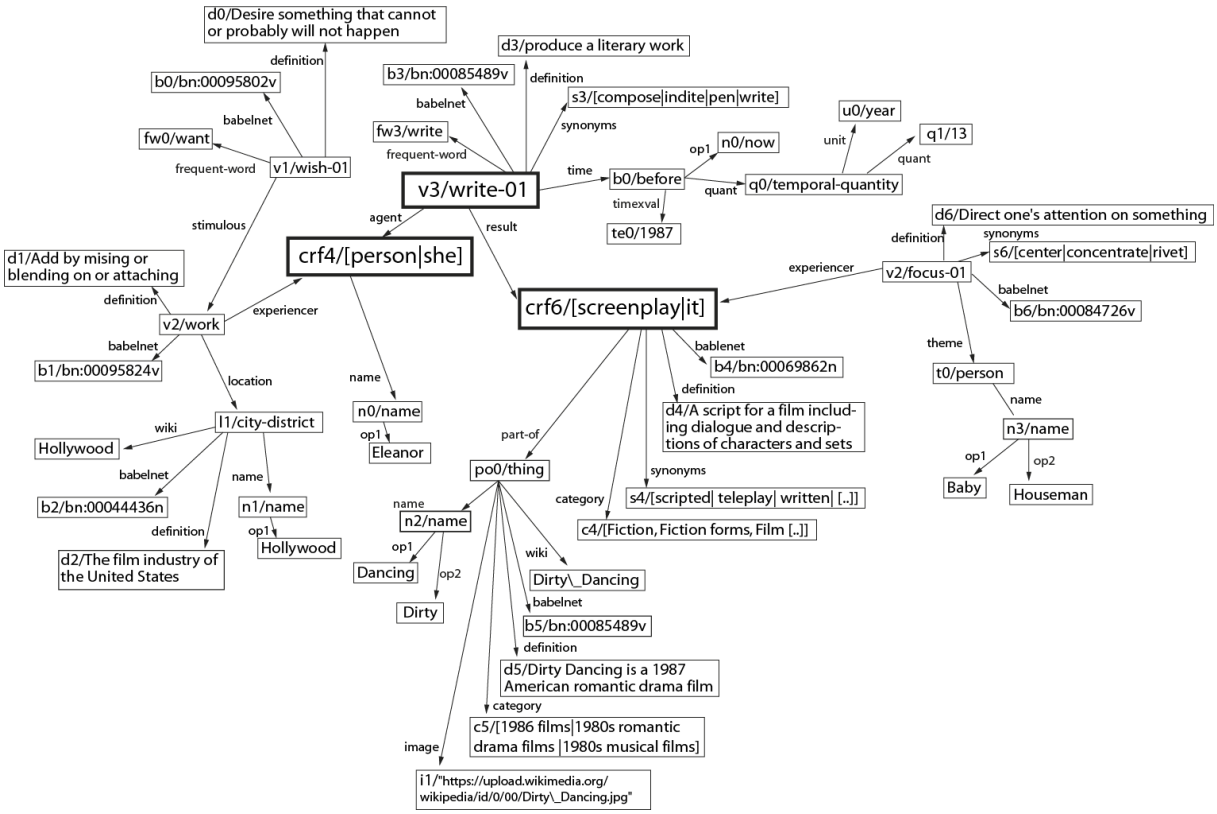

Figure 3: Whole text graph after merging

```
%Eleanor wished to work in Hollywood.
(v1 / wish-01
    :babelnet (b0 / bn:00095802v)
    :definition (d0 / "Desire something that
    cannot or probably will not happen")
    :frequent-word (fw0 / "want")
    :stimulous (v2 / work-01
         :babelnet (b1 / bn:00095824v)
         :definition (d1 / "Add by mising or
            blending on or attaching")
         :experiencer (crf4 / person
            :name (n0 / name :op1 "Eleanor"))
         :location (l5 / city-district
            :babelnet(b2 / bn:00044436n)
            :definition (d2 / "The film
               industry of the United States.")
            :name (n1 / name :op1 "Hollywood")
            :wiki "Hollywood")))
```

Suppose a text with these four sentences: "*Eleanor wished to work in Hollywood. She wrote the screenplay of Dirty Dancing. It was written thirty years ago. The screenplay was focused on Baby Houseman.*". Given such input four direct graphs can be produced such as those shown in Figure 2. We also mark the nodes of the graphs that represent the same concept. These are merging points which allows us to produce the final canonical representation of the whole text.

For example, (*v3/write*) is an AMR node with a variable v3 appearing in two of the sentences, as well as the pairs (*crf4/person – crf4/she*) and (*crf6/it – crf6/screenplay*), labeled by co-reference resolution. Using these merging points a final graph is shown in Figure 3.

Using the canonical representation, the text could be transformed without losing its meaning through the navigation of the AMR text graph. Each transformation could be appropriate in a particular situation or for a specific purpose. Figure 4 shows two possible transformations for our example. The first transformation is an extended text containing the explanation for the term "screenplay" and using synonyms of the original words, whereas the second shows a headline. The former could be useful to someone who is not an expert or has reading comprehension difficulties.

5 Conclusions

In this paper, a text representation using Abstract Meaning Representation (AMR) is proposed. Originally, AMR is intended to represent the concepts and their relationships of one sentence only. In this manner, the set of sentences that compose an entire text results in a set of disconnected AMR graphs. Our proposal consists of an architecture and

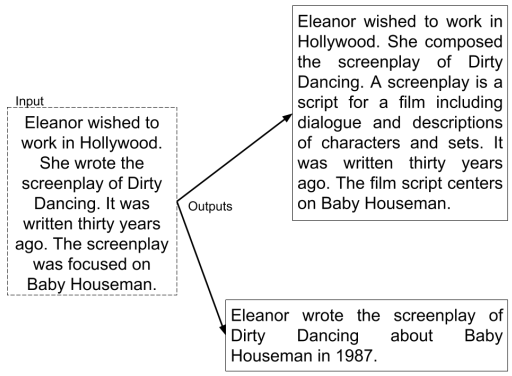

Figure 4: Examples of possible inflexions

a method to add new data to offer more semantic information at sentence level, but to link and merge such graphs too. The final goal is to achieve a unique, invariant and independent standard representation of entire documents. Such canonical representation will allow us to generate new variants of the analyzed text such us summaries, simplifications, etc. to satisfy different user's needs.

In this paper we illustrate this enrichment over four sentences representing a short document, showing the resulting AMR graphs and how they are merged and, finally, an example of text variants that could be generated from this text canonical representation.

In order to do this, several NLP tools were used to enrich the basic AMR representation such as semantic role labelling, entity identification and resolution, temporal information resolution, document categorization and coreference resolution. These added variables and values permit to find linking points on graphs in order to relate as much sentences as possible, as the intuitive notion of what a document is seem to point out. Since the errors made by these NLP tools may affect the accuracy of the information to be represented in the AMR graph, a validation process will be done to detect wrong information with the purpose of avoiding creating a noisy AMR graph. This could be done by first using available annotated corpora as input for the AMR graph and then, comparing the result with the one obtained when using the automatic tools.

In the immediate future, we will focus on analyzing new sources of information to add and managing user's profiles to offer the most useful text variants for them. For the evaluation of our final system, internal metrics will be provided, and a pedagogical point of view will be explored in the form of automatic assessments generation.

Acknowledgments

Research partially supported by the Spanish Government (grants TIN2015-65100-R; TIN2015-65136-C02-2-R).

References

Banarescu, L., C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, August. Association for Computational Linguistics.

Clark, K. and C. D. Manning. 2015. Entity-centric coreference resolution with model stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1415, Beijing, China, July. Association for Computational Linguistics.

Damonte, M., S. B. Cohen, and G. Satta. 2017. An incremental parser for abstract meaning representation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 536–546, Valencia, Spain, April. Association for Computational Linguistics.

Dohare, S. and H. Karnick. 2017. Text summarization using abstract meaning representation. *CoRR*, abs/1706.01678.

Flanigan, J., C. Dyer, N. A. Smith, and J. Carbonell. 2016. Generation from abstract meaning representation using tree transducers. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 731–739, San Diego, California, June. Association for Computational Linguistics.

Flanigan, J., S. Thomson, J. Carbonell, C. Dyer, and N. A. Smith. 2014. A dis-

criminative graph-based parser for the abstract meaning representation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436, Baltimore, Maryland, June. Association for Computational Linguistics.

Goodman, J., A. Vlachos, and J. Naradowsky. 2016. Noise reduction and targeted exploration in imitation learning for abstract meaning representation parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–11, Berlin, Germany, August. Association for Computational Linguistics.

Langkilde, I. and K. Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *COLING-ACL '98, Proceedings of the Conference*, pages 704 – 710, Montreal, Canada, August. Association for Computational Linguistics.

Liu, F., J. Flanigan, S. Thomson, N. Sadeh, and N. A. Smith. 2015. Toward abstractive summarization using semantic representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1077–1086, Denver, Colorado, May–June. Association for Computational Linguistics.

Llorens, H., E. Saquete, and B. Navarro-Colorado. 2013. Applying Semantic Knowledge to the Automatic Processing of Temporal Expressions and Events in Natural Language. *Information Processing & Management*, 49(1):179–197.

Martínez-Barco, P., A. F. Rodríguez, D. Tomás, E. Lloret, E. Saquete, F. Llopis, J. Peral, M. Palomar, J. M. G. Soriano, and M. T. Romá-Ferri. 2013. LEGOLANG: técnicas de deconstrucción aplicadas a las tecnologías del lenguaje humano. *Procesamiento del Lenguaje Natural*, 51:219–222.

Moro, A., F. Cecconi, and R. Navigli. 2014. Multilingual word sense disambiguation and entity linking for everybody. In *Proceedings of the 2014 International Conference on Posters & Demonstrations Track - Volume 1272*, ISWC-PD'14, pages 25–28, Aachen, Germany, Germany. CEUR-WS.org.

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

Pourdamghani, N., K. Knight, and U. Hermjakob. 2016. Generating english from abstract meaning representations. In *Proceedings of the 9th International Natural Language Generation conference*, pages 21–25, Edinburgh, UK, September 5-8. Association for Computational Linguistics.

Saphra, N. and A. Lopez. 2015. Amrica: an amr inspector for cross-language alignments. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 36–40, Denver, Colorado, June. Association for Computational Linguistics.

Schuler, K. K. 2006. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

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

Vanderwende, L., A. Menezes, and C. Quirk. 2015. An amr parser for english, french, german, spanish and japanese and a new amr-annotated corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 26–30, Denver, Colorado, June. Association for Computational Linguistics.

Zhou, J., F. Xu, H. Uszkoreit, W. QU, R. Li, and Y. Gu. 2016. Amr parsing with an incremental joint model. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 680–689, Austin, Texas, November. Association for Computational Linguistics.

Tesis

Procesamiento del Lenguaje Natural, Revista nº 60, marzo de 2018, pp. 71-74 recibido 03-11-2017 revisado 07-01-2018 aceptado 04-02-2018

Nuevos Paradigmas de Análisis Basados en Contenidos para la Detección del Spam en RRSS

New approaches for content-based analysis towards Online Social Network spam detection

Enaitz Ezpeleta
Mondragon Unibertsitatea
Goiru kalea 2, 20500 Arrasate, Spain
eezpeleta@mondragon.edu

Resumen: Tesis doctoral realizada por Enaitz Ezpeleta Gallastegi en Mondragon Unibertsitatea, dentro del grupo de Sistemas Inteligentes para Sistemas Industriales, dirigida por los Doctores Urko Zurutuza Ortega (Mondragon Unibertsitatea) y José María Gómez Hidalgo (Pragsis Technologies). La defensa se efectuó el 30 de septiembre de 2016 en Arrasate. El tribunal estuvo conformado por el Dr. Manel Medina Llinas (Universitat Politecnica de Catalunya), el Dr. Magnus Almgren (Chalmers University of Technology), el Dr. Igor Santos Grueiro (Universidad de Deusto), el Dr. José Ramón Méndez Reboredo (Universidad de Vigo) y el Dr. D. Iñaki Garitano Garitano (Mondragon Unibertsitatea). La tesis obtuvo una calificación de Sobresaliente Cum Laude y la mención "Doctor Europeus".
Palabras clave: Spam, redes sociales, PLN, análisis de sentimiento, polaridad, reconocimiento de personalidad, seguridad

Abstract: PhD Thesis written by Enaitz Ezpeleta Gallastegi at Mondragon University supervised by Dr. Urko Zurutuza Ortega (Mondragon Unibertsitatea) and Dr. José María Gómez Hidalgo (Pragsis Technologies). The viva voce was held on the 30th September 2016 and the members of the commission were Dr. Manel Medina Llinas (Universitat Politecnica de Catalunya), el Dr. Magnus Almgren (Chalmers University of Technology), el Dr. Igor Santos Grueiro (Universidad de Deusto), el Dr. José Ramón Méndez Reboredo (Universidad de Vigo) y el Dr. D. Iñaki Garitano Garitano (Mondragon Unibertsitatea). The thesis obtained the grade of Excellent Cum Laude and the mention "Doctor Europeus".
Keywords: Spam, online social networks, NLP, sentiment analysis, polarity, personality recognition, security

1 Introducción

Las campañas de correo electrónico no deseado siguen siendo una de las mayores amenazas que afectan a millones de usuarios al día. Aunque las técnicas de detección de spam son capaces de detectar un porcentaje muy alto de spam, el problema está lejos de ser solventado, sobre todo por la cantidad tan alta de tráfico spam existente entre el tráfico global de correo electrónico, y las nuevas estrategias utilizadas por los atacantes.

Además, el auge del número de usuarios de las redes sociales (RRSS) en Internet (como Facebook, Twitter, Instagram...), muchos de los cuales publican mucha información de forma abierta en sus perfiles, han proporcionado que estos sitios se conviertan en objetivos atractivos para los atacantes, principalmente por dos razones: la posibilidad de explotar la información pública almacenada en los perfiles de los usuarios, y por la facilidad para entrar en contacto directo con los usuarios mediante los perfiles, los grupos, las páginas... Como consecuencia, cada vez se detectan más actividades ilegales en estas redes. Entre ellas, el spam es una de las que mayor impacto causa. Actualmente, la venta comercial, la creación de alarma social, campañas de sensibilización, distribución de *malware*, etc. son los principales objetivos de los men-

ISSN 1135-5948. DOI 10.26342/2018-60-8
© 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

sajes de spam. Tomando en cuenta esto, partimos de la hipótesis de que su forma de ser escrito conlleva una intencionalidad implícita, que el autor desea explotar para su detección.

Los principales objetivos de esta tesis son: (1) demostrar que es posible desarrollar spam personalizado usando información publicada en redes sociales que eluda los sistemas actuales de detección; y (2) diseñar y validar nuevos métodos para la detección y filtrado de spam usando técnicas de Procesamiento de Lenguaje Natural (PLN). Además, estos sistemas deberán ser efectivos con el spam que se propaga dentro de las redes sociales.

2 Organización de la Tesis

Este trabajo de tesis está organizado en los siguientes capítulos:

1. Indroducción: Se explica la motivación para la realización del trabajo, así como los objetivos y las hipótesis a los que se intenta dar respuesta.
2. Estado de la cuestión: En este capítulo se resume como se ha abordado la detección identificando los diferentes sistemas actuales. También se presentan diferentes propuestas basadas en técnicas de PLN, y se realiza una introducción a los problemas de seguridad de las redes sociales.
3. Efectividad del spam personalizado: Capítulo en el que se presenta el trabajo realizado de cara a demostrar que es posible crear spam personalizado capaz de saltarse los sistemas anti-spam actuales.
4. Análisis de sentimiento: Se resume como se puede conseguir mejorar el filtrado spam utilizando la polaridad de los mensajes.
5. Reconocimiento de personalidad: En este capítulo se describe el modelo creado utilizando las dimensiones de la personalidad del texto.
6. Combinación de ambas técnicas: Presentación del tercer modelo donde se combina la utilización de técnicas de análisis de sentimiento y reconocimiento de personalidad.
7. Conclusiones: Capítulo en el que se resumen las aportaciones más significativas del trabajo, así como las lineas futuras identificadas.

3 Contribuciones y Resultados Experimentales

Para validar el primer objetivo de este trabajo se ha diseñado y desarrollado un sistema que permite enviar campañas de spam personalizado (Ezpeleta, Zurutuza, y Hidalgo, 2015; Ezpeleta, Zurutuza, y Gómez Hidalgo, 2016c). Mediante este sistema, se ha podido demostrar que utilizando información pública personal de los usuarios de las redes sociales (Facebook) es posible crear spam personalizado que alcance ratios de click-through muy superiores a los del spam. Para ello el sistema recolecta direcciones de correo electrónico en Internet, para después extraer la información personal guardada de forma pública por el propietario de la cuenta vinculada a esa dirección en Facebook. Con esa información se crean diferentes perfiles que son usados para enviar correos electrónicos personalizados. Finalmente se han desarrollado experimentos donde se demuestra la eficacia de este tipo de spam frente al spam típico/común. Esta información sirve para subrayar el problema que supone publicar información personal en las redes sociales, así como para entender posibles riesgos futuros a los que la comunidad científica se deberá enfrentar, como es el caso del spam personalizado. Y por último ofrece las bases para el desarrollo de sistemas capaces de detectar este tipo de mensajes, tal y como se ha hecho en la segunda fase de esta tesis.

En la segunda parte de la tesis se presentan tres nuevos modelos para el filtrado de nuevos tipos de spam. Estos métodos tienen como objetivo detectar la intencionalidad comercial no evidente en los textos que luego ayuden a clasificarlos. Siendo este el objetivo, se identificó la necesidad de utilizar técnicas de PLN para analizar el contenido de los mensajes y poder extraer información que pudiera ser interesante a la hora de detectar mensajes no deseados. Debido al auge experimentado por estas técnicas en los últimos años, se ha podido realizar un estudio exhaustivo de gran variedad de técnicas para identificar las que mejor resultado ofrecían para este objetivo. De esta forma se han diseñado dos modelos independientes, donde uno de ellos utiliza Análisis de Sentimiento (AS) y el otro el Reconocimiento de Personalidad (RP) de los mensajes para mejorar la detección del spam.

El AS realizado, extrayendo la polaridad (mensaje positivo, negativo o neutro) de ca-

da mensaje, ofrece a la comunidad científica bases para demostrar que, teniendo los mensajes spam en su mayoría intención de vender productos, el contenido de los mensajes se escribe con una connotación más positiva que en los mensajes legítimos. Gracias a ello, al añadir esta información a los clasificadores de spam, se ha demostrado, tal y como se recoge en (Ezpeleta, Zurutuza, y Gómez Hidalgo, 2016a; Ezpeleta, Zurutuza, y Hidalgo, 2016), que los resultados obtenidos mejoran sustancialmente. Es decir, se ha demostrado que el AS ayuda a mejorar los resultados del filtrado de mensajes no deseados.

En el caso del segundo modelo presentado en (Ezpeleta, Zurutuza, y Gómez Hidalgo, 2016b; Ezpeleta, Zurutuza, y Gómez Hidalgo, 2016), se han mejorado los resultados del filtrado spam añadiendo información sobre la personalidad de cada mensaje, demostrando que las técnicas de RP también resultan de interés a la hora de mejorar los sistemas de detección de spam actuales.

Con la presentación de estos dos nuevos métodos, se ofrece tanto a la comunidad científica, así como a las empresas y organismos del sector, la posibilidad de ofrecer sistemas anti spam más eficaces a los usuarios, aportando seguridad y privacidad a las millones de personas que todos los días sufren las campañas de correo electrónico no deseados.

Finalmente, se ha presentado un nuevo modelo para la detección de spam donde se combinan los dos modelos anteriormente descritos, consiguiendo un sistema más eficaz tal y como se presenta en (Ezpeleta, Zurutuza, y Gómez Hidalgo, 2016d; Ezpeleta et al., 2017; Ezpeleta, Zurutuza, y Gómez Hidalgo, 2017). De esta forma, se demuestra que la combinación de técnicas de AS y RP mejora los resultados de las técnicas actuales de filtrado de spam.

Cabe destacar que las tres técnicas presentadas han sido validadas utilizando diferentes tipos de spam como son el spam en emails, spam en mensajes SMS y spam social o spam recogido en las redes sociales, y además han sido utilizados más de un conjunto de datos por cada tipo, con el objetivo de contrastar y refrendar la validez de los resultados obtenidos.

3.1 Resultados: eficacia del spam personalizado

Para validar el primero de los objetivos, se extrajeron direcciones de correo electrónico a través de un famoso buscador, y se contrastó la existencia de una vinculación a la red social Facebook de cada una de ellas, obteniendo una base de 22.654 usuarios con los cuales se pudieron crear perfiles para llevar a cabo el envío de diferentes campañas.

Los resultados demuestran que el spam personalizado es más eficaz que el spam habitual. Esto se refleja sobre todo en el porcentaje de usuarios que hacen click en la URL personalizada que se incluye en el contenido del correo enviado, siendo 18 veces más alto en el caso del spam personalizado, con un *click-through* del spam típico de un 0,41 % y un 7,62 % en el caso del personalizado.

3.2 Resultados: nuevos modelos para la detección de Spam

Una vez demostrado el riesgo que suponen las nuevas técnicas de creación de spam, se han diseñado y desarrollado tres nuevos modelos para la detección de nuevos tipos de spam. A la hora de realizar los experimentos para evaluar la eficacia de estos modelos, se han aplicado diferentes clasificadores tanto sobre los conjuntos de datos originales de cada tipo de spam, así como sobre los conjuntos de datos creados después de añadir los atributos creados con resultados de las diferentes técnicas utilizadas (AS, RP y combinación de ambas). Finalmente se ha llevado a cabo una comparativa en términos de precisión y el número de falsos positivos.

La Figura 1 muestra la precisión máxima obtenida en los distintos tipos de spam y utilizando los tres modelos presentados en este trabajo. El mejor resultado ha sido obtenido con el modelo que combina ambas técnicas (AS y RP). Cabe destacar que en el caso del número de falsos positivos, este se reduce significativamente en la mayoría de los casos.

4 Conclusiones

Al ser el spam un problema que afecta diariamente a millones de usuarios, la presentación de este tipo de modelos ayuda a que la experiencia de los usuarios vaya mejorando, y que dichos usuarios no sufran de posibles peligros derivados de este tipo de ataques contra su seguridad y privacidad.

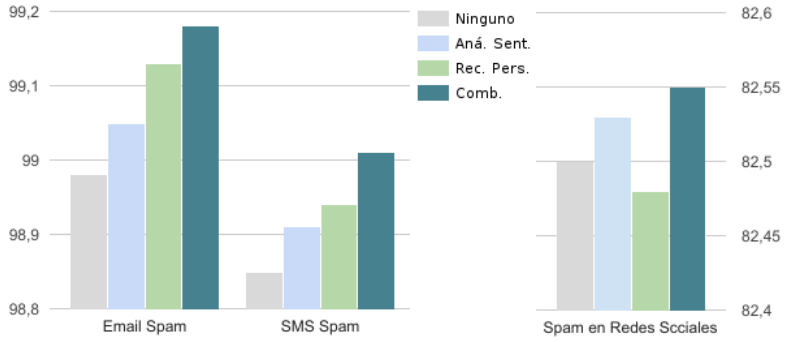

Figura 1: Comparativa de las precisiones obtenidas

En este trabajo se demuestra el potencial de las redes sociales a la hora de crear spam personalizado, el cual no es detectado por los sistemas de detección actuales. Tras presentar tres modelos novedosos en el ámbito de análisis de contenido para la detección del spam, se demuestra que se pueden mejorar los resultados de los sistemas actuales tanto en spam en emails, así como en mensajes SMS, y también en el spam que se propaga dentro de las redes sociales.

Muestra de la aplicabilidad de estos métodos en entornos reales es que actualmente, dentro del proyecto SocialSPAM (PI_2014_1_102), financiado por el Gobierno Vasco, se está desarrollando una aplicación nativa para Facebook. Esta herramienta analiza los mensajes de los usuarios, utilizando los métodos presentados en este trabajo, con el objetivo de detectar posibles mensajes spam, y filtrarlos.

Agradecimientos

Este trabajo ha sido realizado en el grupo de Sistemas Inteligentes para Sistemas Indutriales (Mondragon Unibertsitatea) en el proyecto SocialSPAM(PI_2014_1_102) ambos parcialmente financiados por el Departamento de Educación, Política Lingüística y Cultura del Gobierno Vasco.

Bibliografía

Ezpeleta, E., I. Garitano, I. Arenaza-Nuño, U. Zurutuza, y J. M. Gómez Hidalgo. 2017. Novel comment spam filtering method on youtube: Sentiment analysis and personality recognition. En *Proceedings of Current Trends In Web Engineering - ICWE 2017 International Workshops*.

Ezpeleta, E., U. Zurutuza, y J. M. Gómez Hidalgo. 2016a. Does sentiment analysis help in bayesian spam filtering? En *Springer Int. Publishing*, páginas 79–90.

Ezpeleta, E., U. Zurutuza, y J. M. Gómez Hidalgo. 2016b. Short messages spam filtering using personality recognition. En *Proceedings of the 4th Spanish Conference on Information Retrieval*, CERI '16, páginas 1–7, New York, NY, USA. ACM.

Ezpeleta, E., U. Zurutuza, y J. M. Gómez Hidalgo. 2016c. A study of the personalization of spam content using facebook public information. *Logic Journal of IGPL*.

Ezpeleta, E., U. Zurutuza, y J. M. Gómez Hidalgo. 2016d. Using personality recognition techniques to improve bayesian spam filtering. *Procesamiento del Lenguaje Natural*, 57:125–132.

Ezpeleta, E., U. Zurutuza, y J. M. Gómez Hidalgo. 2017. Short messages spam filtering combining personality recognition and sentiment analysis. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*. In press.

Ezpeleta, E., U. Zurutuza, y J. M. Gómez Hidalgo. 2016. Los spammers no piensan: usando reconocimiento de personalidad para el filtrado de spam en mensajes cortos. En *Actas de la XIV Reunión Española sobre Criptología y Seguridad de la Información*.

Ezpeleta, E., U. Zurutuza, y J. M. G. Hidalgo. 2015. An analysis of the effectiveness of personalized spam using online social network public information. En *Springer Int. Publishing*, páginas 497–506.

Ezpeleta, E., U. Zurutuza, y J. M. G. Hidalgo. 2016. Short messages spam filtering using sentiment analysis. En *Springer Int. Publishing*, páginas 142–153.

Procesamiento del Lenguaje Natural, Revista nº 60, marzo de 2018, pp. 75-78 recibido 21-11-2017 revisado 07-01-2018 aceptado 07-02-2018

Interfaces de Lenguaje Natural para la Consulta y Recuperación de Información de Bases de Conocimiento Basadas en Ontologías

Natural Language Interfaces for Querying and Retrieving Information from Ontology-based Knowledge Bases

Mario Andrés Paredes Valverde
Universidad de Murcia
Facultad de Informática Campus Espinardo
Espinardo, 30100, Murcia, España
marioandres.paredes@um.es

Resumen: Tesis doctoral titulada "Interfaces de lenguaje natural para la consulta y recuperación de información de bases de conocimiento basadas en ontologías ", defendida por Mario Andrés Paredes Valverde en la Universidad de Murcia y elaborada bajo la dirección de los doctores Rafael Valencia García (Universidad de Murcia) y Miguel Ángel Rodríguez García (King Abdullah University of Science & Technology). La defensa tuvo lugar el 23 de mayo de 2017 ante el tribunal formado por los doctores Juan Miguel Gómez Berbís (Presidente, Universidad Carlos III de Madrid), Francisco García Sánchez (Secretario, Universidad de Murcia) y la doctora Catalina Martínez Costa (Vocal, Medical University of Graz) y la tesis obtuvo la mención Cum Laude y Doctor Internacional.
Palabras clave: Procesamiento de lenguaje natural, web semántica, linked data

Abstract: Ph.D. thesis entitled "Natural language interfaces for querying and retrieving information from ontology-based knowledge bases" written by Mario Andrés Paredes Valverde at the University of Murcia under the supervision of the Ph.D. Rafael Valencia García (University of Murcia) and Ph.D. Miguel Ángel Rodríguez García (King Abdullah University of Science & Technology). The viva voice was held on the 23rd May 2017 and the members of the commission were the Ph.D. Juan Miguel Gómez Berbís (President, University Carlos III of Madrid), Ph.D. Francisco García Sánchez (Secretary, University of Murcia) and Ph.D. Catalina Martínez Costa (Vocal, University of Graz) and the thesis obtained the mention Cum Laude and International Doctor.
Keywords: Natural language processing, semantic web, linked data

1 Introducción

El exponencial crecimiento de información disponible en la web e intranets ha dado paso a la necesidad de contar con mecanismos capaces de procesar y comprender dicha información y con ello resolver necesidades específicas. Ante esta situación surge la web semántica, la cual, de acuerdo con Berners-Lee et al. (2001) añade a la información de la web actual una estructura bien definida a través de un conjunto de atributos, valores y relaciones, para lo cual emplea una de las tecnologías más sobresalientes de su arquitectura, que son las ontologías. Diversos individuos y organizaciones de dominios tales como las finanzas (Salas-Zárate et al., 2016) y servicios en la nube (Rodríguez-García et al., 2014) han adoptado las ontologías para publicar su información. Sin embargo, uno de los enfoques más extendidos para el acceso a esta información es el lenguaje formal de consulta SPARQL cuyo uso demanda un alto nivel de conocimiento en tecnologías como RDF y expresiones de lenguaje de consulta, así como el conocimiento previo de la estructura de datos de la base de conocimiento subyacente.

Ante estos hechos, existe la necesidad de hacer accesible la información de la web semántica a todo tipo de usuarios, sean expertos u ocasionales. De acuerdo con Cimiano et al. (2008) el paradigma de recuperación de información basado en lenguaje natural es

ISSN 1135-5948. DOI 10.26342/2018-60-9 © 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

generalmente considerado como el más intuitivo desde un punto de vista de uso, pues oculta al usuario la formalidad de una base de conocimientos basada en ontologías, así como el lenguaje de consulta ejecutable, permitiendo a los usuarios emplear todo el poder comunicativo del lenguaje natural en lugar de verse forzados a utilizar un lenguaje limitado.

En esta tesis doctoral se propone una solución basada en lenguaje natural y ontologías para la consulta y recuperación de información de bases de conocimiento. La solución propuesta aprovecha la tecnología de la web semántica de dos maneras. La primera de ellas consiste en procesar la ontología de la base de conocimiento para generar un vocabulario que le permita conocer los términos comúnmente utilizados por los usuarios en el dominio modelado, y de esta manera poder relacionar los elementos contenidos en la pregunta del usuario con aquellos descritos en la base de conocimiento. La segunda, consiste en utilizar un modelo ontológico independiente del dominio para representar tanto la estructura sintáctica de la pregunta, como el contexto de esta en términos de la base de conocimiento. Para obtener tal representación, se aplican técnicas de procesamiento de lenguaje natural (PLN), entre las que destaca el análisis de dependencias. A través de esta técnica se obtiene una representación sintáctica de la pregunta que guarda una estrecha relación con las tripletas RDF que forman el patrón de grafos a ser obtenido de la base de conocimiento. Este hecho ayuda en gran medida a generar las consultas SPARQL respectivas con base en un conjunto de plantillas de consulta independientes del dominio. A continuación, se describe de manera general el trabajo de investigación doctoral.

2 Objetivos

El objetivo principal de esta tesis es desarrollar soluciones basadas en tecnologías de procesamiento de lenguaje natural y web semántica que permitan reducir la brecha existente entre el usuario y las bases de conocimiento a través del lenguaje natural. Con respecto a los objetivos específicos de la tesis, estos se resumen de la siguiente manera:

1. Diseño e implementación de un modelo ontológico independiente del dominio para la representación de la estructura sintáctica y contexto de la pregunta en lenguaje natural.
2. Diseño de la arquitectura de una interfaz de lenguaje natural para bases de conocimiento basadas en ontologías.
3. Diseño e implementación de un proceso de análisis de preguntas basado en técnicas de procesamiento de lenguaje natural y web semántica.
4. Diseño e implementación de un proceso de generación de consultas SPARQL a partir de una representación semántica de la pregunta en lenguaje natural.
5. Validación de los resultados obtenidos por medio de bases de conocimiento basadas en Linked Data.

3 Estructura de la tesis

La tesis se ha organizado en 6 capítulos que se describen a continuación.

Capítulo 1. Este capítulo provee una breve introducción a las motivaciones del trabajo de investigación y a la metodología seguida para cumplir con los objetivos establecidos.

Capítulo 2. Esta sección proporciona una descripción del estado actual de las tecnologías involucradas en la investigación, que son web semántica, PLN e interfaces de lenguaje natural.

Capítulo 3. Este capítulo discute a detalle la principal motivación para llevar a cabo la investigación. Además, provee tanto el objetivo general como los objetivos específicos establecidos. Finalmente, describe la metodología seguida en esta investigación.

Capítulo 4. Este capítulo describe la arquitectura y funcionamiento general de la interfaz de lenguaje natural para bases de conocimiento basadas en ontologías propuesta en la tesis. También, describe el modelo ontológico de la pregunta que permite describir su estructura y contexto.

Capítulo 5. Este capítulo describe los experimentos de evaluación realizados para medir la efectividad de la interfaz de lenguaje natural, la cual se base en su capacidad de proveer la respuesta correcta a una pregunta en lenguaje natural a partir de una base de conocimientos. Estos experimentos se llevaron a cabo en dos bases de conocimiento con el objetivo adicional de comprobar la portabilidad de la interfaz.

Capítulo 6. Este apartado describe las conclusiones, y discute las principales contribuciones y limitaciones del trabajo

realizado, así como las posibles vías futura que permitan direccionarlas.

4 Contribuciones

Las principales contribuciones de esta tesis doctoral se resumen a continuación.

Modelo ontológico de la pregunta. Modelo ontológico que permite describir la estructura sintáctica de la pregunta, así como el contexto de esta en términos de la base de conocimiento del dominio y de las relaciones existentes entre ellos. La obtención de la estructura sintáctica de la pregunta se basa en la técnica de análisis de dependencias. Esta técnica obtiene relaciones binarias entre los elementos de la pregunta, las cuales, gracias al modelo de la pregunta, pueden ser representadas en forma de tripletas sujeto-predicado-objeto.

Adaptación de una clasificación de preguntas y respuestas al contexto de las bases de conocimiento basadas en ontologías. Esta contribución consiste en adaptar la clasificación de preguntas propuesta por Moldovan et al. (2000). Esta adaptación consistió en sustituir los tipos de respuesta esperados por clases establecidas en ontologías y vocabularios que han sido ampliamente adoptados por individuos y organizaciones para representar su información. Gracias a este proceso, es posible delimitar el espacio de búsqueda de tal forma que los recursos a obtener deberán ser solo aquellos que correspondan con el tipo de datos establecido, o sean subclase de este.

Conjunto de plantillas de tripletas RDF. Esta contribución consiste en un conjunto de plantillas de tripletas RDF las cuales corresponden a las relaciones semánticas existentes entre los elementos de interés identificados en la pregunta que son representados mediante el modelo ontológico de la pregunta propuesta en esta tesis. El conjunto de plantillas ha probado ser independiente del dominio y permite la generación de consultas SPARQL formadas por múltiples tripletas.

Validación de la interfaz en diferentes dominios. El proceso de validación de la interfaz se llevó a cabo en dos dominios bien diferenciados, a saber, DBPedia y MusicBrainz, y cuyos resultados se publicaron en Paredes-Valverde et al. (2015) y Paredes-Valverde et al. (2016) respectivamente. Los experimentos realizados involucraron corpus de preguntas en lenguaje natural utilizados por la comunidad científica para evaluar interfaces de lenguaje natural orientadas a fuentes de datos semánticas, y un conjunto de preguntas en lenguaje natural elaboradas por usuarios potenciales ajenos al trabajo de investigación. Los resultados obtenidos en ambos dominios no varían significativamente uno del otro, lo cual se puede interpretar como un buen nivel de portabilidad por parte de la interfaz desarrollada.

5 Limitaciones

A pesar de que los resultados de evaluación obtenidos por la interfaz propuesta en esta tesis doctoral lucen alentadores, somos conscientes que este enfoque tiene ciertas limitaciones que, sin embargo, pueden ser direccionadas a futuro. Estas limitaciones se describen a continuación.

Tipos de pregunta soportadas. La interfaz de lenguaje natural permite el uso de preguntas factuales, es decir, aquellas que esperan como respuesta un hecho concreto. Por ejemplo, el nombre de una persona o lugar, la altura de una persona, entre otros. Además, la interfaz permite el uso de oraciones imperativas para solicitar información. En este sentido, es importante considerar más tipos de pregunta como de opción múltiple, verdadero/falso, entre otras. Para direccionar esta limitación, se planea el análisis de un corpus de preguntas de este tipo, que nos permita identificar las relaciones de dependencia que ayudarían a obtener una representación semántica de la pregunta.

Problemas de ambigüedad. La ambigüedad se refiere al fenómeno que se presenta cuando una palabra, un sintagma, o una oración puede ser interpretada de más de una forma. A pesar de que la interfaz propuesta implementa mecanismos para afrontar algunos casos de ambigüedad, somos conscientes de que estos presentan limitaciones que les impiden direccionar de mejor manera este problema. Para hacer frente a esta limitación, se plantea la integración de mecanismos de retroalimentación que permitan al usuario desambiguar las preguntas, ya sea a través de la reformulación de la pregunta o a través de seleccionar alguna opción de un conjunto provisto por la interfaz.

Nombre de individuos compuestos por múltiples palabras. La interfaz desarrollada identifica dentro de la pregunta aquellos elementos que hagan referencia a un individuo de la base de conocimiento, como lo puede ser

un libro o una canción. Sin embargo, en ocasiones el nombre del individuo está compuesto por múltiples palabras. Cuando este fenómeno ocurre, la interfaz puede reconocer este nombre, siempre y cuando combine el uso de mayúsculas y minúsculas. Este fenómeno representa un gran reto en el contexto de PLN, tal como se describe en Sag et al. (2002), donde de igual manera se presentan técnicas que podrían ser implementadas en esta interfaz, tal como el uso de reglas o métodos estadísticos.

6 Trabajo a futuro

A continuación, se describen temas que no han sido abordados por la interfaz propuesta y que proporcionan nuevas líneas de investigación.

Conjuntos de datos distribuidos y enlazados. Existe un gran número de individuos y organizaciones que han adoptado ya el enfoque de la web semántica, y en específico, el de Linked Data, para publicar sus datos. Debido a esto, es importante considerar esa distribución al momento de buscar una respuesta, pues en ocasiones esta puede depender de más de una base de conocimientos. En esta línea de investigación, se propone analizar el estado del arte sobre la consulta y recuperación de información de fuentes de información descentralizadas como Linked Data. Algunos de los enfoques más sobresalientes son las consultas federadas a SPARQL endpoints, fragmentos de patrones de tripletas y flujos de Linked Data. Tras el estudio podremos establecer un punto de partida para abordar el problema en cuestión e integrarlo en la interfaz de esta tesis.

Multilingüismo. Actualmente, la mayoría de las interfaces de lenguaje natural orientadas a bases de conocimiento basadas en ontologías no son capaces de responder a preguntas formuladas en múltiples lenguas. Esta tarea requiere que los recursos descritos en las ontologías (clases, propiedades e individuos) cuenten con una propiedad a través de la cual referenciarlos en cada una de las lenguas a considerar. Dicho esto, esta línea de investigación propone adaptar la interfaz propuesta en esta tesis a otra lengua, concretamente, al español. Esto demandará analizar herramientas para el análisis sintáctico de dependencias, y así reducir aún más la brecha existente entre usuarios y bases de conocimiento basadas en ontologías.

Agradecimientos

Mario Andrés Paredes Valverde es apoyado por la Comisión Nacional de Ciencia y Tecnología (CONACyT) y la Secretaria de Educación Pública (SEP).

Bibliografía

Berners-Lee, T., J. Hendler, O. Lassila. 2001. The Semantic Web. *Scientific American*. 284 (5): 28-37.

Cimiano, P., P. Haase, J. Heizmann, M. Mantel, y R. Studer. 2008. Towards portable natural language interfaces to knowledge bases – The case of the ORAKEL system. *Data & Knowledge Engineering*. 65 (2): 325-354.

Moldovan, D., S. Harabagiu, M. Pasca, R. Mihalcea, R. Girju, R. Goodrum, y V. Rus. 2000. The structure and performance of an open-domain question answering system. En *Proceedings of the 38th annual meeting on association for computational linguistics*, páginas 563-570.

Paredes-Valverde, M. A., M. Rodríguez-García, A. Ruiz-Martínez, R. Valencia-García, y G. Alor-Hernández. 2015. ONLI: An Ontology-Based System for Querying DBpedia Using Natural Language Paradigm. *Expert Systems with Applications.* 42 (12): 5163–76.

Paredes-Valverde, M. A., R. Valencia-García, M. Rodríguez-García, R. Colomo-Palacios, y G. Alor-Hernández. 2016. A Semantic-Based Approach for Querying Linked Data Using Natural Language. *Journal of Information Science*. 42 (6): 851–62.

Rodríguez-García, M., R. Valencia-García, F. García-Sánchez, y J. Samper-Zapater. 2014. Ontology-based annotation and retrieval of services in the Cloud. *Knowledge-based systems*. 56:15-25.

Sag, I., T. Baldwin, F. Bond, A. Copestake, y D. Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. En *Computational linguistic and intelligent text processing*, 1-15. Lecture notes in computer science. Springer Berlin Heidelberg.

Salas-Zárate, M., R. Valencia-García, A. Ruíz-Martínez, y R. Colomo-Palacios, 2016. Feature-based opinion mining in financial news: An ontology-driven approach. *Journal of Information Science*.

Procesamiento del Lenguaje Natural, Revista nº 60, marzo de 2018, pp. 79-82 recibido 21-11-2017 revisado 07-01-2018 aceptado 06-02-2018

Detección de Patrones Psicolingüísticos para el Análisis de Lenguaje Subjetivo en Español

Psycholinguistic Patterns Detection for Analyzing the Subjective Language in Spanish

María del Pilar Salaz Zárate
Universidad de Murcia
Facultad de Informática Campus Espinardo
Espinardo, 30100, Murcia, España
mariapilar.salas@um.es

Resumen: Tesis doctoral titulada "Detección de patrones psicolingüísticos para el análisis de lenguaje subjetivo en español", defendida por María del Pilar Salas Zárate en la Universidad de Murcia y elaborada bajo la dirección de los doctores Rafael Valencia García (Universidad de Murcia) y Miguel Ángel Rodríguez García (Universidad King Abdulah). La defensa tuvo lugar el 23 de mayo de 2017 ante el tribunal formado por los doctores Jesualdo Tomás Fernández Breis (Presidente, Universidad de Murcia), Alejandro Rodríguez González (Secretario, Universidad Politécnica de Madrid) y José Antonio Miñarro Giménez (Vocal, Medical University of Graz) y la tesis obtuvo la mención Cum Laude y Doctora Internacional.
Palabras clave: Patrones psicolingüísticos, lenguaje subjetivo, minería de opiniones

Abstract: Ph.D. thesis entitled "Psycholinguistic patterns detection for analyzing the subjective language in Spanish" written by María del Pilar Salas Zárate at the University of Murcia under the supervision of the Ph.D. Rafael Valencia García (University of Murcia) and Ph.D. Miguel Ángel Rodríguez García (University). The viva voice was held on the 23rd may 2017 and the members of the commission were the Ph.D. Jesualdo Tomás Fernández Breis (President, University of Murcia), Ph.D. Alejandro Rodríguez González (Secretary, Polytechnic University of Madrid) and Ph.D. José Antonio Miñarro Giménez (Vocal, University of Graz) and the thesis obtained the mention Cum Laude and International Doctor.
Keywords: Psycholinguistic patterns, subjective language, opinion mining

1 Introducción

Las opiniones son una parte importante en las decisiones del ser humano, cuando una persona desea tomar una decisión se basa en los comentarios de otras personas, por ejemplo, para comprar un producto, seleccionar un destino turístico, incluso para votar por un partido político. Con el surgimiento de la Web 2.0, ya no sólo se dependía de las opiniones de familiares o amigos, sino que se podía acceder a una gran cantidad de información en la Web provista por otros usuarios. Por lo que, actualmente, las personas visitan blogs, foros de discusión o redes sociales con el objetivo de obtener las experiencias de otros usuarios antes de tomar una decisión.

La lingüística es una de las áreas que se ha enfocado en el estudio de la opinión, o mejor dicho del lenguaje subjetivo. Este tipo de lenguaje se emplea para expresar estados personales en el contexto de una conversación o un texto (Wiebe, Wilson, Bruce, Bell, y Martin, 2004; Martínez-Cámara, 2016). Por otro lado, el análisis de sentimientos, también conocido como minería de opiniones, se ha convertido en un tema muy popular que tiene como objetivo el procesar opiniones públicas disponibles en la Web a través de técnicas de procesamiento de lenguaje natural. En este contexto, diferentes propuestas basadas en aprendizaje automático y orientación semántica han surgido en los últimos años. Estos trabajos
abordan problemas tales como análisis y construcción de lexicones de sentimientos,

ISSN 1135-5948. DOI 10.26342/2018-60-10 © 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

evaluación y clasificación de mensajes de Twitter, negación, por mencionar algunos. Otros trabajos se centran en analizar las opiniones en diferentes niveles, a saber, documento, sentencia y aspectos.

A pesar de los esfuerzos llevados a cabo en el análisis de sentimientos existen diversas características psicológicas y lingüísticas que no han sido abordadas. Por lo tanto, la clasificación automática de opiniones requiere un esfuerzo multidisciplinario, donde la lingüística y el procesamiento del lenguaje natural juegan un rol importante. Gracias a estas disciplinas es posible entender el lenguaje humano, clasificar las opiniones y resumir los sentimientos expresados acerca de un producto, servicio o cualquier otro aspecto.

El lenguaje figurado tal como la ironía, el sarcasmo y la sátira juega un papel muy importante en los sistemas de análisis de sentimientos. El doble sentido expresado en una opinión o comentario a través de este lenguaje puede invertir la polaridad de la opinión. Aunque, el lenguaje figurado ha sido ampliamente estudiado por diversas áreas como la lingüística, solo pocos estudios se han enfocado en la detección automática.

Por otro lado, es importante mencionar que pocos trabajos para el análisis de sentimientos, y lenguaje figurado, se han enfocado en el idioma español, quizá debido a la carencia de recursos lingüísticos en ese idioma. Sin embargo, el estudio del español es de lo más importante ya que es uno de los idiomas más utilizados en internet.

Las razones expuestas en los párrafos anteriores han sido la principal motivación para la realización de esta tesis doctoral. Por lo que se propone un método para la detección de patrones psicolingüísticos para el análisis de sentimientos y la detección de la sátira en español. Este método permite, a través de un enfoque automático supervisado, clasificar textos como positivo, negativo, neutro, muy positivo o muy negativo y como satíricos y no satíricos.

2 Objetivos

El objetivo principal de esta tesis doctoral es la detección de patrones psicolingüísticos para el análisis de lenguaje subjetivo en español. Específicamente, se propone el desarrollo de un método para el análisis de sentimientos y la detección de textos satíricos y no satíricos. Por lo tanto, los siguientes puntos fueron abordados en este trabajo.

Determinar qué tan relevantes son las características psicolingüísticas en la clasificación de sentimientos.

Determinar qué tan relevantes son las características psicolingüísticas en la clasificación de textos satíricos.

Identificar cuáles son las características más relevantes para el análisis de sentimientos.

Identificar cuáles son las características más relevantes para la detección de la sátira

3 Estructura de la tesis

La tesis doctoral se divide en cinco capítulos que exponen el estudio que se realizó. A continuación, se describe brevemente el contenido de cada uno de estos capítulos.

Capítulo 1. Este capítulo provee una breve introducción al trabajo de investigación, incluyendo la problemática a abordar.

Capítulo 2. Este apartado consiste en un detallado estudio de la bibliografía relacionada con las tecnologías base para el desarrollo del método propuesto. El estudio inicia con una introducción al lenguaje subjetivo. Posteriormente, se presenta el campo del procesamiento del lenguaje natural, así como los diferentes niveles de procesamiento. Después, se describe el campo del análisis de sentimientos y se proporcionan las definiciones más utilizadas por la comunidad investigadora. Además, se presenta su evolución histórica desde sus inicios en el siglo XX hasta la fecha. Asimismo, se presentan los diferentes niveles de análisis de opiniones, así como los dos principales enfoques en los cuales se basan la mayoría de los estudios, a saber, orientación semántica y aprendizaje automático. En el penúltimo apartado, se provee una introducción al lenguaje figurado, específicamente la ironía, el sarcasmo y la sátira. Finalmente, se presenta la importancia de las características psicolingüísticas en el lenguaje humano, y se introduce la herramienta LIWC, la cual permite obtener variables psicolingüísticas desde un texto escrito.

Capítulo 3. Este capítulo describe el método para el análisis de sentimientos y la detección de la sátira propuesto en este trabajo de investigación.

Capítulo 4. Esta sección se centra en la validación del método propuesto para el análisis de sentimientos y detección de la sátira. Este

capítulo se divide en dos apartados. En el primer apartado se presentan y discuten los resultados del análisis de sentimientos en dos dominios, a saber, películas y turismo. En el segundo apartado se presentan los resultados y discusión para la detección de la sátira, el cual fue validado en el dominio de noticias.

Capítulo 5. Finalmente, en este capítulo presentan las conclusiones obtenidas del trabajo de investigación y las posibles vías futuras.

4 Contribuciones

Las principales aportaciones de esta tesis se resumen a continuación.

Desarrollo de un método para la clasificación de sentimientos y detección de la sátira. Este método permite clasificar opiniones como positivas, negativas, neutras, muy positivas y muy negativas y tweets como satíricos y no satíricos. El método puede ser adaptado a diversos problemas de clasificación de textos e idiomas. Sin embargo, este requiere un corpus etiquetado como entrada.

Proceso para el preprocesamiento de tweets en español: La normalización de textos extraídos de redes sociales tal como Twitter suele ser más difícil debido a que los usuarios suelen abreviar palabras y usar jerga debido a la limitación de 140 caracteres que tienen los tweets. Actualmente, existen pocas herramientas del procesamiento del lenguaje natural que permiten normalizar estos textos en español. Para ello, nosotros definimos un proceso que permite normalizar los tweets para procesarlos posteriormente como un texto normal. El proceso consiste en tres principales pasos: 1) tokenización del texto y detección de entidades tales como URLs, menciones y etiquetas; 2) eliminar los elementos detectados en el paso 1 con excepción de etiquetas donde sólo es eliminado el "#"; y 3) extensión de abreviaturas y corrección de ortografía. Este proceso es de suma importancia, ya que actualmente Twitter está siendo un foco de investigación debido a la gran cantidad de información subjetiva contenida en estas redes sociales, la cual está constituida principalmente por opiniones.

Desarrollo de un corpus en el dominio del turismo. Los corpora son un recurso importante en el análisis de sentimientos. Por un lado, los métodos basados en un enfoque de aprendizaje automático requieren de corpus etiquetados con el objetivo de entrenar algoritmos de clasificación. Por otro lado, estos corpus sirven como base para la evaluación de sistemas de análisis de sentimientos. El desarrollo de un corpus requiere esfuerzo y tiempo debido a que el etiquetado se realiza manualmente con el objetivo de obtener un corpus de calidad. Sin embargo, hoy en día existen pocos corpus disponibles en español en la comunidad investigadora, es por ello, que el corpus obtenido en este trabajo de tesis supone una gran aportación.

Desarrollo de un corpus de tweets satíricos. Este corpus consiste en un conjunto de tweets etiquetados como satíricos y no satíricos extraídos desde diversas cuentas de Twitter. Actualmente existen algunos corpus del lenguaje figurado como ironía y sarcasmo. Sin embargo, hay una carencia de corpus con información satírica, sobre todo en español. Este corpus además de ser en este idioma está dividido en sátira mexicana y satírica española.

Detección de características psicolingüísticas para el análisis de sentimientos. Otra aportación relevante de esta tesis se centra en la identificación y extracción de características psicolingüísticas que son más discriminantes para el análisis de sentimientos y detección de la sátira.

5 Líneas futuras

Con respecto a investigación futura, se proveen varios aspectos que no han sido considerados como parte de esta tesis. Sin embargo, son considerados como líneas de investigación futuras a explorar. A continuación, se detalla cada uno de estos aspectos.

Integrar técnicas que permitan proveer un mejor soporte del proceso de normalización ante casos como tweets. La normalización de textos como tweets es una tarea muy difícil debido a que normalmente son textos con palabras abreviadas y con faltas de ortografía. En este trabajo de tesis, se propone un proceso para su normalización. Sin embargo, tiene una limitación en cuanto al procesamiento de etiquetas en inglés hashtags. Una etiqueta puede contener múltiples palabras juntas. Por lo que considerar técnicas tales como la presentada en (Bejcek, Stranák, y Pecina, 2013), permitirá abordar este problema.

Aplicación del método a diversos dominios. El método de análisis de sentimiento desarrollado en esta tesis ha sido favorablemente aplicado en los dominios

turístico y de películas. Por lo que la aplicación a otros dominios sería otra de las líneas de investigación a explorar como posible línea futura para tener en cuenta. Sin embargo, como se mencionó anteriormente, se requiere de un corpus del dominio etiquetado. Es por ello, que se propone el desarrollo de nuevos corpus en diversos dominios. Un área de especial interés es el dominio médico, el cual ha sido poco explorado. Sin embargo, las opiniones pueden ser de gran interés entre pacientes sobre todo cuando padecen de enfermedades que requieren de autogestión como la diabetes, asma, cáncer, hipertensión, etc. Por otro lado, en cuanto al lenguaje figurado, este trabajo está enfocado en la detección de la sátira, por lo que la creación de nuevos corpus para la ironía, sarcasmo y sátira en diversos dominios permitiría el desarrollo y evaluación de nuevos sistemas de análisis de sentimiento.

Aplicación del método en otros idiomas. La aplicación del método se ha enfocado en el idioma español, por lo que como trabajo a futuro se aplicará este método a otros idiomas tales como inglés, francés, árabe y a diversas variedades de español, como el que se habla en Argentina, Uruguay, Venezuela, etc. Esto permitirá determinar si los patrones psicolingüísticos detectados en esta tesis pueden contribuir también a la detección del análisis de sentimientos y sátira en diferentes idiomas y culturas.

Detección de patrones psicolingüísticos para el sarcasmo y la ironía. El procedimiento para detectar patrones psicolingüísticos únicamente ha sido diseñado para la sátira. Por lo que sería muy interesante también detectar patrones psicolingüísticos para el sarcasmo e ironía. Además, esto permitiría determinar el nivel de similitud entre estos tipos de lenguaje figurado, es decir, determinar qué categorías psicolingüísticas comparten.

Integración del sistema de detección del lenguaje figurado en el análisis de sentimientos. Los sistemas presentados en este trabajo de tesis doctoral son independientes. Por lo que la incorporación de un módulo que permita detectar la ironía, el sarcasmo y la sátira, así como otros tipos de lenguaje figurado como el humor en el sistema de análisis de sentimientos, permitirá no sólo detectar la polaridad de la opinión, sino también detectar si el texto es literal, irónico, sarcástico o satírico.

Contribuir al enriquecimiento de LIWC en español. La extracción de características depende en gran medida del diccionario de LIWC. Sin embargo, este diccionario carece de algunas palabras del español como verbos y de una gran variedad de palabras utilizadas en diversos países como Venezuela, Colombia, Ecuador, etc. El diccionario de LIWC puede ser enriquecido con otras palabras, lo cual se traduciría en una mejor extracción de características y, por tanto, una mejor precisión del sistema.

Agradecimientos

María del Pilar Salas Zárate es apoyada por la Comisión Nacional de Ciencia y Tecnología (CONACyT) y la Secretaria de Educación Pública (SEP).

Bibliografía

Bejcek, E., P. Stranák, y P. Pecina. 2013. Syntactic identification of occurrences of multiword expressions in text using a lexicon with dependency structures. En *Proceedings of the 9th Workshop on multiword expressions,* páginas 106-115.

Martínez-Cámara, E. 2015. Análisis de opinions en Español (Tesis de doctorado). Universidad de Jaén. Departamento de Informática. Obtenido de http://rua.ua.es/dspace/handle/10045/53569

Wiebe, J., T. Wilson, R. Bruce, M. Bell, y M. Martin. 2004. Learning Subjective Language. *Computational Linguistics.*, 30(3), 277-308.

Salas-Zárate, M. d. P., M. A. Paredes-Valverde, M. A. Rodriguez-Garcia, R. Valencia-García, y G. Alor-Hernández. 2017. Automatic detection of satire in Twitter: A psycholinguistic-based approach. *Knowledge-Based Systems*, 128, 20-33.

Salas-Zárate, M. d. P., R. Valencia-García, A. Ruiz-Martínez, y R. Colomo-Palacios. 2017. Feature-based opinion mining in financial news: an ontology-driven approach. *Journal of Information Science*, 43(4), 458-479.

Salas-Zarate, M. d. P., M. A. Paredes-Valverde, J. Limon, D. A. Tlapa, & Y. A. Báez. 2016. Sentiment Classification of Spanish Reviews: An Approach based on Feature Selection and Machine Learning Methods. *Journal of Universal Computer Science*, 22(5), 691-708.

Información General

SEPLN 2018

XXXIV CONGRESO INTERNACIONAL DE LA SOCIEDAD ESPAÑOLA PARA EL PROCESAMIENTO DEL LENGUAJE NATURAL

Hospital Universitario Virgen del Rocío – Sevilla (España)
19-21 de septiembre 2018
http://www.sepln.org/ y http://www.sepln2018.com/

1 Presentación

La XXXIV edición del Congreso Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) se celebrará los días 19, 20 y 21 de septiembre de 2018 en el Hospital Universitario Virgen del Rocío.

La ingente cantidad de información disponible en formato digital y en las distintas lenguas que hablamos hace imprescindible disponer de sistemas que permitan acceder a esa enorme biblioteca que es Internet de manera cada vez más estructurada.

En este mismo escenario, hay un interés renovado por la solución de los problemas de accesibilidad a la información y de mejora de explotación de la misma en entornos multilingües. Muchas de las bases formales para abordar adecuadamente estas necesidades han sido y siguen siendo establecidas en el marco del procesamiento del lenguaje natural y de sus múltiples vertientes: Extracción y recuperación de información, Sistemas de búsqueda de respuestas, Traducción automática, Análisis automático del contenido textual, Resumen automático, Generación textual y Reconocimiento y síntesis de voz.

2 Objetivos

El objetivo principal del congreso es ofrecer un foro para presentar las últimas investigaciones y desarrollos en el ámbito de trabajo del Procesamiento del Lenguaje Natural (PLN) tanto a la comunidad científica como a las empresas del sector. También se pretende mostrar las posibilidades reales de aplicación y conocer nuevos proyectos I+D en este campo.

Además, como en anteriores ediciones, se desea identificar las futuras directrices de la investigación básica y de las aplicaciones previstas por los profesionales, con el fin de contrastarlas con las necesidades reales del mercado. Finalmente, el congreso pretende ser un marco propicio para introducir a otras personas interesadas en esta área de conocimiento

3 Áreas Temáticas

Se anima a grupos e investigadores a enviar comunicaciones, resúmenes de proyectos o demostraciones en alguna de las áreas temáticas siguientes, entre otras:

- Modelos lingüísticos, matemáticos y psicolingüísticos del lenguaje
- Desarrollo de recursos y herramientas lingüísticas.
- Gramáticas y formalismos para el análisis morfológico y sintáctico.
- Semántica, pragmática y discurso.
- Resolución de la ambigüedad léxica.
- Generación textual monolingüe y multilingüe
- Traducción automática
- Síntesis del habla
- Sistemas de diálogo
- Indexado de audio
- Identificación idioma
- Extracción y recuperación de información monolingüe y multilingüe
- Sistemas de búsqueda de respuestas.
- Evaluación de sistemas de PLN.

ISSN: 1135-5948
© 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

- Análisis automático del contenido textual.
- Análisis de sentimientos y opiniones.
- Análisis de plagio.
- Minería de texto en blogosfera y redes sociales.
- Generación de Resúmenes.
- PLN para la generación de recursos educativos.
- PLN para lenguas con recursos limitados.
- Aplicaciones industriales del PLN.

4 Formato del Congreso

La duración prevista del congreso será de tres días, con sesiones dedicadas a la presentación de artículos, pósters, proyectos de investigación en marcha y demostraciones de aplicaciones. Además, prevemos la organización de talleres-workshops satélites para el día 18 de septiembre.

5 Comité ejecutivo SEPLN 2018

Presidente del Comité Organizador

- Carlos Luis Parra Calderón (Hospital Universitario Virgen del Rocío)

Colaboradores

- Mª Cabeza Gutiérrez Ruiz (Hospital Universitario Virgen del Rocío)
- Noa Patricia Cruz Díaz (Savana)
- José F. Quesada (Universidad de Sevilla)
- José Antonio Troyano (Universidad de Sevilla)
- Alicia Martínez García (FISEVI)
- Francisco Núñez Benjumea (FISEVI)

6 Consejo Asesor

Miembros:

- Manuel de Buenaga Rodríguez (Universidad Europea de Madrid, España)
- Sylviane Cardey-Greenfield (Centre de recherche en linguistique et traitement automatique des langues, Lucien Tesnière. Besançon, Francia)
- Irene Castellón Masalles (Universidad de Barcelona, España)
- Arantza Díaz de Ilarraza (Universidad del País Vasco, España)
- Antonio Ferrández Rodríguez (Universidad de Alicante, España)
- Alexander Gelbukh (Instituto Politécnico Nacional, México)
- Koldo Gojenola Galletebeitia (Universidad del País Vasco, España)
- Xavier Gómez Guinovart (Universidad de Vigo, España)
- José Miguel Goñi Menoyo (Universidad Politécnica de Madrid, España)
- Bernardo Magnini (Fondazione Bruno Kessler, Italia)
- Nuno J. Mamede (Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa, Portugal)
- M. Antònia Martí Antonín (Universidad de Barcelona, España)
- M. Teresa Martín Valdivia (Universidad de Jaén, España)
- Patricio Martínez Barco (Universidad de Alicante, España)
- Paloma Martínez Fernández (Universidad Carlos III, España)
- Raquel Martínez Unanue (Universidad Nacional de Educación a Distancia, España)
- Ruslan Mitkov (University of Wolverhampton, Reino Unido)
- Leonel Ruiz Miyares (Centro de Lingüística Aplicada de Santiago de Cuba, Cuba)
- Manuel Montes y Gómez (Instituto Nacional de Astrofísica, Óptica y Electrónica, México)
- Lidia Ana Moreno Boronat (Universidad Politécnica de Valencia, España)
- Lluis Padró Cirera (Universidad Politécnica de Cataluña, España)
- Manuel Palomar Sanz (Universidad de Alicante, España)
- Ferrán Pla (Universidad Politécnica de Valencia, España)
- Germán Rigau Claramunt (Universidad del País Vasco, España)
- Horacio Rodríguez Hontoria (Universidad Politécnica de Cataluña, España)
- Paolo Rosso (Universidad Politécnica de Valencia, España)
- Emilio Sanchís (Universidad Politécnica de Valencia, España)
- Kepa Sarasola Gabiola (Universidad del País Vasco, España)
- Encarna Segarra (Universidad Politécnica de Valencia, España)
- Thamar Solorio (University of Houston, Estados Unidos de América)

- Maite Taboada (Simon Fraser University, Canadá)
- Mariona Taulé (Universidad de Barcelona, España)
- Juan-Manuel Torres-Moreno (Laboratoire Informatique d'Avignon / Université d'Avignon, Francia)
- José Antonio Troyano Jiménez (Universidad de Sevilla, España)
- L. Alfonso Ureña López (Universidad de Jaén, España)
- Rafael Valencia García (Universidad de Murcia, España)
- M. Felisa Verdejo Maíllo (Universidad Nacional de Educación a Distancia, España)
- Manuel Vilares Ferro (Universidad de la Coruña, España)
- Luis Villaseñor-Pineda (Instituto Nacional de Astrofísica, Óptica y Electrónica, México)

7 Fechas importantes

Fechas para la presentación y aceptación de comunicaciones:

- Fecha límite para la entrega de comunicaciones: 21 de marzo de 2018.
- Notificación de aceptación: 2 de mayo de 2018.
- Fecha límite para entrega de la versión definitiva: 17 de mayo de 2018.
- Fecha límite para propuesta de talleres y tutoriales: 24 de marzo de 2018.

Información para los Autores

Formato de los Trabajos

- La longitud máxima admitida para las contribuciones será de 8 páginas DIN A4 (210 x 297 mm.), incluidas referencias y figuras.
- Los artículos pueden estar escritos en inglés o español. El título, resumen y palabras clave deben escribirse en ambas lenguas.
- El formato será en Word ó LaTeX

Envío de los Trabajos

- El envío de los trabajos se realizará electrónicamente a través de la página web de la Sociedad Española para el Procesamiento del Lenguaje Natural (http://www.sepln.org)
- Para los trabajos con formato LaTeX se mandará el archivo PDF junto a todos los fuentes necesarios para compilación LaTex
- Para los trabajos con formato Word se mandará el archivo PDF junto al DOC o RTF
- Para más información http://www.sepln.org/home-2/revista/instrucciones-autor/

Información Adicional

Funciones del Consejo de Redacción

Las funciones del Consejo de Redacción o Editorial de la revista SEPLN son las siguientes:

- Controlar la selección y tomar las decisiones en la publicación de los contenidos que han de conformar cada número de la revista
- Política editorial
- Preparación de cada número
- Relación con los evaluadores y autores
- Relación con el comité científico

El consejo de redacción está formado por los siguientes miembros

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

Funciones del Consejo Asesor

Las funciones del Consejo Asesor o Científico de la revista SEPLN son las siguientes:

- Marcar, orientar y redireccionar la política científica de la revista y las líneas de investigación a potenciar
- Representación
- Impulso a la difusión internacional
- Capacidad de atracción de autores
- Evaluación
- Composición
- Prestigio
- Alta especialización
- Internacionalidad

El Consejo Asesor está formado por los siguientes miembros:

Manuel de Buenaga	Universidad Europea de Madrid (España)
Sylviane Cardey-Greenfield	Centre de recherche en linguistique et traitement automatique des langues (Francia)
Irene Castellón	Universidad de Barcelona (España)
Arantza Díaz de Ilarraza	Universidad del País Vasco (España)
Antonio Ferrández	Universidad de Alicante (España)
Alexander Gelbukh	Instituto Politécnico Nacional (México)
Koldo Gojenola	Universidad del País Vasco (España)
Xavier Gómez Guinovart	Universidad de Vigo (España)
José Miguel Goñi	Universidad Politécnica de Madrid (España)
Bernardo Magnini	Fondazione Bruno Kessler (Italia)

© 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

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

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.
Si desea inscribirse como socio de la Sociedad Española del Procesamiento del Lenguaje Natural puede realizarlo a través del formulario web que se encuentra en esta dirección http://www.sepln.org/socios/inscripcion-para-socios/
Los números anteriores de la revista se encuentran disponibles en la revista electrónica: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/issue/archive
Las funciones del Consejo de Redacción están disponibles en Internet a través de http://www.sepln.org/category/revista/consejo_redaccion/
Las funciones del Consejo Asesor están disponibles Internet a través de la página http://www.sepln.org/home-2/revista/consejo-asesor/
La inscripción como nuevo socio de la SEPLN se puede realizar a través de la página http://www.sepln.org/socios/inscripcion-para-socios/

© 2018 Sociedad Española para el Procesamiento del Lenguaje Natural