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**

**Extracción y recuperación de Información**



**Análisis de textos médicos**



**Análisis del habla**



**Aprendizaje automático en PLN**

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



**ISSN**: 1135-5948
**ISSN electrónico**: 1989-7553
**Depósito Legal**: B:3941-91
**Editado en:** Universidad de Jaén
**Año de edición:** 2018
**Editores:** Carlos Luis Parra-Calderón Hospital Universitario Virgen del Rocío
carlos.parra.sspa@juntadeandalucia.es
Alicia Martínez García Hospital Universitario Virgen del Rocío
alicia.martinez.exts@juntadeandalucia.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

| | |
|---|---|
| Paloma Martínez Fernández | Universidad Carlos III (España) |
| Raquel Martínez Unanue | Universidad Nacional de Educación a Distancia (España) |
| Ruslan Mitkov | University of Wolverhampton (Reino Unido) |
| Manuel Montes y Gómez | Instituto Nacional de Astrofísica, Óptica y Electrónica (México) |
| 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**

| | |
|---|---|
| Lara Gil Vallejo | Universitat Oberta de Cataluña (España) |
| Fernando Martínez Santiago | Universidad de Jaén (España) |
| Alessandra Cignarella | Università di Torino (Italia) |
| Soto Montalvo | Universidad Nacional de Educación a Distancia (España) |
| Mariluz Morales Botello | Universidad Europea de Madrid (España) |
| Mikel Larrañaga Olagaray | Universidad de País Vasco (España) |
| Rafael Muñoz Gil | Universidad Europea de Madrid (España) |
| Borja Navarro Colorado | Universidad de Alicante (España) |
| Bilal Ghanem | Universidad Politécnica de Valencia (España) |

Sociedad Española para el Procesamiento del Lenguaje Natural

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

ISSN: 1135-5948

## Preámbulo

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

Las áreas temáticas tratadas son las siguientes:

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

El ejemplar número 61 de la revista *Procesamiento del Lenguaje Natural* contiene trabajos correspondientes a tres apartados diferentes: comunicaciones científicas, resúmenes de proyectos de investigación y descripciones de aplicaciones informáticas (demostraciones).

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

Se recibieron 44 trabajos para este número, de los cuales 26 eran artículos científicos y 18 correspondían a resúmenes de proyectos de investigación y descripciones de aplicaciones informáticas. De entre los 26 artículos recibidos, 13 han sido finalmente seleccionados para su publicación, lo cual fija una tasa de aceptación del 50%.

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

Septiembre de 2018
Los editores

Sociedad Española para el Procesamiento del Lenguaje Natural

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

ISSN: 1135-5948

## Preamble

The *Natural Language Processing* journal aims to be a forum for the publication of 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 61th issue of the *Procesamiento del Lenguaje Natural* journal contains scientific papers, research project summaries and description of Natural Language Processing software tools. 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.

Fourty-four papers were submitted for this issue, from which twenty-six were scientific papers and eighteen were either projects or tool description summaries. From these twenty-six papers, we selected thirteen (50%) 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.

September 2018
Editorial board

Sociedad Española para el Procesamiento del Lenguaje Natural

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

ISSN: 1135-5948

**Artículos**

**Extracción y recuperación de Información**



**Análisis de textos médicos**



**Análisis del habla**



**Aprendizaje automático en PLN**



**Proyectos**

# *Artículos*

# *Extracción y recuperación de información*

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 15-22 recibido 20-03-2018 revisado 01-05-2018 aceptado 03-05-2018

# Text-to-Pictogram Summarization for Augmentative and Alternative Communication

## *Resúmenes Texto-a-Pictograma para Comunicación Aumentativa y Alternativa*

**Laura Cabello, Eduardo Lleida, Javier Simón, Antonio Miguel, Alfonso Ortega**
ViVoLab-Universidad de Zaragoza C/ María de Luna 1. 50018 Zaragoza
laura92cp2@gmail.com, (lleida,jasimon,amiguel,ortega)@unizar.es, http://vivolab.unizar.es

**Abstract:** Many people suffer from language disorders that affect their communicative capabilities. Augmentative and alternative communication devices assist learning process through graphical representation of common words. In this article, we present a complete text-to-pictogram system able to simplify complex texts and ease its comprehension with pictograms.
**Keywords:** pictogram, summarization, embeddings, augmentative and alternative communication, AAC, natural language processing, NLP

**Resumen:** Numerosas personas padecen trastornos del habla que merman su capacidad de comunicación. Su proceso de aprendizaje se apoya en el uso de dispositivos para la comunicación aumentativa y alternativa con símbolos gráficos. En este artículo presentamos un sistema texto-a-pictograma completo, capaz de simplificar textos complejos y facilitar su comprensión con pictogramas.
**Palabras clave:** pictograma, resumen, comunicación aumentativa y alternativa, CAA, procesado del lenguaje, PLN

## *1 Introduction*

Texts are often illustrated because images help people to comprehend and remember the content. Moreover, graphic resources facilitate communication for individuals with severe language and speech disorders. Augmentative and alternative communication (AAC) encompasses all sort of communication methods that supplement and ease, or even replace, spoken and/or written language.

We present a novel text-to-pictogram system that aims to support AAC by conveying meaningful information to Spanish-speaking people with language impairments. Our system relies on pictograms or simple images to build sentences from a summarized text that captures the core meaning of an input document. Its main application is to serve as a reading aid to help understanding the theme from the input source and enhance learning process. Use case examples range from tales summaries for children to newspaper summaries for adults.

Since generating graphical replacement for different text genres is not an easy task, firstly we propose to summarize the input document to discard redundant information while retaining the core message. We are mainly concerned with what summary content should be regardless the form, so we define an extractive summarization method to extract salient sentences from the input source. Then, we retrieve a sequence of representative images for each word or constituent, i.e. group of words that function as a single unit within a hierarchical structure. This stage employs syntax information from part-of-speech (POS, from henceforth) labels, semantic features encoded in word embeddings, and context information from topic modeling. In order to improve understanding, we omit stopwords and further process complex text structures.

It is important to note that this work is not a study of summarization methods nor language modeling and feature learning techniques, but a complete text-to-pictogram system. The specific implementations presented here were chosen from literature after testing several options. Individually, they are simple but effective approaches to our ulti-

ISSN 1135-5948. DOI 10.26342/2018-61-1

mate objective of depicting a given article as a sequence of images.

## *2 Related Work*

Previous research has been done in the field of text-to-pictogram systems. The current work differs from others in its nature, aiming to be an assistive tool to help people with language disorders understand complex texts. Also following same approach to extract pictures in a word by word basis, García-Cumbreras et al. (2016) presented an AAC system meant to be used in a controlled environment; Mihalcea and Leong (2006) studied a translation system through pictures; Zhu et al. (2007) rendered a nonlinear layout based on web search; UzZaman, Bigham, and Allen (2011) focused on multimodal summarization through images; Martínez-Santiago et al. (2015) proposes an upper ontology with linguistic knowledge used to model the usual language of beginning communicators. Another line of work is meant to represent a the gist of the text, known as text-to-scene systems. WordsEye (Coyne and Sproat, 2001) is one of the best known systems, able to synthesize realistic 3D scenes for descriptive sentences. **AraWord**[1] and **DictaPicto**[2] are examples for simple word by word basis translation to pictograms.

## *3 Method*

Our text-to-pictogram system can be divided into three general phases, detailed in subsequent parts and depicted in Figure 2. Data preprocessing is a critical step in any Natural Language Processing (NLP) application. Thereafter, we use a sentence-ranking mechanism to form an extractive summary from an input document. Then, for each selected keyphrase, we exploit several NLP techniques -such as word embeddings and topic models-, deployed earlier at preprocessing stage to convert text to images. An output of our system can be seen in Figure 1.

### 3.1 Databases

Database **NEWS1709** was used to train word embeddings and LDA model. A custom web crawler searched through 32 different Spanish newspapers and magazines throughout 2017. We collected 850633 articles from January to September, covering a wide range of themes including but not limited to economy, politics, sports, forecast, culture and opinion. An initial preprocessing was required in order to get rid of advertisements, duplicate articles and non Spanish texts.

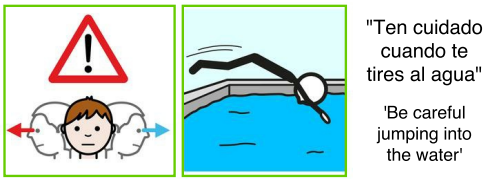


Figure 1: Example of a depicted sentence. Our system has detected the reflexive verb *'tirarse'* and output a single picture for the action *'tirarse al agua'* (*'jumping into water'*)

With regard to pictograms, **ARASAAC**[3] (Aragonese Portal of Augmentative and Alternative Communication) provides a package of 9564 color pictograms labeled in Spanish. Each pictogram is compose of three attributes: name, a set of labels or tags semantically related and the image itself. However, since some images can illustrate more than one concept, we are able to represent over 13000 different terms. This database is our main resource of images.

### 3.2 Preprocessing

Before directly tackling the problem of text-to-pictogram conversion, a preprocessing of our training data is required in order to define a voca-bulary enhanced with POS tags. We use the NEWS1709 database which is composed of raw text from newspaper articles, as explained in previous section. In addition, a set of word embeddings and a topic model are also created before being used in the process of selecting a suitable pictogram.

NLP tasks involve vast, not fixed vocabulary, since common expressions evolve and differ over time. Aiming to delimit our, we made use of a *lemmatizer* that removed inflectional endings to return the dictionary form of words, known as *lemma*. Concurrently, we gathered word's POS information such as category and type for those terms that appeared more than 100 times within the entire corpus. This way we filtered irrelevant terms and misspelling errors out and

[1] http://aulaabierta.arasaac.org/araword_inicio

[2] http://www.fundacionorange.es/aplicaciones/dictapicto-tea/

[3] www.arasaac.org

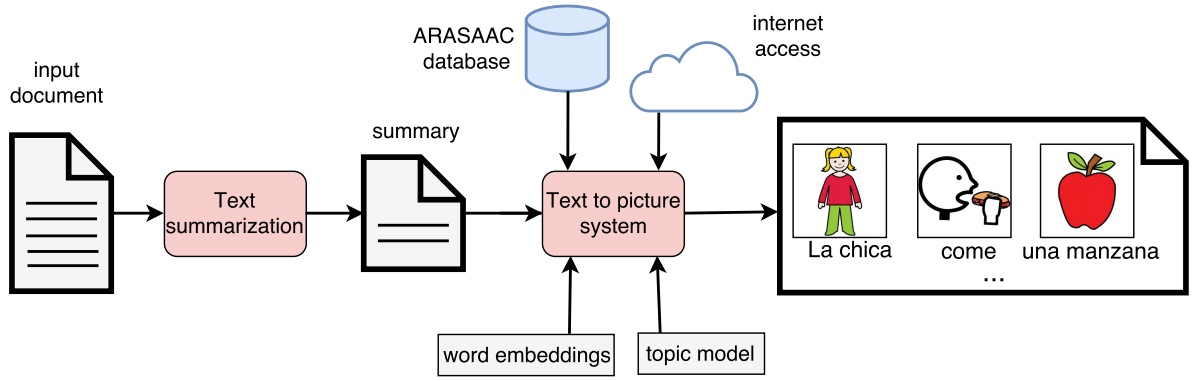


Figure 2: Block diagram depicting an overview of the end-to-end system

created a dictionary with 51358 terms. We stored 32 different labels after collecting all categories and types provided[4]. Both lemmatized content and word POS tags were achieved using the open-source suite FreeLing.

### Word embeddings

Embeddings are a mathematic representation of words that captures a high degree of syntactic and semantic information. A word embedding $\vec{w}$ is a $D$-dimensional distributed representation of word $w$, from a vocabulary $V$, in a real valued vector space so as to $V \rightarrow \mathbb{R}^D : w \mapsto \vec{w}$.

Mikolov et al. (2013a) introduced word2vec, one of the first unsupervised embedding methods built upon a probabilistic prediction model, the continuous bag-of-words or CBOW. As shown in Section 4, we evaluated different approaches based on this model and observed that methods involving morphological information from POS tagging outperformed other implementations stu-died. Best results were provided by *wordtag* method, whose embeddings were trained over tagged text. For example, the sentence "el perro come" ("the dog eats") would be feed as "el/DA perro/NC come/VM", where DA: Determiner Article, NC: Noun Common, VM: Verb Modal.

Word embeddings are used in our end-to-end system in two scenarios. If the target word or constituent is not directly linked to a pictogram, we attempt to find a synonym from our vocabulary of pictograms defined by ARASAAC. We compute nearest neighbors to the target expression through the commonly used cosine similarity measure, defined as $sim(w_1, w_2) = \frac{\vec{w_1}\vec{w_2}}{||\vec{w_1}||||\vec{w_2}||}$. Another usage scenario applies if we're dealing with *polysemy*, this is, target expression potentially match more than one pictogram. In this case, word embeddings combined with a topic model are used to produce a sentence embedding, therefore allowing to retrieve the most suitable image as detailed in Section 3.4.

### Topic modeling

Topic modeling provides methods for automatically organizing, searching or even understanding a large amount of data archives. Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan, 2003) is a popular method of topic modeling. It defines a generative probabilistic model of a corpus where each document is considered to exhibit multiple themes (or topics). Uncovering the hidden thematic structure from a considerable document collection allows us to generalize and subsequently classify unseen texts into predefined topics.

Despite topic modeling is inherent to extract representative content from documents, results strongly depend on the value given to hyperparameters $\alpha$ and $\eta$ in Dirichlet distribution. Since we are working with a large, heterogeneous corpus, we tune $\alpha$ close to zero resembling a *mixture model* where only one topic is assigned per document: newspaper articles do not exhibit a mixture of many topics, but are more specific. $\alpha = 0.05$ draws a sparse probability density function ($\Theta_d$), which means only few topics -usually one to three- will have positive probability within each document. Thus, expected distribution is not centered in the topic simplex. On the

[4]See manual of FreeLing for further information: https://talp-upc.gitbooks.io/freeling-user-manual/content/tagsets/tagset-es.html

other hand, $\eta$ alters words distribution $\beta_{d,k}$. We set a high value ($\eta = 1.0$), so each topic is likely to contain a mixture of most of the words and not only a specific set of few words. This means all weights $w_d \in \beta_{d,k}$ are drawn from a probability distribution, which is important when solving the problem of *polysemy*.

Our proposal is to use LDA distribution of words over topics assigned to a document to weight words in a sentence. Therefore, most relevant words are given more importance and polysemic words are better disambiguated by context gathered in each topic.

Training of LDA model was performed over NEWS1709 dataset. Test and hold-out sets were created following same approach as in NEWS1709 but with press from October and November respectively. We monitored perplexity over test and hold-out corpora to ensure convergence and generalization of topics. In total, training represents 80% of the data, test and hold-out 10% each. Selected LDA model distinguishes $K = 50$ topics and it was trained with 10 passes over the entire corpus.

### 3.3 Text Summarization

Document summarization has been an active research area of NLP since the late 1950s (Luhn, 1958). Different approaches to make well formed summaries (Chang and Chien, 2009; Gong and Liu, 2001; Bian, Jiang, and Chen, 2014; Ozsoy, Alpaslan, and Cicekli, 2011) seek concise texts that convey important information in the original document(s). Most of them are extractive methods based on Latent Semantic Analysis (LSA) or LDA algorithms. We opted for a solution based on LSA over LDA due to the lack of necessity for previous training to tweak algorithm hyperparameters, which has a major impact in performance and often depends on external information such as the corpora used. Furthermore, it leads to a faster overall runtime (an average of 120s *vs* 2s in our tests).

LSA is an algebraic method applied to text summarization by Gong and Liu (2001). They applied singular value decomposition (SVD) to generic text summarization, and designed an unsupervised approach which does not need any previous training or outer information. After that, different LSA approaches have been proposed (Gong and Liu, 2001; Steinberger and Jezek, 2004; Murray, Renals, and Carletta, 2005; Ozsoy, Alpaslan, and Cicekli, 2011) which usually contain three common steps (Ozsoy, Alpaslan, and Cicekli, 2011):

i. *Define an input matrix.* Text must be mathematically readable. The process starts with creation of a term-sentence matrix $\mathbf{A} = [A_1, A_2, ..., A_n]$ whose columns define the importance of every word in each sentence. Cells can be filled in following different approaches, such as computing the word frequency in a sentence, the log-entropy value, or, as we did, the Tf-Idf (Term Frequency-Inverse Document Frecuency) if we count on a corpus with several documents. If there are a total of $m$ terms and $n$ sentences in the document, then we will have an $m$x$n$ matrix $\mathbf{A}$ for the document, where without loss of generality, $m \geq n$. Since all words are not seen in all sentences, the matrix is usually sparse.

ii. *Apply SVD.* SVD is an algebraic method that decomposed the given matrix $\mathbf{A}$ into three new matrices, defined as follows:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V^T} \quad (1)$$

where $\mathbf{U}$ is an $m$x$k$ column-orthonormal matrix containing $k$ underlying concepts ($m > k$), $\mathbf{\Sigma}$ is a $k$x$k$ diagonal matrix whose elements are non-negative singular values sorted in descending order, and $\mathbf{V^T}$ is a $k$x$k$ orthonormal matrix whose columns are sentence singular vectors.

The interpretation of applying SVD to a term-sentence matrix is two-folded. From transformation point of view, it leads to a dimensionality reduction from an *m-dimensional* to a *k-dimensional* vector space. From semantic point of view, the SVD derives the latent semantic structure from the document represented by matrix $\mathbf{A}$ (Steinberger and Jezek, 2004).

iii. *Select salient sentences.* Different set of sentences are drawn depending on how we use the results from SVD. Several approaches solely use information within $\mathbf{V^T}$ matrix (Gong and Liu, 2001; Ozsoy, Cicekli, and Alpaslan, 2010), others also make use of $\mathbf{\Sigma}$ to emphasize most important concepts (Steinberger and Jezek, 2004; Murray, Renals, and Carletta, 2005). We implemented Topic method and Cross method proposed in (Ozsoy, Cicekli, and Alpaslan, 2010), and

chose the latter after comparing their performance. Cross method uses $\mathbf{V^T}$ matrix for sentence selection purposes.

### 3.4 Selecting Images

Once we have detected the main concepts from the input document, next stage is to find a sequence of pictograms or images to represent them. We retrieve a list of pictogram candidates from ARASAAC database and then call function in Algorithm 1 to get the most suitable image. If a word is out of ARASAAC dictionary of pictograms, we try to find a synonym computing cosine distance among word embeddings. Only in the event that a proper noun is not represented by ARASAAC pictograms, we do image search through a custom search RESTful API and get a public domain picture. Verbal periphrasis and compound verbal forms are depicted as a whole, making an emphasis in the main verb. Finally, stopwords are left without visual representation.

Our proposal to solve polysemy combines morphological information from POS tags, semantic information encoded in word embeddings and weights assigned to each word related to its importance within the $K_d$ latent topics in the document, i.e., a hint of how relevant is *that* word in *that* document. Therefore, keywords are highlighted and words linked to more than one pictogram -typically polysemic words- are better distinguished. Let us consider the following example to endorse this statement. Note that it is an adaptation in English to ease reader's understanding. Suppose we are depicting the sentence 'Every year *cranes* return to the *wetlands*' from an article talking about wildlife, where keywords are in *italics*. Making use of our trained LDA model, two topics are assigned to it: *topic 1* related to nature and *topic 5* related to climate change, because the article also reads about it. Then, suppose we are searching for the best pictogram to represent the word *crane* (lemma from *cranes*) and we have reached *line 14* in *Algorithm* 1, which means there are potentially multiple suitable images.

Now, our approach to selecting the best image combines word embeddings and LDA topics. It consists of the following steps. First, we compute the sentence embedding $\vec{s}$ as a weighted sum of its $L$ word embeddings. Word weights $\vec{z_s}$ take into account the topic-distribution in the document ($\Theta_{d,k}$ for topic $k$ in document $d$) and term-distribution for each topic ($\beta_{d,k}$). Mathematically,

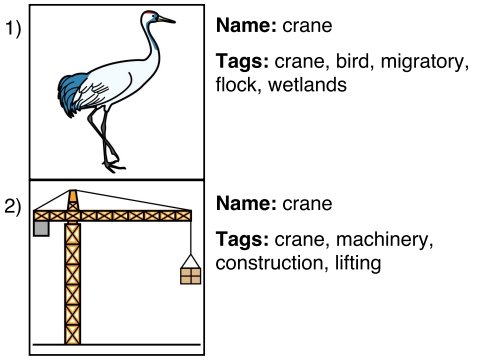


Figure 3: Retrieved pictos for the word *crane* from *ARASAAC* database

$$\vec{s} = \frac{\vec{z_s}}{||\vec{z_s}||}\mathbf{W}, \tag{2}$$

with

$$\vec{z_s} = \sum_{n=0}^{L-1} \sum_{k \in K_d} \Theta_{d,k}\beta_{d,k}z_{d,n}, \tag{3}$$

where $z_{d,n}$ is a binary variable that weights stopwords and words out of vocabulary by zero; $\mathbf{W}$ is an $L$x$D$ matrix whose rows are word embeddings. The inner sum in Equation 3 looks at the topics from document $d$, i.e., topic 1 and 5 in the example above.

Next, we perform alike with tags attached to every image. Following our example, the word *cranes* have two different pictograms shown in Figure 3. Thus, we apply Equation 2 over topics from the original article to each set of tags and create $\vec{t_1}$ and $\vec{t_2}$. Word tags from picture 1 are closely related to words from original sentence and they are likely to have a higher weight in topics 1 and 5 than those from picture 2. This leads to a final embedding $\vec{t_1}$ that encodes a semantic akin to that in $\vec{s}$. Finally, we employ cosine similarity for similarity computations in the embedding space. Results in this particular example show that $sim(s, \vec{t_1}) > sim(s, \vec{t_2})$, so the right image is selected.

## 4 Evaluation

Successive evaluation metrics refer to different steps in our system pipeline. Before selecting the final implementation, we compared word embeddings with respect to specific queries and tested the quality of summaries given by different LSA approaches.

**Algorithm 1** Select Picto

```
1: function SELECTPICTO(x, osent, lsent)
   Input x: list of picto candidates to represent the target word, osent: original sentence lsent: lemmatized sentence
   Output: selected picto
2:     if x.length = 1 then return x.picto
3:     context ← osent[w-W:w+W]
4:     initialize counter[] to zero
5:     for picto in x do
6:         if picto.name is compound and picto.name is in osent then
7:             return picto
8:         else
9:             if any word in context in picto.tags then
10:                counter[i] ← counter[i] + 1
11:    if max(counter) > 0 then
12:        return x[argmax(counter)]
13:    else
14:        picto ← word2vec.eval_context_with_LDA(x.tags, osent, lsent)
15:        return picto
```

Finally, we performed an overall test to objectively evaluate our text-to-pictogram system based on resources for AAC available in ARASAAC site.

## 4.1 Embeddings Evaluation

Table 1 displays a comparative among approaches proposed to enhance baseline word2vec performance with POS features. While in *wordtag* the POS is directly appended to training text as shown in Section 3.2, in *vectortag* it is encoded as one-hot vector and then concatenated to baseline word2vec embeddings. All embeddings mapped words into a 200-dimensional space, except *vectortag* that results in 200+32 dimensions, trained over 10 iterations with symmetric window of 4 samples and implemented negative sampling (Mikolov et al., 2013b) with 13 negative samples. Following previous work (Schnabel, Mimno, and Joachims, 2015; Mikolov, Yih, and Zweig, 2013), we conducted experiments on word analogy, relatedness and coherence tasks.

Word analogy assesses the capability of word embeddings to deduce semantic relationships. This task satisfies the statement *"if a:b, then x:y"* where $y$ is unknown. Most suitable word $y$ is found using word embeddings in the following function proposed by Mikolov, Yih, and Zweig (2013):

$$y^* = argmax_y sim(y, b) - sim(y, a) + sim(y, x) \quad (4)$$

We evaluate this metric on an Spanish version of Google analogy questions set proposed by Mikolov et al. (2013a).

Word relatedness and Coherence query inventories were created following work in (Schnabel, Mimno, and Joachims, 2015). We gathered 100 query words that balance frequency, POS (adjectives, adverbs, nouns and verbs) and concreteness. With regard to intrinsic evaluation of *relatedness*, the four nearest neighbors were retrieved for each of the 100 query words; 4 volunteers were then requested to pick the term that is most similar to the target word according to their perception. To evaluate *coherence*, we assess whether small, local clusters of words in the embedding space are mutually related. Same voluntary users were presented four words the day after, three of which are close neighbors (each query word from our 100 query words set and its two nearest neighbors) and one of which is an "intruder". Intruder word was selected to normalize frequency-based effects as in cited article.

| | Analogy | Relatedness | Coherence |
|---|---|---|---|
| word2vec (baseline) | **49.4** | 47.5 (67.0) | 92.7 |
| wordtag | 49.3 | **53.0 (75.5)** | **93.5** |
| vectortag | 44.8 | 50.4 (70.7) | 91.5 |

Table 1: Average accuracy scores (%) for each embedding method. Numbers between brackets show accuracy if acknowledging 2nd-nearest neighbor as valid. Best results for each metric are highlighted in bold

## 4.2 Summarization Evaluation

ROUGE evaluation tool (Lin, 2004) was adopted to measure summarization performance over a set of one hundred Spanish news extracted from NEWS1709 database. Articles were selected to have between 200 and 900 words and randomly belong to either one of the following categories: economy, international, national, society or sports. Golden standard summaries were provided by a linguistic expert. Table 2 contains different ROUGE evaluations that prove Cross

slightly better than Topic method, as presented in the original article. LSA algorithms were tuned so each summary covers up to half of the original length.

| | LSA-Cross | LSA-Topic |
|---|---|---|
| ROUGE-1 R | 0.7242 | 0.6913 |
| ROUGE-1 P | 0.5605 | 0.5611 |
| ROUGE-1 F | 0.6082 | 0.5840 |
| ROUGE-L R | 0.7015 | 0.6715 |
| ROUGE-L P | 0.5441 | 0.5450 |
| ROUGE-L F | 0.5900 | 0.5670 |
| ROUGE-S R | 0.5371 | 0.5138 |
| ROUGE-S P | 0.3453 | 0.3478 |
| ROUGE-S F | 0.3711 | 0.3470 |

Table 2: Results show a comparison of average recall (R), precision (P) and F-measure (F) for two different algorithms and ROUGE measures

ROUGE-1 measures unigram overlap between reference and automatic summaries, ROUGE-L measures the longest common subsequence of words and ROUGE-S scores overlap of word pairs (skip bigram) that can have an unlimited set of gaps in between words (Lin, 2004).

### 4.3 Text-to-pictogram Evaluation

It is not easy to devise an objective measure to quantify a text-to-pictogram system performance. In order to manage so, we counted on $N = 53$ sentences with 245 different lemmas and an average length of 10.5 words per sentence. Sentences have been extracted from two sources: an adapted document from the 2010 Convention on the Rights of Persons with Disabilities[5] and children's tales written by Douglas Wright[6], which were manually illustrated by ARASAAC. We adopted a binary evaluation and scored if a pictogram predicted by our system matched the one assigned by the illustrator. Since ARASAAC database includes different pictograms to depict exactly the same concept, we also counted them in as valid results. We achieved an averaged accuracy of 74% computed as $\frac{1}{N}\sum_i \frac{\#correct\ pictograms\ in\ sentence\ i}{\#total\ available\ pictograms\ in\ sentence\ i}$ where the denominator excludes some pictograms used but not included in ARASAAC database (public domain pictures, see Section 3.4).

## 5 Conclusion and Future Work

In this paper, we presented an assistive text-to-pictogram system to help people with language disorders to understand complex texts. The system integrates various phases combining automatic text summarization, natural language processing and a depicting algorithm that translates input text into images. We approached the issue of automatic summarization as a simple sentence ranking mechanism and then processed the summary with pre-trained embeddings and a topic model. Polysemy supposes a challenge in many aspects of NLP. We proposed to combine syntactic and morphological information from the local context of a polysemous word. Despite the fact that automatic illustration is an inherently subjective task, we conducted an overall test on children's book illustration and evaluated the different modules merged in our system. We realized the importance of a well defined evaluation method and corpus to foster the research in the area.

Our future compromise is to work on this line to provide a reference corpus for text-to-pictogram translation and summarization. Also we will concentrate on improving reading comprehension for natural language represented by pictograms alone by working jointly the ARASAAC professionals and experts on AAC systems at the Alborada [7] special education school in Zaragoza (Spain).

## Acknowledgments

This work has been supported by the Spanish Government through project TIN2017-85854-C4-1-R and by the European Union's FP7 Marie Curie action, IAPP under grant agreement no.610986. We are grateful to ARASAAC for their determination to provide graphical resources for AAC. We are also thankful to our voluntary evaluators.

## References

Bian, J., Z. Jiang, and Q. Chen. 2014. Research on multi-document summarization based on lda topic model. In *2014*

[5] http://www.ceapat.es/InterPresent2/groups/imserso/documents/binario/convencion_accesible2.pdf

[6] http://www.arasaac.org/materiales.php?id_material=578

[7] https://cpeealborada.wordpress.com/

*Sixth International Conference on Intelligent Human-Machine Systems and Cybernetics*, volume 2, pages 113–116, August.

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

Chang, Y.-L. and J. T. Chien. 2009. Latent dirichlet learning for document summarization. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1689–1692, April.

Coyne, B. and R. Sproat. 2001. Toward communicating simple sentences using pictorial representations. In *Proceedings of 28th Conference on Computer Graphics and Interactive Techniques*, pages 487–496.

García-Cumbreras, M., F. Martínez-Santiago, A. Montejo-Ráez, M. Díaz-Galiano, and M. Vega. 2016. Pictogrammar, comunicación basada en pictogramas con conocimiento lingüístico. *Procesamiento del Lenguaje Natural*, 57:185–188.

Gong, Y. and X. Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25.

Lin, C. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out: Proceedings of ACL-04 Workshop*, pages 74–81.

Luhn, H. 1958. The automatic creation of literature abstracts. *IBM Journal of Research Development*, 2(2):159–165.

Martínez-Santiago, F., M. C. Díaz-Galiano, U. na López L. A., and R. Mitkov. 2015. A semantic grammar for beginning communicators. *Knowledge-Based Systems*, 86:158–172.

Mihalcea, R. and B. Leong. 2006. Toward communicating simple sentences using pictorial representations. In *Association of Machine Translation in the Americas*.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations*, ICLR.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, volume 2 of *NIPS'13*, pages 3111–3119.

Mikolov, T., W.-T. Yih, and G. Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, HLT-NAACL, pages 746–751.

Murray, G., S. Renals, and J. Carletta. 2005. Extractive summarization of meeting recordings. In *6th Interspeech and 9th European Conference on Speech Communication and Technology*, pages 593–596.

Ozsoy, M., F. Alpaslan, and I. Cicekli. 2011. Text summarization using latent semantic analysis. In *Journal of Information Science*, volume 37, pages 405–417.

Ozsoy, M., I. Cicekli, and F. Alpaslan. 2010. Text summarization of turkish texts using latent semantic analysis. In *23rd International Conference on Computational Linguistics*, pages 869–876, August.

Schnabel, T., I. L. D. Mimno, and T. Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, pages 298–307.

Steinberger, J. and K. Jezek. 2004. Using latent semantic analysis in text summarization and summary evaluation. In *Proceedings of ISIM'04*, pages 93–100.

UzZaman, N., J. P. Bigham, and J. F. Allen. 2011. Multimodal summarization of complex sentences. In *Proceedings of the 16th International Conference on Intelligent User Interfaces*, IUI '11, pages 43–52.

Zhu, X., A. B. Goldberg, M. Eldawy, C. R. Dyer, and B. Strock. 2007. A text-to-picture synthesis system for augmenting communication. In *Proceedings of the 22nd National Conference on Artificial Intelligence*, AAAI'07, pages 1590–1595.

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 23-30 recibido 26-03-2018 revisado 02-05-2018 aceptado 03-05-2018

# Detecting the Central Units of Brazilian Portuguese argumentative answer texts

## *Detección de las unidades centrales para textos de respuesta argumentativa en Portugués-Brasileño*

**Kepa Bengoetxea**[1], **Mikel Iruskieta**[1], **Juliano Antonio**[2]
[1]University of the Basque Country (UPV/EHU). Ixa Group
[2]Universidade Estadual de Maringá (UEM). Funcpar.
kepa.bengoetxea@ehu.eus, mikel.iruskieta@ehu.eus, jdantonio@uem.br

**Abstract:** Understanding or writing properly the main idea or the Central Unit (CU) of a text is a very important task in exams. So, detecting automatically the CU may be of interest in language evaluation tasks. This paper presents a CU detector based on machine learning techniques for argumentative answer texts in Brazilian Portuguese. Results show that the detection of CUs following machine learning techniques in argumentative answer texts is better that those using rules.
**Keywords:** central unit, RST, argumentative answer texts

**Resumen:** Comprender o escribir correctamente la idea principal o Unidad Central (UC) de un texto es una tarea muy importante en los exámenes. Así, la detección automática de la UC puede ser de interés en las tareas de evaluación del lenguaje. Este artículo presenta un detector de UCs basado en aprendizaje automático para textos de respuesta argumentativa en Brasileño. Los resultados muestran que la detección de las UCs utilizando aprendizaje automático en brasileño y textos de respuesta argumentativa obtienen mejores resultados que los basados en reglas.
**Palabras clave:** unidad central, RST, textos de respuesta argumentativa

## *1 Introduction*

Information about discourse structure can improve the realization of complex linguistic tasks such as automatic summarization, text generation, segmentation, information extraction, sentence compression, automatic translation, paraphrasing and even evaluating texts. To do such tasks, we need annotated corpora to develop and test an automatic discourse parser. Regarding Brazilian Portuguese, CST News Corpus (Aleixo and Pardo, 2008) is a corpora with information about discourse structure, which consists of news texts annotated with Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and Dizer (Pardo and Nunes, 2004) is an automatic discourse analyzer for this language.

Both the corpora and the parser are based on RST, a theory that investigates the coherence relations which arise from the combination between text spans (Mann and Thompson, 1988). A very important notion for the theory is nuclearity. In asymmetric relations, the nuclear span is the member of the pair that is more essential to the writer's purpose than the others. As RST relations can be recursive, nuclearity works at different levels (top and bottom) of the relational discourse structure, which is hierarchical due to these asymmetric relations.

RST diagrams are represented as trees (henceforth RS-trees), and the central unit (henceforth CU; the most salient node of the rhetorical structure tree, which is at the top of the RS-tree) is an elementary discourse unit (henceforth EDU) which is not satellite of any other unit or text span.

The corpus studied was created with texts written by candidates for university entrance exams. The texts of the corpus were produced as an answer to the question "What's the secret of Vestibular: intelligence, effort or luck?". According to Menegassi (2011), argumentative answer genre belongs to scholar/academic sphere. It is initiated by the resumption of the question followed by the answer to the question, which is the thesis defended by the author. The remainder of the text presents arguments that support the thesis in order to try to convince or persuade

ISSN 1135-5948. DOI 10.26342/2018-61-2

the reader.

The aim of this paper is to build up a CU detector based on machine learning techniques for argumentative answer text in Brazilian Portuguese. The identification of the CU is relevant for the training of the Portuguese teachers committee which corrects the texts manually or also to the students who need some help in indicating the central idea in this genre, which can be crucial in their future,[1] and it is also crucial to develop a better discourse parser.

The option for building a corpus and not using an annotated existing corpus is derived from the goal of developing an application for the automatic detection of the CU of argumentative answer texts and also to show that an ADA have to detect first the CU and after elaborate the RS-tree.

## 2 *Theoretical background and related work*

In Antonio and Santos (2014) description of the rhetorical structure of argumentative answer genre, the initial statement is the CU of the text and its development is the satellite. In the corpus investigated by the authors, EVIDENCE relation is very common to be held between the satellite and the nucleus as the writer intends to increase the reader's belief on the content of the nucleus. ELABORATION relation also occurs in the corpus when writers intend to present additional information to the content of the nucleus. It must be noticed that not only these relations may be held between the central unit and the remainder of the text in argumentative answer texts. An example of argumentative answer text of our corpus is presented in Figure 1.

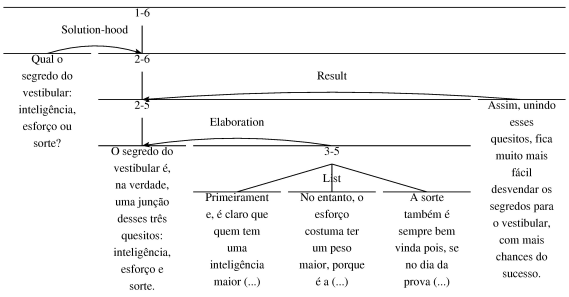


Figure 1: RS-tree of an argumentative answer text [M21294]

[1]The gold standard data can be consulted at http://ixa2.si.ehu.eus/rst/pt/.

There, after the question that the student has to answer ($span_1$), the text is divided into 5 spans. $Span_2$ is the CU, i.e., the unit which presents the main idea of the text. As the other spans are satellites regarding the CU, the arrows point towards $span_2$. $Spans_{3-5}$ hold ELABORATION relation with the nucleus (which is also the CU or the RS-tree). In ELABORATION relation, the satellite provides additional details about the elements of the nucleus (Mann and Thompson, 1988). In other words, the writer provides more information about the three secrets to achieve success in the entrance exams presented in $span_2$. In $span_3$ he elaborates intelligence, in $span_4$ he elaborates effort and in $span_5$ he elaborates luck. The relation held among $span_3$, $span_4$ and $span_5$ is LIST, a multinuclear rhetorical relation. Finally, $span_6$ is a RESULT satellite. The question that the student has to answer is related to the argumentative answer as a SOLUTION-HOOD relation, which is at the top of the RS-tree.

Regarding the identification of the most important discourse unit within the framework of RST, some applications have been developed: Pardo, Rino and Nunes (2003) developed an extractive summarizer for texts of any domain and Pardo an Nunes (2004) created DiZer, an ADA, both of them for Brazilian Portuguese.

The automatic discourse analyzer DiZer for BP (adapted especially to scientific texts) found a CU for this example[2] shown in Example (1):

(1) *Qual o segredo do vestibular: inteligência, esforço ou sorte?*

DiZer detected the question itself of the argumentative text as the CU, but if we introduce only the argumentative text in DiZer, the CU is the following: '*O segredo de o vestibular é, em a verdade, uma junção de esses três quesitos:*'.[3] The difference with the manual annotation is because of a different segmentation. DiZer left out the following part of text: *inteligência, esforço e sorte* from the $Span_1$ (in this case $EDU_1$). Whereas the CU for this text is: *O segredo do*

[2]There are several repositories and ways to use DiZer, we have used the *Portuguese by Dizer* and a greedy method to built the tree.

[3] English translation: The secret of vestibular is, in fact, a junction of three features:

*vestibular é, na verdade, uma junção desses três quesitos: inteligência, esforço e sorte.*[4]

Another similar program for detecting the most important idea in BP is GistSumm (Pardo, Rino, and Nunes, 2003). Example (2) was summarized with GistSumm at a 0.80 compresion rate.

(2) [*No entanto, o esforço costuma ter um peso maior, porque é a partir dele que os vestibulandos descobrem a melhor forma de estudar, muitas vezes abrem mão das horas de lazer para passarem mais tempo com os livros, enfim, quem é esforçado entende que não adianta ficar acomodado, é preciso dedicação para que se alcance o fim desejado.*]$_4$

GistSumm extracts the most important sentences using a ranking system based on words and sentence position. In the case of Example (2), GistSumm extracted the text span$_4$ in which the writer highlights the most important factor: 'effort'. Although the writer mentioned in the beginning of the text that the secret of *Vestibular* is a junction of the three factors (intelligence, effort and luck), in the span extracted by GistSumm, the writer assigns a bigger weight to effort.

There are similar works that detect the most important ideas of texts in different languages and domains: for English abstracts (Burstein and Marcu, 2003), for Spanish scientific texts (Bengoetxea and Iruskieta, 2018) and for Basque abstracts (Bengoetxea, Atutxa, and Iruskieta, 2017) developed a CU detector for scientific abstracts. For Brazilian Portuguese Iruskieta et al. (2016) used a rule-based automatic detector and the system gets a F-score 0.553 in the test dataset of this corpus.

## *3 Methodology*

### 3.1 Corpus and annotators

The corpus used in this paper consists of 100 texts written by candidates for Summer 2013 entrance exams[5] at *Universidade Estadual de Maringá* (UEM). There are excerpts the candidates can base upon to write the texts demanded by the instructions. On Summer 2013 the question to be answered by candidates was metadiscoursive. They had to write about the factors that lead to success in Vestibular (university entrance exams). The instructions were: "*As a candidate, write, using up to 15 lines, an argumentative answer to the question '**What is the secret of Vestibular: intelligence, effort or luck?**'*. You can base upon the information of the excerpts, but you cannot copy them".

| Corpus | Texts | Tokens | EDUs | CUs |
|---|---|---|---|---|
| Train | 60 | 8,499 | 846 | 69 |
| Test | 40 | 6,511 | 576 | 50 |
| Total | 100 | 15,010 | 1,422 | 119 |

Table 1: Corpus information

The gold standard we created contains 1,422 EDUs and 100 texts, each with its CUs (see Table 3.1). 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 there are 11.77% more texts with multiple CU (in total 20%) and 5.2% less CUs in the EDU$_1$ position (in total 60%).

This corpus was divided into 2 non-overlapping datasets: the first 60 texts as a training dataset and the last 40 texts as test dataset.

The annotation phases were as follows:

*i*. A corpus of 100 argumentative answer texts was collected.

*ii*. Four annotators segmented the texts manually into EDUs and, afterwards, a super annotator harmonized the segmentation.

*iii*. Four annotators determined the CU of each text, and finally the texts were harmonized.

### 3.2 Linguistic Features (LF)

Following Antonio (2015), the development of the CU detector was based on the frequency of some lexical and grammatical items or indicators which were, therefore, chosen as features (see Table 2).

The first feature is a list of nouns with a high frequency in the CU. They are re-

[4]English translation: The secret of vestibular is, in fact, a junction of three features: intelligence, effort and luck.

[5]The exams are available at `http://www.vestibular.uem.br/2013-V/uemV2013p2g1.pdf`.

lated to the meaning of junction or combination, such as *junção* 'junction', *combinação* 'combination', *união* 'union', *conjuntura* 'conjuncture', *miscigenação* 'miscigenation' (which was misused by the writer), *mistura* 'mixture', *soma* 'sum' and *mescla* 'mix'.

Regarding verbs, two types of verbs were also used to characterize the CUs: (*i*) copula verbs: *ser* 'to be' and *estar* 'to be', and (*ii*) evidential verbs which express propositional attitude (Dall'Aglio-Hattnher, 2007) such as *acreditar* 'to believe', *crer* 'to believe' and *pensar* 'to think' in first person singular.

Another group of features are bonus words: adverbs and adverbial phrases. These were also used by the writers in some CUs. It is important to remark that all epistemic adverbs used by writers were asseverative in an attempt to make their propositions more credible.

In argumentative answer genre, it is expected that writers resume the question before the answer. Thus, feature "title words" contains the words which were in the question that the writers were supposed to answer.

Finally, we found that the likelihood of a CU occurring at the beginning of the answer was quite high in the annotated data. To account for this, we used one feature that reflected the position of each EDU in argumentative answer texts.

Another group of features regards words which have low frequency in the CU and they are signals for other discourse structures (stop words). Thus, they can be used as cues in order not to identify the EDUs in which they appear as CU. It is plausible the fact that the conclusion may be stronger than the initial statement in terms of expressiveness. The conclusion is usually started by a discourse marker or a finisher expression such as *portanto* 'therefore', *enfim* 'ultimately', *a partir disso* 'taking this into account', *sendo assim* 'thus'. Besides that, the fact that the answer and the stronger arguments are restated, and this makes the conclusion seem more assertive than the initial statement. The same case happens when the writer repeats the question in the text, this kind of segment can be detected with an question mark "?" or interrogative pronoun (in this case, *qual* 'what').

Table 2 summarizes all the features we used in machine learning techniques.

| Group | Subgroup | Words |
|---|---|---|
| Nouns | junction of factors | junção, combinação, conjuntura, miscigenação, mistura, mix, soma, união, mescla |
| Verbs | copula | ser, estar |
| | evidential | acredito, creio, penso |
| Bonus words | epistemic | indubitavelmente, certamente, decerto, seguramente, obviamente, naturalmente, asseguradamente, inquestionavelmente, positivamente, decisivamente, incontestavelmente |
| | adverbial | sem dúvida, com certeza, de certeza, na verdade |
| Title words | resumption | segredo, vestibular |
| | factors | inteligência, esforço, sorte |
| Segment position | | the position of each segment in the text |
| Not in CU | question | question mark(?) and qual |
| Non CU connectors | conclusion | portanto, enfim, sendo assim, por isso, contudo, pois, a partir de, de este maneira, me o qual, então, e que, de este modo, assim, porém |

Table 2: Features extracted by a linguist from the training dataset

These features in argument answer texts are very different from other works that analyze other genre (Bengoetxea, Atutxa, and Iruskieta, 2017). So, we think that the indicators of the CUs are sensible to domain and genre. But there are some features that are common in most of them, such as: title words, segment position, epistemic adverbs, copula verbs and evidential verbs in first person singular.

### 3.3 Automatic Features (AF)

To detect the best features to tag the CU automatically we performed the following steps:

- 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 words. Finally, the training set dictionary obtained using this scheme contains 1000 features.
- We converted all letters to lower case.
- We followed bag of words approach and used tokens (unigrams, bigrams, trigrams and fourgrams) as features, where a classification instance is a vector of tokens appearing in the segmented text.
- We added segment position and title word occurrence information to the feature vector. Using weka's "string to word vector", text was converted into feature vector using TF-IDF as feature.
- We removed 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. We have tested the two most effective feature selection methods: *a*) Chi square and *b*) information gain using different set of attributes: 50, 100, 500 and 1000. Lastly, we performed all the classifiers using Chi square using a set of 100 attributes.

– Finally, the training set dictionary obtained using this scheme contains 100 features; the same dictionary was used for the test set.

## 4 The systems

In this paper, we conducted experiment in the Weka toolkit by using several supervised learning classifiers based on Support Vector Machines (SVM), Artificial Neural Networks, Bayes, Decision Tree and Rule-based method on the CU detector.

With the aim of selecting the best classifier, we have trained several learning classifiers on the indicators defined in Table 2 following Antonio (2015) using 10-fold cross-validation. The best systems are: *i*) Sequential Minimal Optimization (SMO) (Platt, 1998), *ii*) Multinomial Naive Bayes (MNB) and *iii*) Bernoulli Naive Bayes (BNB) system.

We have compared the results of three systems with a baseline and a rule-based automatic CU detector. A simple, but powerful baseline is based on the position of the given EDU into the whole document. The position is an important indicator, because we found that the likelihood of a CU occurring at the beginning of the answer was 65.2% in the training set. So we consider that the first segment is the only CU of the text as our baseline. The best rule-based system related in Iruskieta et al. (2016) also uses EDU position to detect the CU in argumentative answer texts. The best features in this system were: the EDU position (from 1 to 2) and at least 3 nouns or 3 title nouns of Table 2.

Our system has a module with 2 different stages to post-process the results from the classifiers:

*i*) The first stage deselects the conclusion segment (Post1): The most frequent error that the classifier performs in the cross-validation dataset is to select the conclusion segment as CU. So, if a EDU was considered as CU by the system and if it starts by a conclusion discourse marker (for example, 'therefore', 'thus', 'ultimately'), then the CU is deselected.

*ii*) The second stage select at least one CU (Post2): Sometimes the systems classify all the segments of a text as non-CU or the first post-processing stage has deselected the CU. Depending on the classifier, we can apply 2 different techniques to select at least one CU:

  – A statistical post-processing to select the CU. In the case of MNB and BNB, the classifier always returns the probability of an EDU to be labelled as CU. So, the statistical post-processing use this value to select at least the most likely EDU to be labeled as CU.

  – The first EDU as CU. In the case of SMO, we consider that the first EDU is a unique CU of the text, but in the case that the first segment has an interrogative mark, the second EDU will be chosen by the system as the CU of the text. We also applied this technique in BNB and MNB systems, selecting the best one in each case.

Both stages (Post1 and Post2) are applied sequentially to process the outputs of the classifiers.

## 5 Results

We estimated the performance of ours systems using Linguistic Features (LF) and Automatic Features (AF). We partitioned the 60 texts into 10 groups. We trained 10 times on 9/10 of the labeled data and evaluated the performance on the other 1/10 of the data.

The evaluation results in Table 3 show the average performance of our classifier using traditional recall (Rec.), precision (Prec.), and F-score ($F_1$) metrics.

As we reported in Table 3.1, out of a total of 846 EDUs there are 67 CUs on the 10-fold cross-validation dataset (0.079 diffi-

| System | | D | C | E | M | P | R | $F_1$ |
|---|---|---|---|---|---|---|---|---|
| B.line | | C | 45 | 15 | 24 | 0.750 | 0.652 | 0.697 |
| | | T | 24 | 16 | 26 | 0.600 | 0.480 | 0.533 |
| Rule-based | | C | | | | 0.824 | 0.627 | 0.712 |
| | | T | | | | 0.778 | 0.429 | 0.553 |
| LF+ BNB | W | C | 48 | 21 | 21 | 0.695 | 0.695 | 0.695 |
| | | T | 27 | 21 | 23 | 0.562 | 0.540 | 0.551 |
| | P1 | C | 48 | 18 | 21 | 0.727 | 0.695 | 0.711 |
| | | T | 27 | 13 | 23 | 0.675 | 0.540 | 0.600 |
| | P2 | C | 48 | 19 | 21 | 0.716 | 0.695 | 0.705 |
| | | T | 27 | 16 | 23 | 0.627 | 0.540 | 0.580 |
| LF+ SMO | W | C | 46 | 7 | 23 | 0.867 | 0.666 | **0.754** |
| | | T | 25 | 3 | 25 | 0.892 | 0.500 | **0.640** |
| | P1 | C | 46 | 6 | 23 | 0.884 | 0.666 | **0.760** |
| | | T | 25 | 1 | 25 | 0.961 | 0.500 | **0.657** |
| | P2 | C | 47 | 13 | 22 | 0.783 | 0.681 | **0.728** |
| | | T | 29 | 11 | 21 | 0.725 | 0.580 | **0.644** |
| AF+ SMO | W | C | 42 | 12 | 27 | 0.777 | 0.608 | 0.682 |
| | | T | 19 | 9 | 31 | 0.678 | 0.380 | 0.487 |
| | P1 | C | 41 | 6 | 28 | 0.872 | 0.594 | 0.706 |
| | | T | 19 | 4 | 31 | 0.826 | 0.380 | 0.520 |
| | P2 | C | 46 | 15 | 23 | 0.754 | 0.666 | 0.707 |
| | | T | 27 | 15 | 23 | 0.642 | 0.540 | 0.586 |
| AF+ MNB | W | C | 51 | 23 | 18 | 0.689 | 0.739 | 0.713 |
| | | T | 25 | 21 | 25 | 0.543 | 0.500 | 0.520 |
| | P1 | C | 49 | 15 | 20 | 0.765 | 0.710 | 0.736 |
| | | T | 25 | 11 | 25 | 0.694 | 0.500 | 0.581 |
| | P2 | C | 50 | 20 | 19 | 0.714 | 0.724 | 0.719 |
| | | T | 28 | 17 | 22 | 0.622 | 0.560 | 0.589 |

Table 3: Results obtained on cross-validation (C) and test (T) sets without any post-process (W), with the first stage of the post-process (P1) and with the both stages of the post-process (P2)

culty)[6] and 576 EDUs there are 49 CUs on the test dataset (0.085 difficulty).

Table 3 shows results obtained on the 10-fold cross-validation and test sets using *i*) a baseline (all the first segments are considered as the unique CU of each text), *ii*) the best rule-based system (Iruskieta, Antonio, and Labaka, 2016) and *iii*) the best different machine learning methods using linguistics features and automatic features. These machine learning methods are MNB, BNB and SMO, each of them has three stages applied sequentially: *i*) the initial stage without any post-process (Without), *ii*) a first post-process stage (Post1) and *iii*) a second post-process stage (Post2). We can say that the rule-based detector has similar results in comparison with system's with Post1, because the system does not ensure that each text has a least one CU and all the conclusions were avoided. We want to note that the task to detect a CU means that the system has to assign at least one CU for each text, which is necessary to start building a RS-tree. Taking this into account we have compared the systems into 2 groups:

- RB and system's with Post1 that may not return any CU for each text: The best model is LF+SMO+Post1 which provides 0.76 in cross-validation and 0.657 in test. Table 3 shows that LF+SMO+Post1 system is better than rule-based system in 0.048 points in cross-validation and 0.104 points in test dataset. In the second position is AF+MNB+Post1 model which is better than rule-based system and LF+BNB+Post1 system.
- Baseline and system's with Post2 that return at least one CU for each text: The best model is LF+SMO+Post2 which provides 0.728 in cross-validation and 0.644 in test. We can observe that LF+SMO+Post2 system is better than baseline system in 0.031 points in cross-validation and 0.111 points in test dataset. In the second position is AF+MNB+Post2 model which is better than baseline and LF+BNB+Post2 system.

To conclude, if we compare the system's with Post1 and Post2, the task of return at least one CU for each text reduces the accuracy in almost all the systems.

## 5.1 A comparison using box plot

To show how robust all the algorithms 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 2).

To visualize the performance of the 14 systems (Baseline (B), Rule-based (RB),[7] LF+BNB (LBNB), LF+BNB+Post1 (LBNBP1), LF+BNB+Post2 (LBNBP2), LF+SMO (LSMO), LF+SMO+Post1 (LSMOP1), LF+SMO+Post2 (LSMOP2), AF+MNB (AMNB), AF+MNB+Post1 (AMNBP1), AF+MNB+Post2 (AMNBP2), AF+SMO (ASMO), AF+SMO+Post1 (ASMOP1), AF+SMO+Post2 (ASMOP2)), we

[6] Where the nearer it is from 1 the easier it is to determine the CU.

[7] Results obtained in previous work using a rule-based system was 0.712 for train and 0.553 for test.

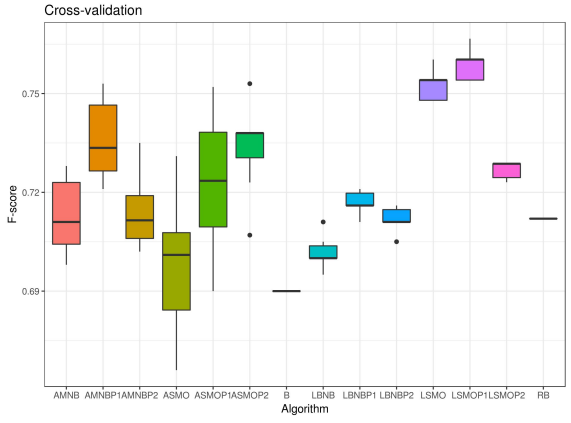


Figure 2: Exploring F-score distribution on the 10-fold cross-validation using 10 random seeds with Box Plot

have summarized the distribution of F-score values using box plots. 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 or outliers.

Taking this into account we have compared the systems into 2 groups:

- RB and system's with Post1 that may not return any CU for each text: The best model is LF+SMO+Post1 which has a F-score median value of 0.76. All the system's with Post1 show a greater F-score median value than RB. In the second position is a system that use automatic features like AF+MNB+Post1 which has a F-score median value of 0.723.

- Baseline and system's with Post2 that return at least one CU for each text: the best model is AF+SMO+Post2 (which has a F-score median value of 0.738) and it is close to LF+SMO+Post2 (which provides a F-score median value of 0.728).

If the system has to assign at least one CU for each text, which is necessary to start building a RS-tree, we finally have selected LF+SMO+Post2 system to avoid outliers.

To understand these results, we present an error analysis in the following subsection.

| **Correct CUs** | | **Wrong CUs due to** | |
|---|---|---|---|
| Total agreem. | Partial agreem. | Wrong Structure | Segment. errors |
| 25 | 4 | 10 | 1 |

Table 4: SMO's post-processed method error analysis of the test dataset

### 5.2 Error analysis

First of all, we checked manually if each text follows the patterns specified by Antonio (2015) and we found that the 10-fold cross-validation dataset follows in a better way than the test dataset. 25% of the texts do not have the prototypical characteristics of the CUs in the 10-fold cross-validation dataset, whereas in the test dataset 55% of the texts do not have the prototypical characteristics of the CUs, because the analyzed texts were written by students.

Secondly, in the following error analysis in Table 4, we analyze why the SMO+Post2 system does not detect all the CUs of the 40 texts from the test dataset and, mainly, errors happen with texts which do not present the expected structure for the argumentative answer genre.

## *6 Conclusions and future work*

In this paper we have introduced the first Central Unit (CU) detector based on machine learning techniques for Brazilian argumentative answer texts. The CU detector can be tested at http://ixa2.si.ehu.eus/rst/tresnak/cu-detector/bp/.

The results from our work indicate that identifying the CU discourse segment in argumentative answer text is well defined and the optimal set of features to classify the CUs are: title words, segment position, epistemic adverbs, copula verbs and evidential verbs in first person singular.

Detecting the CUs in real exams that have been written by students is difficult because sometimes they do not follow the discourse structure of an argumentative answer text, but of a dissertation. Different features have to be taken into account when we are detecting the CU in different genres.

We conclude that this system ensures, with a post-process stage, that each text has at least one central unit and we obtain better results using machine learning techniques (results with BNB+Post2 were 0.58 and with SMO+Post2 0.644) than using a rule-based

approach (0.553) (Iruskieta, Antonio, and Labaka, 2016). We think that there is some room to improve using punctuation information as other features in the post-processing, as we have to work with students argumentative answer texts.

We have shown that a CU detector based on machine learning techniques can be built by easily extracting manually some indicator as in Antonio (2015). We think that some parsers can benefit from this methodological step to find the CU after automatic segmentation and before linking the rhetorical relations of an RS-tree.

The work carried out will be useful if we can provide a fair evaluation of the argumentative answer texts, assigning better grades to those texts which follow the patterns of the CU and giving some indicators or clues to students to write the CU in a better way. We intend to develop different studies of how we can detect the CU in other languages, genres and domains taking into account annotated data and features developed here.

## *Acknowledgments*

This study was carried out within the framework of the following projects: IXA Group: natural language processing (GIU16/16) [UPV/EHU], QUALES KK-2017/00094 (Gobierno Vasco) and TUNER TIN2015-65308-C5-1-R (MINECO/FEDER, UE).

## *References*

Aleixo, P. and T. Pardo. 2008. CSTNews: um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (Cross-Document Structure Theory). Technical Report ICMC-USP.

Antonio, J. 2015. Detecting central units in argumentative answer genre: signals that influence annotators' agreement. In *5th Workshop "RST and Discourse Studies" in Actas del XXXI Congreso de la Sociedad Española del Procesamiento del Lenguaje Natural*. SEPLN.

Antonio, J. D. and J. A. Santos. 2014. A estrutura retórica do gênero resposta argumentativa. *Signum: Estudos da Linguagem*, 17(2):193–223.

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.

Bengoetxea, K. and M. Iruskieta. 2018. A supervised central unit detector for spanish. *Procesamiento del Lenguaje Natural 60: 29–36. ISSN 1135-5948. DOI 10.26342/2018-60-3.*

Burstein, J. and D. Marcu. 2003. A machine learning approach for identification of thesis and conclusion statements in student essays. *Computers and the Humanities*, 37(4):455–467.

Dall'Aglio-Hattnher, M. 2007. Pesquisas em sintaxe: a abordagem funcionalista da evidencialidade. *Trilhas de Mattoso Câmara e outras trilhas: fonologia, morfologia e sintaxe. Araraquara: Cultura Acadêmica Editora*, 12:103–145.

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.

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

Menegassi, R. J. 2011. A escrita na formação docente inicial: influências da iniciação à pesquisa. *Signum: Estudos da Linguagem*, 14(1):387–419.

Pardo, T. and M. Nunes. 2004. Dizer - um analisador discursivo automático para o português do brasil. In *In Anais do IX Workshop de Teses e Dissertações do Instituto de Ciências Matemáticas e de Computação*, pages 1–3, São Carlos-SP, Brasil. 19 a 20 de Novembro.

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

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

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 31-38 recibido 03-04-2018 revisado 02-05-2018 aceptado 04-05-2018

# Aproximación a un modelo de recuperación de información personalizada basado en el análisis semántico del contenido

## *An Approach of a Personalized Information Retrieval Model based on Contents Semantic Analysis*

**Eric Utrera Sust[1], Alfredo Simón-Cuevas[2], Jose A. Olivas[3] and Francisco P. Romero[3]**

[1]Universidad de las Ciencias Informáticas. Carretera a San Antonio de los Baños, km 2 ½, Torrens, Boyeros, CP.: 19370, La Habana, Cuba
ebutrera@uci.cu

[2]Universidad Tecnológica de La Habana "José Antonio Echeverría", Cujae
Ave. 114, No. 11901, CP: 19390, La Habana, Cuba
asimon@ceis.cujae.edu.cu

[3]Universidad de Castilla La Mancha
Paseo de la Universidad, 4, Ciudad Real, España
{JoseAngel.Olivas, FranciscoP.Romero}@uclm.es

**Resumen:** En este trabajo se presenta una primera aproximación de un modelo de recuperación de información personalizada basado en el procesamiento semántico del contenido. El modelo propuesto reduce la sobrecarga de información innecesaria para los usuarios y mejora los resultados recuperados mediante la combinación de un procesamiento semántico de contenido aplicado a las consultas y documentos indexados, y la información de los perfiles de usuarios. La aplicabilidad de la propuesta fue evaluada en el contexto de un motor de búsqueda real, a través de consultas diseñadas por expertos en diferentes dominios y la medición de su rendimiento. Los resultados obtenidos fueron comparados con los del motor de búsqueda puesto a prueba, lográndose mejoras en cuanto a la precisión y exhaustividad.

**Palabras clave:** recuperación de información personalizada, análisis semántico

**Abstract:** In this paper, an approach of a personalized information retrieval model based on the semantic processing of the content is proposed. The proposed model reduces the unnecessary information overload for users and improves the retrieval results through combining a content semantic processing applied to the queries and indexed documents, and information user processing from different perspectives. The applicability of the proposal was evaluated in the context of a real web search engine, through several queries designed by experts and associated to differents topics, and the measurement of their performance. The results were compared to those obtained by the search engine put to the test, achieving improvements the retrieval results.

**Keywords:** information retrieval systems, search engines, semantic processing

## 1 *Introducción*

Actualmente, los buscadores web no siempre ofrecen la información que el usuario necesita como resultado de una consulta. Algunos de los factores que inciden en esta situación son: el análisis del contenido (consulta – documento indexado) aún suele estar basado, fundamentalmente, en análisis sintáctico del contenido textual, sin tener en cuenta la semántica subyacente (Klusch, Kapahnke, Schulte, et al., 2016); no se tienen en cuenta los intereses de los usuarios (Jay, Shah, Makvana, et al., 2015); no existe tratamiento de la ambigüedad inherente al lenguaje natural (Shou, Bai, Chen, et al., 2014) y baja calidad en la formulación de consultas (Singh, Dey, Ashour, et al., 2017).

La personalización de la recuperación de información constituye una de las líneas que

ISSN 1135-5948. DOI 10.26342/2018-61-3 

actualmente se está trabajando para incrementar la calidad de los resultados de los buscadores web (Singh, Dey, Ashour, et al., 2017). En este sentido, se reporta el uso de varias técnicas, tales como: la expansión de consultas, la desambiguación de consultas, el uso de bases de conocimiento (ej. taxonomías, ontologías, etc.) y modelos de datos enlazados para mejorar los resultados de búsqueda (Tanaka, Spyratos, Yoshida, et al., 2015). El proceso de personalización cuando se integra a un Sistema de Recuperación de Información (SRI) de dominio general, para diferentes consultas y en diferentes contextos de búsqueda, es consistentemente menos efectivo que cuando se centra en un solo dominio (Makwana, Patel y Parth, 2017). Por otra parte, las consultas ambiguas, incompletas y breves, provocan que en ocasiones estos modelos de personalización generen grados de interés sobre documentos que no son los correctos, debido a que el usuario introduce la misma consulta de diferentes maneras y los documentos tienen diferentes contextos de búsqueda. Todo esto dificulta predecir el comportamiento y grado de interés de los usuarios sobre los documentos de forma adecuada y obtener nuevos conocimientos a través de esta interacción (Sharma y Rana, 2017). Se ha demostrado que el procesamiento de consultas y documentos influye en el proceso de mejorar la relevancia de los resultados (Hahm, Yi, Lee, et al., 2014; Corcoglioniti, Dragoni y Rospocher, 2016). En este sentido varios investigadores han propuesto soluciones enfocadas a SRI de dominio general (Zhang, Yan-hong, Wei-jun, et al., 2013; Preetha y Shankar, 2014; Shafiq, Alhajj y Rokne, 2015; Zhou, Lawless, Wu1, et al, 2016; Makwana, Patel y Parth, 2017) pero aún el tratamiento de las consultas y documentos no es suficiente.

En este trabajo se propone un Modelo de Recuperación de Información Personalizada (MRIP) basado en el procesamiento semántico del contenido para mejorar la eficacia de los SRI. MRIP está enfocado a un SRI de dominio general y se compone de 4 procesos: *Analizador de Contenidos*, *Generador de Perfiles*, *Personalizador y Generador de Ranking*. El *Analizador de Contenidos* procesa las consultas que se encuentran en el motor de consultas y los documentos almacenados en el índice del SRI. El *Generador de perfiles* recibe como entrada las acciones de los usuarios en el sistema y representa los perfiles de usuarios. El *Personalizador* es responsable de predecir qué documentos son interesantes para los usuarios y el *Generador de Ranking* es responsable de crear un ranking de los documentos recuperados por el MRIP. La aplicabilidad del MRIP fue evaluada por un total de 12 usuarios expertos y se integró en un motor de búsqueda web. Los resultados obtenidos fueron satisfactorios, mejorando la precisión y exhaustividad del motor de búsqueda.

El resto del documento está organizado de la siguiente manera: en la Sección 2 se analizan y caracterizan los trabajos relacionados con la problemática abordada; en la Sección 3 se presenta el modelo propuesto; en la Sección 4 se exponen y analizan los resultados del caso de estudio desarrollado para evaluar la aplicabilidad del modelo propuesto; y en la Sección 5 se presentan las conclusiones.

## *2 Trabajos relacionados*

Los modelos de personalización tienen como objetivo caracterizar a los usuarios que interactúan con el sistema desde la generación de un perfil que está en correspondencia con sus gustos e intereses (Zhou, Lawless, Wul, et al., 2016). En (Zhang, Yan-hong, Wei-jun, et al., 2013) se propone un motor de búsqueda personalizado distribuido, que se utiliza para minar el historial web del usuario y crear una base de datos de patrones de interés. El modelo de interés de los usuarios se expresa mediante una tríada ordenada de la forma: palabra interesada, peso de la palabra y grado de estreno de la palabra. En (Shafiq, Alhajj y Rokne, 2015) se propone un enfoque para encontrar los intereses personales y los contextos sociales de los usuarios, basándose en las actividades de los usuarios en sus redes sociales, Desarrollan un mecanismo que extrae información de la red social de un usuario y la utiliza para volver a clasificar los resultados de un motor de búsqueda. Trabajo similar el de (Zhou, Lawless, Wu1, et al, 2016), donde se construyen perfiles de usuario mejorados a partir de un conjunto de anotaciones y recursos que los usuarios han marcado. Presentan dos modelos probabilísticos para incorporar simultáneamente anotaciones sociales, documentos y la base de conocimiento externa, y un modelo de expansión de consulta para mejorar la búsqueda. El proceso de expansión se hace a partir de un conjunto de palabras en el perfil de usuario, con el objetivo de devolver una lista ordenada de términos de perfil que se agregarán a la consulta. En

(Makwana, Patel y Parth, 2017) se analizan los clics del usuario y se crean grupos de usuarios similares utilizando la técnica de agrupamiento de C-means difusa. La consulta pasa por un proceso de eliminación de ambigüedades a partir de criterios que se han especificado en otras búsquedas. Para cada término de búsqueda relevante, se registran los enlaces en los que hizo clic ese usuario y también calcula el valor de interés.

La mayoría de los trabajos presentados analizan el grado de interés de los usuarios sobre los documentos basándose en los términos claves con los que se relaciona, sin tener en cuenta que un documento puede tratar de "java" y sin embargo ese término puede tener una frecuencia de aparición baja. Por otra parte, el proceso de expansión y desambiguación de consultas se hace utilizando búsquedas anteriores, a través de un análisis léxico-sintáctico que no tiene en cuenta el sentido de las palabras.

## *3 Modelo de Recuperación de Información Personalizada (MRIP)*

MRIP (ver Figura 1), se compone de 4 procesos: *Analizador de Contenidos*, *Generador de Perfiles*, *Personalizador* y *Generador de Ranking.*

### 3.1 Analizador de contenidos

Este proceso se encarga de analizar las consultas que se encuentran en el motor de consultas y los documentos indexados. Se compone de dos subprocesos: *Procesamiento de consultas* y *Procesamiento de documentos*.

#### 3.1.1 Procesamiento de consultas

La gran parte de las consultas depende de la falta de contexto de búsqueda. Para solucionar este problema el subproceso para analizar las consultas está compuesto por 4 subprocesos: *Análisis*, *Expansión*, *Almacenamiento* y *Cálculo de similitud*. El subproceso *Análisis* apoyándose en Freeling (Padró, y Stanilovsky, 2012) y BabelNet (Navigli, y Ponzetto, 2012), identifica el idioma de la consulta, tokeniza, elimina palabras vacías, extrae las palabras claves, e identifica entidades nombradas. Para identificar el idioma se compara el texto entrado con los módulos disponibles para diferentes idiomas en Freeling y devuelve el idioma en el que está escrito el texto. Para extraer los tokens se utilizan las reglas de tokenización propuestas por Freeling y para identificar las entidades nombradas se utiliza BabelNet y sus servicios de máxima entropía para detectar personas, nombres y organizaciones sobre las palabras claves. El subproceso *Expansión* extrae las relaciones semánticas entre los términos utilizando BabelNet, extrayendo sinónimos, hipónimos e hiperónimos para cada término de la consulta. La tercera etapa *Almacenamiento* colecciona la consulta expandida para el usuario en los repositorios del modelo de la Figura 1.

El cálculo de similitud semántica es aplicado en diferentes áreas del conocimiento y en el caso específico de los SRI posibilita que se encuentren resultados de búsquedas similares a

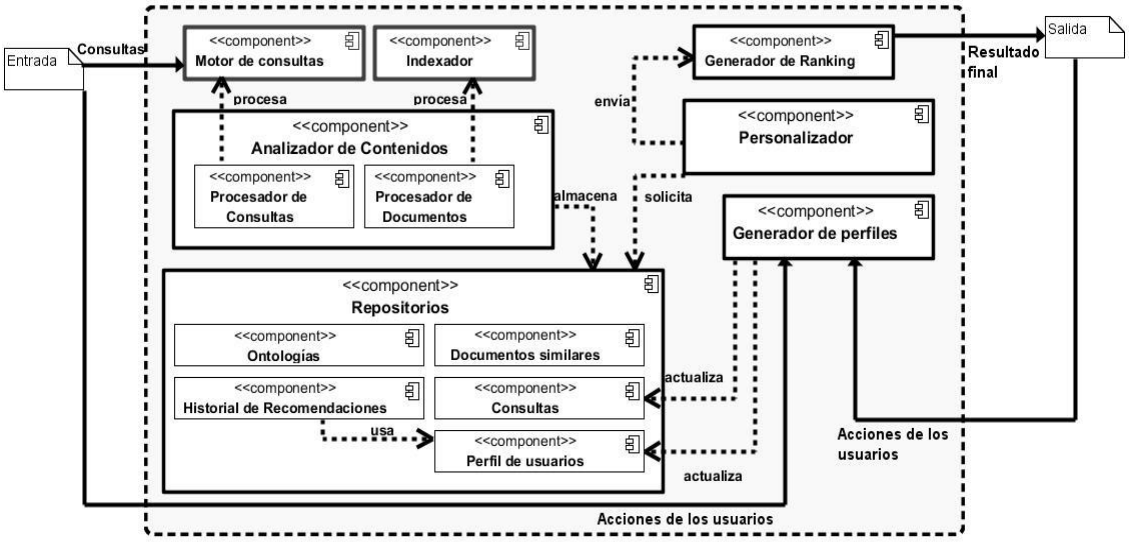


Figura 1. MRIP basado en el procesamiento semántico del contenido

las consultas de los usuarios. El subproceso *Cálculo de similitud* utiliza la medida propuesta por (Mihalcea, y Strapparava, 2006) por su utilidad en el cálculo de similitud entre cadenas de textos (ver fórmula (1)). Este cálculo se almacena en los Repositorios de la figura 1 como una matriz $M_{2x2}(\mathbb{R})$.

$$\begin{aligned}&\mathrm{sim}(C_1, C_2)\\&= \frac{1}{2}\left(\frac{\sum_{w\in\{C_1\}}(maxSim(w, C_2) * idf(w))}{\sum_{w\in\{C_1\}} idf(w)}\right.\\&\left. + \frac{\sum_{w\in\{C_2\}}(maxSim(w, C_1) * idf(w))}{\sum_{w\in\{C_2\}} idf(w)}\right) \quad (1)\end{aligned}$$

donde:
*C1 y C2:* consultas a analizar.
*maxSim (w, C1):* similitud semántica máxima entre la palabra w en la consulta *C1* y cada una de las palabras en la consulta *C2*.
*idf (w):* frecuencia inversa en el documento de la palabra w.

### 3.1.2 Procesamiento de documentos

Este proceso, aplicado a los documentos indexados, incluyen las siguientes tareas: (1) categorización y (2) análisis de similitud.

La categorización de los documentos indexados consta de 5 pasos: pre-procesamiento, extracción de bi-gramas, construcción de la colección, construcción del categorizador y categorización del documento. Los documentos son categorizados teniendo en cuenta 150 categorías (ej. Economía, Justicia, Biología, Ciencias de la información, etc.), las cuales se tomaron de Wikipedia (en Español e Inglés). En la etapa de pre-procesamiento el primer paso es la conversión de palabras a minúsculas, luego se eliminan las palabras vacías y se eliminan los acentos. En el proceso de creación de la colección se le aplica el pre-procesamiento a cada documento de una categoría, se extraen los bigramas de ese documento y se agregan al vector de contenido del documento como una palabra unida. Luego se crea un vector formado por los términos y bigramas (unión de las palabras de un bigrama) y su frecuencia de aparición. La colección de entrenamiento se crea con todos estos vectores formados por las palabras claves y su frecuencia de aparición en el documento. Finalmente, el proceso de categorización consiste en cargar la colección de entrenamiento, realizarle el proceso de pre-procesamiento al texto que se desea categorizar, extraer los bigramas del texto, agregar la unión de un bigrama como término independiente al vector de texto, crear el vector con palabras y su frecuencia de aparición y la aplicación del algoritmo Naïve Bayes Multinominal para la predicción de la categoría a la que pertenece.

*Similitud entre los documentos.* La similitud semántica mide la fuerza de las relaciones semánticas entre los conceptos incluidos en un documento. En este proceso, los documentos a evaluar son representados en forma de vectores, los cuales están formados por las palabras claves presentes en dichos documentos, y su extensión con otros términos relacionados capturados de BabelNet. La construcción de los vectores se lleva a cabo a partir de: eliminación de los términos repetidos en los documentos, extracción de las palabras claves según frecuencia de ocurrencia de los términos en el documento, etiquetado POS para obtener únicamente nombres, verbos y adjetivos de los documentos de entrada, la desambiguación de los términos usando Babelfy (Moro, Cecconi, y Navigli, 2014), y finalmente, la captura de BabelNet de otros términos sinónimos, hipónimos e hiperónimos, relacionados con las palabras claves identificadas. La evaluación de la similitud entre dos documentos se lleva a cabo usando el coeficiente de similitud de Jaccard (Chunzi y Wang, 2017) sobre los vectores característicos de los documentos. Finalmente, la similitud entre documentos es almacenada en los *Repositorios* de la Figura 1 como una matriz $M_{2x2}(\mathbb{R})$.

## 3.2 Generador de perfiles

Los perfiles se generan tanto para usuarios registrados como no registrados. Cuando el usuario no está registrado, se crea un perfil utilizando las cookies del navegador. El Generador de Perfiles recibe como entrada las acciones implícitas (información capturada a través de las acciones que realiza el usuario sobre los resultados de la búsqueda) y explícitas (temas de interés y datos personales registrados por los usuarios en el sistema, por ejemplo: datos demográficos, fecha de nacimiento, sexo, educación y temáticas favoritas). El perfil del usuario $P_u$ es descrito por los siguientes elementos:

*Id_Usuario, Profesión, Fecha_Nacimiento, Localidad, Consulta, Documentos_Consultados, Grado_Interés, Temáticas_Preferidas, Factor_Olvido, Expiración_Cookie.*

Para la generación de los perfiles de los usuarios se aplica la técnica de aprendizaje automático basada en el Modelo Espacio-Vectorial, donde las variables de $P_u$ se representan como vectores.

## 3.3 Personalización de la Recuperación

El proceso de la personalización de los resultados de la recuperación en el MRIP está basado en un análisis que combina la modelación de las relaciones Usuario-Consulta, Usuario-Documentos y Usuario-Temática, a partir de las preferencias capturadas del usuario. Para generar *Grado_Interés*, $P_u$ necesita de tres variables $P_u$ ($u_id_j$, $u_iq_j$, $u_it_j$), donde $u_id_j$ representa las acciones realizadas por el usuario $u_i$ sobre los documentos consumidos $d_j$, $u_iq_j$ representa las consultas $q_j$ realizadas por el usuario $u_i$ y $u_it_j$ representa el porciento de búsqueda del usuario $u_i$ por la temática $t_j$.

*Correlación Usuario-Documento ($u_id_j$).* Se encarga de predecir cuáles de los documentos $d_i$ puede ser interesante para el usuario $u_i$, basándose en un entorno de conocimiento que representa las acciones de $u_i$ sobre $d_i$. Este entorno es creado basándose en el grado de interés que tiene un usuario sobre un documento recuperado, capturado mediante un conjunto de reglas. En estas reglas se modela el comportamiento del usuario sobre un documento considerando acciones tales como: *Like (L), Dislike (DL), Share (S), Do not share (NS), Visit (V),* y *Do not visit (NV)* y se infiere su grado de interés (GI). Las reglas definidas son:

1. R1: *Si L ∧ NV ∧ NS, entonces GI = Irrelevante*;
2. R2: *Si DL ∧ NV ∧ NS, entonces GI = Irrelevante*;
3. R3: *Si S ∧ NV, entonces GI = Irrelevante;*
4. R4: *Si S ∧ DL ∧ NV, entonces GI = Irrelevante;*
5. R5: *Si DL ∧ V ∧ NS, entonces GI = Relevancia baja;*
6. R6: *Si S ∧ L ∧ NV, entonces GI = Relevancia baja;*
7. R7: *Si S ∧ V, entonces GI = Relevancia Media;*
8. R8: *Si S ∧ DL ∧ V, entonces GI = Relevancia Media;*
9. R9: *Si S ∧ L ∧ V, entonces GI = Relevancia Alta; y*
10. R10: *Si NS ∧ L ∧ V, entonces GI = Relevancia Alta.*

El subproceso por otra parte utiliza una matriz donde se representan las reglas aplicadas por los usuarios sobre los documentos a partir de una consulta. Luego predice a partir del grado de similitud entre los documentos, las reglas que puede tener en cuenta un usuario con respecto a un documento que no ha visitado.

Finalmente, para buscar la similitud entre los vectores de calificación de dos usuarios *u* y *v* se utiliza el coeficiente de correlación de Pearson (Desrosiers y Karypis, 2011). La fórmula 1 muestra como calcular el Coeficiente de correlación de Pearson.

$$\mathrm{PC}(u,v) = \frac{\sum_{i\in I_{uv}}(r_{ui}-\bar{r}_u)(r_{vi}-\bar{r}_v)}{\sqrt{\sum_{i\in I_{uv}}(r_{ui}-\bar{r}_u)^2\sum_{i\in I_{uv}}(r_{vi}-\bar{r}_v)^2}} \quad (1)$$

donde:
$r_{ui}$ : valoración que ha dado el usuario u documento i.
$r_{vi}$: valoración que ha dado el usuario v al documento i.
$\bar{r}_u$ y $\bar{r}_v$ : total de documentos valorados en común por los usuarios u y v respectivamente.

*Correlación Usuario-Consulta ($u_iq_j$).* Este subproceso es el encargado de medir los usuarios similares al usuario actual a partir de una matriz donde se almacenan la similitud entre las consultas. Posteriormente al igual que el subproceso visto anteriormente busca la similitud entre el usuario actual y los demás usuarios utilizando el Coeficiente de correlación de Pearson.

*Correlación Usuario-Temática ($u_it_j$).* Tiene la responsabilidad de identificar los usuarios $u_i$ que son similares al usuario actual a partir del porciento de temáticas T (Economía, Justicia, Biología, Ciencias de la información, etc.), preferidas por cada uno. Al igual que los subprocesos anteriores, busca la correlación lineal entre el usuario actual y los demás usuarios utilizando el Coeficiente de correlación de Pearson.

*Decisor*. Se encarga de decidir finalmente el grado de interés de un usuario sobre un documento. Luego de obtener la similitud entre el usuario actual y los demás por cada uno de los subproceso anteriores, este calcula un promedio de similitud general $sim(u,v)$. Posteriormente selecciona los *k-vecinos* usando la técnica *Máximo Número de Vecinos* la cual consiste en seleccionar los *k* usuarios que son más similares

al usuario activo, donde *k* será un parámetro del algoritmo. Finalmente, para predecir el Grado de Interés que un usuario *u* tendría sobre un documento $d_i$ que no ha visitado, fue usado la fórmula de la *Media Ponderada* (fórmula (2)) (Ricci, Rokach, Shapira, et al., 2010):

$$\mathrm{MP(u,i)} = \frac{\sum_{v \in G_{u,i}} sim(u,v) * r_{v,i}}{\sum_{v \in G_{u,i}} sim(u,v)} (2)$$

donde:
*v*: grupo de usuarios que han valorado un documento *i*.
$r_{v,i}$: voto del usuario *v* al documento *i*.
*sim(u,v)*: valor de correlación calculado anteriormente entre el usuario *u* y *v*.

Debido a que el interés del usuario sobre un documento puede disminuir, se incluye el cálculo del factor olvido puesto en práctica por (Wang, Li., Lin, et al., 2017).

$$F(d) = e^{\frac{\log_2(\Delta t)}{f}} \quad (3)$$

donde:
*f*: cantidad de días
$\Delta t$: período de tiempo entre la última actualización del grado de interés sobre el documento di y el grado de interés actual.

Este valor es actualizado en el campo *Factor_Olvido* registrado en el perfil del usuario. El nuevo grado de interés (NGI) sobre el documento $d_j$ se calcula:

$$NGI(d) = VGI(d) * F(d) \quad (4)$$

donde:
*VGI(d):* viejo grado de interés sobre $d_j$.
*F(d):* factor olvido calculado en la fórmula 2.

## 3.4 Generador de ranking

El componente *Generador de Ranking* es responsable de hacer un nuevo ranking entre los documentos recuperados por el MRIP a partir de una consulta. Posteriormente, consulta la matriz de documentos con sus similitudes (ya almacenadas en los repositorios) y finalmente devuelve los documentos de mayor similitud en la parte superior de los resultados.

# *4 Caso de estudio: Red Cuba*

El modelo propuesto fue implementado e integrado a un buscador web denominado Red Cuba (https://www.redcuba.cu/) con el objetivo de evaluar su aplicabilidad en un escenario real. Red Cuba está basado en un modelo de Recuperación de Información Híbrido (Booleano-Espacio vectorial) que se basa en el análisis de la frecuencia de los términos en los documentos. Está llamado a ser la principal fuente de acceso a información cubana en Internet y actualmente cuenta con más de 2 millones de contenidos indexados, de los cuales más del 90% no son indexados por buscadores internacionales. Se realizó un experimento en el que participaron 12 expertos en diferentes temáticas, (Educación, Turismo, Comercio, Juegos, Cultura, Cocina, Deporte, Política, Tecnología, Moda, Dirección, y Pedagogía), cada uno de los cuales definió una consulta asociada a su área de experiencia. Se propuso de esta forma para lograr tener un control de las consultas asociadas a documentos que responden a temáticas de interés para un experto.

Cada usuario seleccionó los documentos más relevantes (DR) para dicha consulta (respuesta deseada), según lo indexado en el buscador; de ellos solo 6 crearon su perfil (P) en el buscador. En la Tabla 1 se describen y caracterizan dichas consultas.

| Id. | Palabras claves | P | DR |
|---|---|---|---|
| **Q1** | Libro, autor, P. J. Deitel | Si | 35 |
| **Q2** | Hotel, La Habana, 5 estrellas | Si | 30 |
| **Q3** | Tienda, camisas, blancas | Si | 39 |
| **Q4** | Juegos, Android, ciencia | Si | 48 |
| **Q5** | Escritor, poemas, cubanos | Si | 28 |
| **Q6** | Pasta, Bocaditos, helado | Si | 39 |
| **Q7** | Director, equipo, beisbol, cubano, sub23 | No | 28 |
| **Q8** | Aborígenes, Cuba | No | 32 |
| **Q9** | Creadores, computadora, cubana | No | 38 |
| **Q10** | Pelo, plancha, suavizador | No | 12 |
| **Q11** | Leyes, gaceta, oficial | No | 28 |
| **Q12** | Modelo, formación, integral, estudiantes | No | 45 |

Tabla 1. Consultas y tipos de usuarios para el experimento

Las métricas de precisión y exhaustividad, fueron utilizadas para medir y comparar la eficacia del buscador Red Cuba, aplicando el modelo propuesto y sin aplicarlo, tomando de referencia los documentos relevantes identificados por los expertos. En la medición de los resultados solo se tuvo en cuenta los 50

primeros documentos recuperados y los resultados se muestran en la Figura 2 y 3.

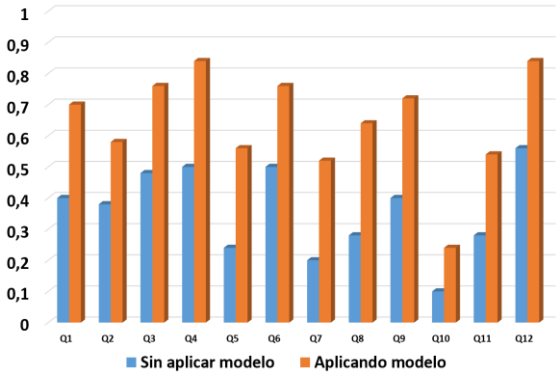


Figura 2. Resultados de la precisión

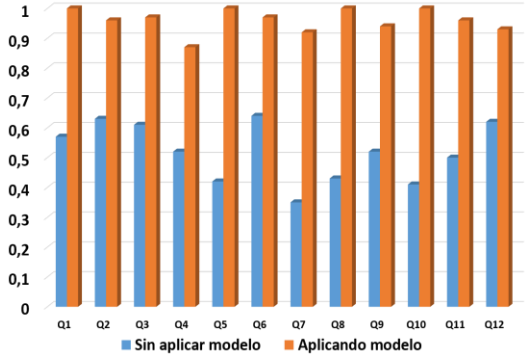


Figura 3. Resultados de la exhaustividad

Se puede observar como los resultados de precisión y exhaustividad aplicando el modelo al motor de búsqueda son mejores que sin aplicarlo, lo que demuestra que la integración de todos estos componentes contribuye a mejorar los resultados. Se observa además como los resultados de exhaustividad en el buscador de las consultas Q1, Q5, Q8 y Q10 aplicando el modelo llegan a tomar valores de 1. Esto constituye un resultado positivo ya que se logra maximizar el nivel de satisfacción del usuario en estas consultas.

Se está trabajando en la publicación de un subconjunto de datos que permita replicar los experimentos. Mientras tanto se podría replicar el experimento rastreando el buscador e indexando sus contenidos.

## 5 *Conclusiones*

Con los elementos teóricos y prácticos más actuales en el campo de los buscadores web y los sistemas de personalización, se desarrolló un modelo para la recuperación de información personalizada basada en el procesamiento semántico del contenido. El modelo propuesto además de ser un sistema que utiliza técnicas características de sistemas de personalización y de web semántica, posee mecanismos que permiten personalizar los resultados basándose en el perfil del usuario y el significado que tiene el contenido de su preferencia con el que interactúa. El uso de la semántica propicia tener la disponibilidad de un importante cúmulo de conocimiento multi-dominio. La evaluación final comprobó que la personalización de los resultados, guiando al usuario desde el inicio con sus intereses y conociendo el sentido semántico de las consultas y documentos, se logran mejores resultados de búsqueda, disminuyendo en gran medida la sobrecarga de información innecesaria.

## *Agradecimientos*

Este trabajo ha sido parcialmente financiado por el proyecto METODOS RIGUROSOS PARA EL INTERNET DEL FUTURO (MERINET), financiado por el Fondo Europeo de Desarrollo Regional (FEDER) y el Ministerio de Economía y Competitividad (MINECO), Ref. TIN2016-76843-C4-2-R.

## 6 *Bibliografía*

Chunzi, W. y B. Wang. 2017. Extracting Topics Based On Word2vec and Improved Jaccard Similarity Coefficient. En *Proceedings of the IEEE 2nd International Conference On Data Science in Cyberspace*, páginas 389-397.

Corcoglioniti, F., M. Dragoni y M. Rospocher. 2016. Knowledge extraction for information retrieval. En *Proceedings of the International Semantic Web Conference. Springer, Cham*, páginas 317-333.

Desrosiers, Ch. y G. Karypis. 2011. A Comprehensive Survey of Neighborhood-Based Recommendation Methods. *Recommender Systems Handbook. Springer*, páginas 107-144.

Hahm, G. J., Yi, M. Y., Lee, J. H., Suh, H. W. 2014. A personalized query expansion approach for engineering document retrieval. *Advanced Engineering Informatics*, 28(4): 344-359.

Johnson., M.S. 2016. Personalized Recommendation System for Custom Google Search. *International Journal of Computer & Mathematical Sciences*, 5:2347–8527.

Jay, P., P. Shah, K. Makvana y P. Shah. 2015. Review On Web Search Personalization Through Semantic Data. En *Proceedings of the IEEE International Conference On Electrical, Computer and Communication Technologies*, páginas 1-6.

Klusch, M., P. Kapahnke, S. Schulte, F. Lecue y A. Bernstein. 2016. Semantic Web Service Search: A Brief Survey. *Ki-Künstliche Intelligenz*, 30(2):139-147.

Mihalcea, R. y C. Strapparava. 2006. Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity. AAAI, páginas 775-780.

Moro, A., F. Cecconi, R. Navigli. 2014. Multilingual Word Sense Disambiguation and Entity Linking for Everybody. En *Proceedings of the International Semantic Web Conference*, páginas 25-28.

Makwana, K., J. Patel y S. Parth. 2017. An Ontology Based Recommender System to Mitigate the Cold Start Problem in Personalized Web Search. En *Proceedings of the International Conference on Information and Communication Technology for Intelligent Systems. Springer*, páginas 120-127.

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

Padró, L. y E. Stanilovsky. 2012. Freeling 3.0: Towards Wider Multilinguality. En *Proceedings of the International Conference On Language Resources and Evaluation Lrec2012*.

Preetha, S. y V. Shankar. 2014. Personalized search engines on mining user preferences using click through data. En *Proceedings of the Information Communication and Embedded Systems, IEEE*, páginas 1-6.

Ricci, F., L. Rokach, B. Shapira y P.B. Kantor. 2010. Recommender Systems Handbook. *Springe*r.

Singh, A., N. Dey, A. Ashour y V. Santhi. 2017. Web Semantics for Personalized Information Retrieval. *IGI Global. Information Science Reference*, páginas 166-186.

Shou, L., H. Bai, K. Chen y Ch. Chen. 2014. Supporting Privacy Protection in Personalized Web Search. *IEEE Transactions On Knowledge and Data Engineering*. 26(2): 453-467.

Sharma, S. y V. Rana. 2017. Web Personalization through Semantic Annotation System. *Advances in Computational Sciences and Technology*, 10(6):1683-1690.

Shafiq, O., R. Alhajj. y J. G. Rokne. 2015. On personalizing Web search using social network analysis. *Information Sciences*, 314: 55-76.

Tanaka, Y., Spyratos, N., Yoshida, T., Meghini, C. 2015. En *Proceedings of the Information Search, Integration and Personalization*. páginas 1–2.

Verma, D. y B. Kochar. 2016. Multi Agent Architecture for Search Engine. *International Journal of Advanced Computer Science and Applications*, 7(3): 224-229.

Wang, M., Q. Li., Y. Lin, y B. Zhou. 2017. A personalized result merging method for metasearch engine. En *Proceedings of the 6th International Conference on Software and Computer Applications. ACM*, páginas 203-207.

Zhou, D., S. Lawless, X. Wu1, W. Zhao y J. Liu. 2016. Enhanced Personalized Search Using Social Data. En *Proceedings of the Conference On Empirical Methods in Natural Language Processing*, páginas 700-710.

Zhang, H., M. Yan-hong, M. Wei-jun, y B. Zhong-xian. 2013. Study of Distributed Personalized Search Engine. *Advanced Materials Research. Trans Tech Publications*, páginas 1035-1039.

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 39-46 recibido 04-04-2018 revisado 30-04-2018 aceptado 04-05-2018

# Getting answers from semantic repositories: a keywords-based approach

## *Obteniendo respuestas de repositorios semánticos usando palabras clave*

**Francisco Abad Navarro**[1], **Jesualdo Tomás Fernández Breis**[1]
[1] Departamento de Informática y Sistemas, Universidad de Murcia,
IMIB-Arrixaca, CP 30100 Spain
email:francisco.abad@um.es, jfernand@um.es

**Abstract:** The Web of Data proposes to publish and connect data by applying the semantic web technologies for the representation of knowledge and data and the definition of queries. The success of the Web of Data requires that both humans and machines are able to extract information from such semantic repositories. For this purpose, query interfaces for humans must make the interaction with the repository as transparent as possible. In this work, we present a generic method for querying semantic repositories based on processing and recognizing the keywords input by the users as entities of the ontology used in the description of the data. A SPARQL query is automatically derived from the query graph extracted from the list of keywords. We also describe the application of the method to three different semantic repositories from different domains.
**Keywords:** Semantic web, question answering, ontology, sparql, rdf, automatic query construction, semantic exploitation

**Resumen:** La Web de Datos propone publicar y conectar los datos utilizando las tecnologías de la web semántica para la representación del conocimiento, de los datos y la especificación de consultas. El éxito de la Web de Datos requiere que tanto los humanos como las máquinas sean capaces de obtener información en estos repositorios semánticos. Para ello, los interfaces de consulta para humanos deben hacer lo más transparente posible el proceso de interacción con el repositorio. En este trabajo presentamos un método genérico para la consulta de repositorios semánticos basado en el reconocimiento de las keywords introducidas por los usuarios como entidades de la ontología utilizada para la descripción de los datos. Del procesamiento del grafo de consulta generado se deriva automáticamente la consulta SPARQL que se ejecuta contra el repositorio. Describimos el uso del método con tres repositorios de distintos dominios.
**Palabras clave:** Web semántica, pregunta-respuesta, ontología, sparql, rdf, construcción automática de consultas, explotación semántica

## *1 Introduction*

The Semantic Web (Berners-Lee, Hendler, and Lassila, 2001) is a set of technologies whose principal goal is to describe the data on the web in such way that a machine can read and process it easily. Ontologies (Bechhofer, 2009) play a main role as part of this set of technologies. They describe how the data is organized through concepts, their properties and relations with each others expressed by logical axioms. The other great pillar of the Semantic Web is RDF (Resource Description Framework) (Klyne and Carroll, 2006). RDF is used to represent the data as triples formed by subject, predicate and object, where subjects are concrete instances of concepts from the ontology, predicates are relations or properties described in the ontology and objects could be another instance of a concept or a primitive value. All the data in a RDF repository could be represented as a graph, where nodes are instances or primitive values and edges are relations or properties connecting the nodes.

ISSN 1135-5948. DOI 10.26342/2018-61-4

The development of Web of Data is the major objective of the Semantic Web initiative called Linked Open Data (LOD) (Bizer et al., 2008). There, data is ideally shared using formats like RDF, and the meaning of the entities is provided by an ontology. The LOD has penetrated in many domains such as biology (Consortium, 2016) or music (Swartz, 2002). The data stored in these repositories can be navigated through a web browser or queried using SPARQL. The exploitation of such repositories by non-semantic web experts is hampered by the need of knowing such language, so there is a need for making semantic web technologies as transparent as possible for human users interested in exploiting semantic repositories.

In the last years there have been different approaches related to controlled or natural language interfaces that automatically generate SPARQL queries, such as autoSPARQL (Lehmann and Bühmann, 2011), FREyA (Damljanovic, Agatonovic, and Cunningham, 2011) or OWLPath (Valencia-García et al., 2011), but they do not emulate the keywords-based search users are familiar with.

In this paper we propose a method which processes the keywords input by the users, generates the tree that interprets the query and automatically designs and executes SPARQL queries. The application of the method to three freely available, not developed by us, SPARQL endpoints in Spanish and English language is also reported in this paper.

## *2 Method*

In this section we describe our method, which can be applied to any RDF dataset whose vocabulary is provided by an ontology and which offers a SPARQL endpoint for querying. Our query model assumes that the input consists of keywords in natural language, which have to be processed, transformed into semantic entities and then used for the creation of SPARQL queries.

The method consists of the following modules: text normalizer, index builder, named entity recognizer, tree generator and SPARQL query generator. These modules are summarized in Figure 1 and detailed next:

### 2.1 Text normalizer

We use a language dependent text normalization in order to transform natural language text into a canonical form, so different forms of the same word are translated into a unique representation. Currently Spanish and English are the languages supported. This process is performed in two steps:

1. Preprocessing: the input is converted to lower case and common words and special characters as *, + or & are removed.
2. Normalization: the preprocessed text is the input for a Stanford NLP pipeline (Manning et al., 2014), which provides the following annotation tools, which are configured depending on the language used: token annotator, sentence annotator and part of speech annotator. Also a custom stem annotator was implemented through Snowball stemmer API [1] and appended to the Stanford pipeline.

At the end of the normalization step, the detected stems are concatenated with a blank space between them. Table 1 shows several examples of the text normalization.

| **Language** | **Original Text** | **Normalized Text** |
|---|---|---|
| Spanish | Empresas SL | empres |
| Spanish | Pirineos (los) | pirine |
| Spanish | teléfono | telefon |
| English | Companies | compani |
| English | United States (the) | unit state |
| English | telephone | telephon |

Table 1: Examples of text normalization y Spanish and English language

### 2.2 Index builder

We use a text index for the recognition of the named entities in the input text. This index is built by extracting the labels of the ontology classes and properties, as well as of the individuals stored in the dataset. The index contains the following fields:

- URI. The uniform resource identifier of an element. For example `<http://opendata.caceres.es/recurso/cultura-ocio/museos/Museo/10-museo-de-armas>`.

[1] `http://snowballstem.org/`

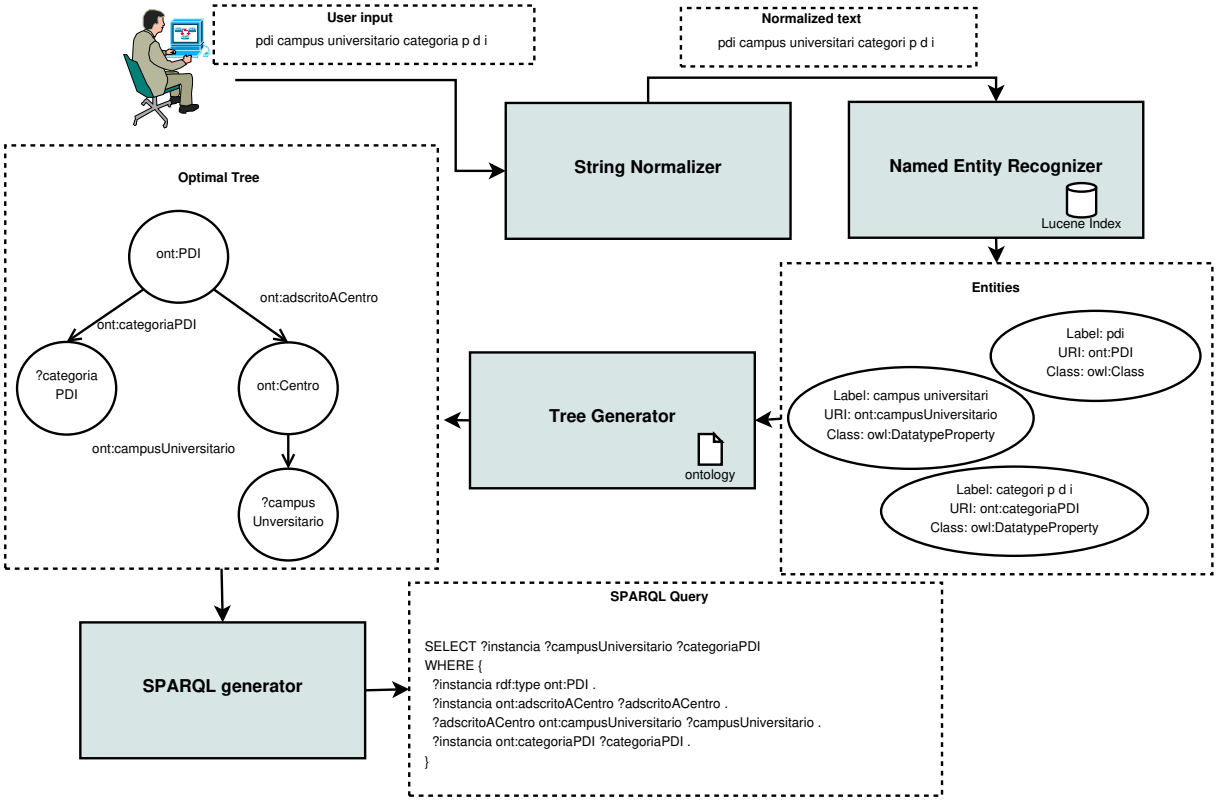


Figure 1: Method pipeline example

- Local name. The last part of the URI. For example `10-museo-de-armas` in the previous URI.
- Class. The class of the ontology that the element belongs to. In this case, `10-museo-de-armas` belongs to `<http://opendata.caceres.es/def/ontomunicipio#Museo>`. If the element is not an individual, this field is set as `owl:Class` for classes, `owl:DatatypeProperty` for properties, `owl:ObjectProperty` for relations.
- Type. The type of element, that could include "INSTANCE", if the element is an individual or "CLASS", "PROPERTY" or "RELATION" if the element is an ontology class. In this case the type of `10-museo-de-armas` would be "INSTANCE".
- Label. The original label associated with the element. For example "Museo de Armas". The value of this field is obtained from the `rdfs:label` annotations, filtering by the language used.
- Preprocessed label. The label after the preprocessing described in Section 2.1. For example "museo de armas".
- Normalized Label. The label after the normalization described in Section 2.1. For example "muse de armas".

While classes, relations and properties are extracted from the ontology owl file directly through `rdfs:label` annotations, a query is performed for each class to retrieve its individuals and their labels, also filtered by language. It is possible to specify properties that could play a textual label role, for instance `foaf:name`. For example, when indexing individuals that belong to a hypothetical class `ont:Person` from a Spanish RDF repository that uses `foaf:name` for representing the name of a person, the following query would be performed (prefix definition is omitted):

```
SELECT DISTINCT ?uri ?class ?label
WHERE{
  VALUES ?class { ont:Person } .
  ?uri a ?class .
  {
    ?uri rdfs:label ?label .
    FILTER (lang(?label) = 'es') .
  }
  UNION {
    ?uri foaf:name ?label .
```

```
    FILTER (lang(?label) = 'es') .
  }
}
```

Finally, the result of the query has enough information to add the individuals of the class `ont:Person` to the index.

### 2.3 Named entity recognizer

The named entity recognizer is implemented through a language dependent Stanford pipeline. This pipeline has the same annotators than the pipeline used for text normalization with a new custom annotator at the end, whose objective is to perform the entity recognition.

The implementation of this custom annotator is based on the standard named entity recognizer `TokensRegexNERAnnotator` from Stanford NLP. The original one uses a gazetteer file that contains, for each line, a regular expression associated with a type. Our customization consists in using the index instead of the original gazetteer file. This annotator is configured to use the index fields `normalized label`, `URI` and `class` as textual label representing an element, its identifier and its type, respectively.

By using this pipeline, the input text is segmented in named entities. Each named entity may match more than one entity in the index so each one will have a list of matched entities. For instance, if an ontology contains the classes `ont:CountryCapital`, referring the capital of a country, and `ont:Capital`, referring the money used in business, both annotated with the `rdfs:label` "capital", the entity recognizer would return both classes for the input text "capital".

Table 2 shows the named entities together with their related entities (URI and class) found in the Spanish version of DBpedia for the input "ocupación personas España".

| **Named entity** | **URI** | **Class** |
|---|---|---|
| "ocupación" | ont:occupation | owl:ObjectProperty |
| "personas" | ont:person<br>ont:Person | owl:ObjectProperty<br>owl:Class |
| "España" | res:España<br>res:España<br>res:España | ont:Country<br>ont:PopulatedPlace<br>ont:Place |

Table 2: Example of named entity recognized in the input "ocupación personas España" using the Spanish DBpedia

### 2.4 Tree generator

The next step is to try to connect all the entities extracted in the previous steps by using the ontology. For this purpose a tree that describes the query is built and the goal of this step is to obtain the tree with the minimum height that connects all the entities through the relations in the ontology, checking the domain and the range of these relations.

Firstly, the system tries to remove redundant information in order to get more precise results and to reduce the size of the problem. In this process, for each named entity detected, we check the hierarchy of ontology classes, properties and relations associated with it. The most specific classes are maintained while the general classes are removed. For example, in DBpedia, the resource `res:Napoleon` belongs to the classes `ont:Royalty`, `ont:Person` and `ont:Agent`. In this example, the method would keep `ont:Royalty` since it is the most concrete one. With this action, the system is able to find more specific trees, which are translated into more precise queries. Contrariwise, the method selects the most general property or relations when a hierarchy of properties or relations is detected. For example, if a hypothetical ontology has the properties `personalPhoneNumber` and `workPhoneNumber`, which are sub-properties of `phoneNumber`, the method will maintain `phoneNumber`. We expect that a query performed against a SPARQL endpoint by using a top-level relation or property returns, at least, the union of the results that have all its sub-properties as predicate. That is, a query about `phoneNumber` should include `personalPhoneNumber` and `workPhoneNumber`.

As a named entity can be finally associated with more than one entity, a backtracking process is performed to iterate over all possible combinations. Each one is evaluated by calculating the height of the different trees, which result from computing the shortest path between each entity (playing the role of root element) to the others (acting as leaf nodes) by using breadth first algorithm. Finally, the combination whose evaluation results in a tree with the minimum height is selected.

This process is performed twice using two different strategies to compute the shortest path between the ontology elements. On one hand, paths between root and leaf elements are computed by using the exact classes defined in the range and the domain of the ontology relations.On the other hand, ontology classes that are compatible with the domain and range are taken into account in order to expand the tree during their construction. The compatible classes of a class are defined as the union of the super and subclasses of that class. Finally, if only one strategy has found a tree, that tree is selected. If both strategies have found a tree, that whose height is lower is selected. At equal height, the tree found by the most conservative strategy is chosen. If none of the strategies finds a solution, then the elements are not connected and the method will not be able to generate a SPARQL query.

Following the example of DBpedia commented in Section 2.3, Figure 2 shows the tree obtained from the named entities described in Table 2.

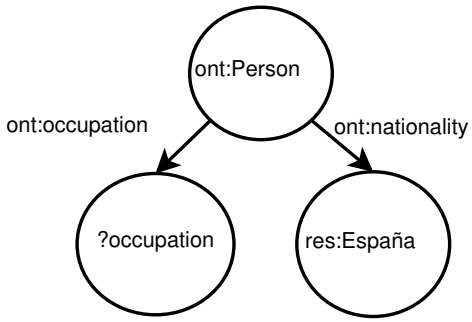


Figure 2: Example of tree construction by using the input "ocupación personas España" with the Spanish DBpedia

## 2.5 SPARQL query generator

If only one named entity is detected in the input, the method generates a query depending on the type of the first entity associated with the named entity. If the entity is a property or a relation, the SPARQL query retrieves all RDF subjects and objects connected by that property or relation. For instance, if only `ont:birthYear` property is detected, the following query is generated:

```
SELECT ?instance ?object
WHERE {
  ?instance ont:birthYear ?object .
}
```

If the entity is a class, the SPARQL query retrieves the individuals that belong to that class. For instance, if only `ont:Writer` class is recognized in the user input, the following query is generated:

```
SELECT ?instance
WHERE {
  ?instance rdf:type ont:Writer .
}
```

If the entity is an individual, firstly a SPARQL query is executed to retrieve all the RDF predicates that link the individual to other entities or primitive types. Then, the returned predicates are used for generating the final query. For example, if the individual `res:Napoleon` is detected in the input, the following query will extract the predicates that are linked to the individual:

```
SELECT DISTINCT ?predicate WHERE {
res:Napoleon ?predicate ?value .
}
```

Then, the final query is built using the predicates. For example, if the returned predicates are `rdf:type`,`ont:deathPlace` and `ont:deathDate`, the following query will be generated:

```
SELECT ?instance
       ?type
       ?deathPlace
       ?deathDate
WHERE {
  VALUES ?instance { res:Napoleon } .
  ?instance rdf:type ?type .
  ?instance ont:deathPlace ?deathPlace .
  ?instance ont:deathDate ?deathDate .
}
```

In the case of more than one named entity detected in the input, the SPARQL query generator transforms the tree of entities into a SPARQL query by applying the following heuristics:

1. The root element of the tree is the element from which the user wants to retrieve information. This element could be a concrete individual or a class. In the latter case the method would return a list of the individuals that belong to that class.
2. The leaf nodes of the tree could be classes, relations or properties in the ontology. In this case they act as the information that the user wants to retrieve

from the root element. On the other hand, a leaf node could be a concrete individual, so playing the role of a filter, imposing that the root element had to be related with the individual.

This kind of trees fits very well to SPARQL queries. Root and leaf elements of the tree are associated with variables that appear in `SELECT` clause. If the root element is a concrete individual, its URI identifier will be stored in the associated variable through a `VALUES` expression. Otherwise, if the root element is a class, the associated variable will contain all individuals that belong to the class through an statement of type `?var rdf:type classURI`. Then, this root variable is connected to the leaf variables according the paths in the tree. Finally, if a leaf element is an individual, their associated variable is set with its URI through a `VALUES` clause, acting as filter. Otherwise, the leaf variable will contain all elements related with the root variable through the relations found in the tree.

According to the example described in Section 2.4, whose tree is shown in Figure 2, the root element is the class `ont:Person` and the leaf nodes are the relation `ont:occupation` and the individual `res:España`, which is reached through the relation `ont:nationality`. Therefore, based on the heuristics, this tree indicates that the user wants to retrieve a list of persons from Spain together with their occupations. The final SPARQL query built from this tree is the following (prefix definition is omitted):

```
SELECT ?instance ?occupation ?Country
WHERE {
?instance rdf:type ont:Person .
?instance ont:occupation ?occupation .
?instance ont:nationality res:España .
VALUES ?Country { res:España } .
}
```

## *3 Results*

A web application (available at `http://sele.inf.um.es/sesssamo-demo/`) has been developed implementing the described method. Three repositories have been configured: the University of Extremadura, the city of Cáceres and both the English and Spanish versions of DBpedia. Next, we describe some examples of queries executed against each repository. All the queries generated a tree of height 2 except one, which generated a query of height 3. The time to build these queries was 435.5 ms on average, with a median of 297 ms. These tests were performed on a laptop with 12 GB of RAM and an Intel Core i5-3337u processor.

### 3.1 University of Extremadura

This university publishes RDF data in Spanish through a SPARQL endpoint (`http://opendata.unex.es/sparql`). The data is organized according to ontouniversidad ontology (`http://opendata.unex.es/def/ontouniversidad.html`)

The following list shows some examples of queries:

- **directores departamento**. List of departments together with their heads.
- **asignaturas isabel cuadrado gordillo**. The subjects from which Isabel Cuadrado Gordillo is teacher of.
- **publicacion isabel cuadrado gordillo**. List of publications of Isabel Cuadrado Gordillo.
- **asignaturas pdi**. List of professors together with their subjects.
- **pdi campus universitario**. Returns a list of researchers together with the university campus that they belong to. This query generates a tree of height 3.

### 3.2 City of Cáceres

The city of Cáceres has a SPARQL endpoint (`http://opendata.caceres.es/sparql`) in which RDF data in Spanish are published according to ontomunicipio ontology (`http://opendata.caceres.es/def/ontomunicipio.html`).

The following queries were tested:

- **cofradias nombre asociacion numero miembros**. List of brotherhoods together with their official name and number of members.
- **monumentos enlace s i g**. List of monuments with their Geographic Information System link.
- **barrios centro**. List of neighborhoods that belong to the center district.
- **hoteles via**. List of hotels and the street name where they are situated.

- **cofradias procesiones**. List of brotherhoods together with the processions they organize.

### 3.3 DBpedia

DBpedia is an encyclopedic knowledge base whose data is extracted from Wikipedia through automatic processes (Auer et al., 2007). Different SPARQL endpoints are available depending on the language, but all conform to an unique ontology, available at `http://downloads.dbpedia.org/2016-04/dbpedia_2016-04.owl`.

Spanish labels were added to the ontology automatically via Google Translator API in those cases where it exists an English but not a Spanish label. Moreover, only elements from the ontology were added to the index. The named entity recognizer uses the index for identifying these elements (concepts, relations and properties) and the DBpedia Spotlight API (Mendes et al., 2011) for detecting concrete individuals.

Some examples of queries using Spanish and English versions of DBpedia are shown below.

#### 3.3.1 Spanish version

- **progenitores Juan Carlos I**. Parents of Juan Carlos I of Spain (Juan de Borbón, María de las Mercedes de Borbón-Dos Sicilias).
- **ocupación personas España**. List of Spanish persons together with their occupation.
- **superficie Atlanta**. The area of Atlanta (3.412E8).
- **esposa John Lennon**. The spouse of John Lennon (Yoko Ono).

#### 3.3.2 English version

- **The Beatles former band members**. The name of the persons who formed The Beatles (George Harrison, John Lennon, Ringo Starr, Paul McCartney).
- **Einstein spouse**. The spouse of Albert Einstein (Mileva Marić, Elsa Löwenthal).
- **Stephen Hawking doctoral advisor**. The advisor of Stephen Hawking's PhD thesis (Dennis William Sciama).
- **mean temperature k Mars**. The mean temperature of Mars in Kelvin degrees (210.0, 210.15).

## *4 Discussion*

In this paper we have presented a method whose goal is to facilitate human users the exploitation of semantic repositories, among whose main applications we can identify semantic search and question answering systems. State of the art solutions in this field usually require a complete, well written question in natural language. This permits a more exhaustive analysis of the query, and shows good results, but this is not the usual interaction way between users and search engines. Our work assumes that we should use the same type of interaction when exploiting semantic repositories, and that this is an essential feature for the success of semantic web technologies.

Our work is in line with the method shown in (Tran et al., 2007), which proposes a searcher based on keywords. This method also uses trees to describe queries. As in our work, and based on the locality principle, the tree with the lowest height is consider more likely to contain the answer the user wants to know. Nonetheless, in this work only one entity is associated with each keyword before computing the tree. This may cause problems if a keyword is ambiguous and the selected entity is not what the user had in mind. Our method takes this into account and, as previously mentioned in Section 2.4, it uses all possible entities matching a keyword in order to compute the tree. Finally, based on locality, the correct entity will be selected. Our method renovates the technological approach developed by our research group in (Valencia-García et al., 2011), since now the query design is not driven by the ontology but by the user input.

One limitation of our current implementation is due to ambiguity in cases where two concepts are linked by more than one relation. That situation permits to generate different trees with the same height, and we are currently executing only one of them. For example, in DBpedia, the input "writers Spain" shows no results because the system links `ont:writer` and `res:Spain` through `ont:livingPlace` instead of `ont:nationality`. As future work we will explore the execution of both queries

or offering the users the two interpretations of the queries and asking them to select the desired one. This way of solving ambiguity has already been used (Damljanovic, Agatonovic, and Cunningham, 2011; Lehmann and Bühmann, 2011). Another limitation is the use of synonyms that are not indexed. We plan to enrich the index by using tools like WordNet (Miller, 1995) to provide these synonyms. Once these limitations have been overcome we hope to obtain good results by using the Question Answering over Linked Data (QALD) evaluation system (Lopez et al., 2013).

## *Acknowledgements*

This work has been funded by the Spanish Ministry of Economy, Industry and Competitiveness, the European Regional Development Fund (ERDF) Programme and the Fundación Séneca through grants TIN2014-53749-C2-2-R and 19371/PI/14.

## *References*

Auer, S., C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*. Springer, pages 722–735.

Bechhofer, S. 2009. Owl: Web ontology language. In *Encyclopedia of database systems*. Springer, pages 2008–2009.

Berners-Lee, T., J. Hendler, and O. Lassila. 2001. The semantic web. *Scientific american*, 284(5):34–43.

Bizer, C., T. Heath, K. Idehen, and T. Berners-Lee. 2008. Linked data on the web (ldow2008). In *Proceedings of the 17th international conference on World Wide Web*, pages 1265–1266. ACM.

Consortium, U. 2016. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 45(D1):D158–D169.

Damljanovic, D., M. Agatonovic, and H. Cunningham. 2011. Freya: An interactive way of querying linked data using natural language. In *Extended Semantic Web Conference*, pages 125–138. Springer.

Klyne, G. and J. J. Carroll. 2006. Resource description framework (RDF): Concepts and abstract syntax. Technical report, W3C.

Lehmann, J. and L. Bühmann. 2011. Autosparql: Let users query your knowledge base. In *Extended Semantic Web Conference*, pages 63–79. Springer.

Lopez, V., C. Unger, P. Cimiano, and E. Motta. 2013. Evaluating question answering over linked data. *Web Semantics Science Services And Agents On The World Wide Web*, 21:3–13.

Manning, C., M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Mendes, P. N., M. Jakob, A. García-Silva, and C. Bizer. 2011. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, pages 1–8. ACM.

Miller, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Swartz, A. 2002. Musicbrainz: A semantic web service. *IEEE Intelligent Systems*, 17(1):76–77.

Tran, T., P. Cimiano, S. Rudolph, and R. Studer. 2007. Ontology-based interpretation of keywords for semantic search. In *The Semantic Web*. Springer, pages 523–536.

Valencia-García, R., F. García-Sánchez, D. Castellanos-Nieves, et al. 2011. Owlpath: An owl ontology-guided query editor. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 41(1):121–136.

# *Análisis de textos médicos*

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 49-56 recibido 04-04-2018 revisado 01-05-2018 aceptado 04-05-2018

# Construcción de recursos terminológicos médicos para el español: el sistema de extracción de términos CUTEXT y los repositorios de términos biomédicos

## *Construction of medical terminological resources for Spanish: the CUTEXT term extraction system and biomedical term repositories*

**Jesús Santamaría**[1,2]
**Martin Krallinger**[1,2]
[1]CNIO (Centro Nacional de Investigaciones Oncológicas)
[2]BSC (Barcelona Supercomputing Center)
jsantamaria@cnio.es
krallinger.martin@gmail.com

**Resumen:** El uso frecuente de términos médicos motivó la construcción de grandes recursos terminológicos para el inglés, como el Unified Medical Language System (UMLS) o las ontologies Open Biological and Biomedical Ontology (OBO). La construcción exclusivamente manual de recursos terminológicos es en sí misma muy valiosa, pero constituye (1) un proceso laborioso que requiere mucho tiempo, (2) no garantiza que los conceptos o términos incluidos se 'alineen' realmente con el lenguaje médico y los términos que se usan en los documentos clínicos escritos por los profesionales de la salud y (3) requiere actualización constante y revisión debido a los cambios y la aparición de nuevos conceptos biomédicos. En este artículo presentamos una herramienta de extracción de términos médicos multilingüe, llamada CUTEXT (CValue Utilizado para Extraer Términos), un recurso promovido por el Plan de Impulso de las Tecnologías del Lenguaje (Villegas et al., 2017), disponible en: `https://github.com/Med-TL/Plan-TL/tree/master/CUTEXT`
**Palabras clave:** CValue, términos, extracción automática de términos

**Abstract:** The heavy use of medical terms motivated the construction of large terminological resources for English, such as the Unified Medical Language System (UMLS) or the Open Biological and Biomedical Ontology (OBO) ontologies. Purely manual construction of terminological resources is by itself very valuable, but constitutes (1) a highly time-consuming process, (2) it does not guarantee that included concepts or terms do actually align with the medical language and terms as they are being used in clinical documents by healthcare professionals and (3) requires constant update and revision due to changes and emergence of new biomedical concepts over time. In this paper we present a multilingual medical term extraction tool, called CUTEXT (Cvalue Used To Extract Terms), a resource promoted by the Spanish National Plan for the Advancement of Language Technology (Villegas et al., 2017), available at: `https://github.com/Med-TL/Plan-TL/tree/master/CUTEXT`
**Keywords:** CValue, terms, automatic terms extraction

## *1 Introducción*

Una característica común de los textos biomédicos y clínicos es el uso excepcionalmente frecuente de términos técnicos específicos de dominio. Esta característica es compartida por la mayoría de los documentos médicos, independientemente de si están escritos en inglés, español u otros idiomas. Además, en el caso del dominio médico, existe una considerable brecha de recursos terminológicos para todos los demás idiomas en comparación con el inglés. La detección y el procesamiento eficientes de términos técnicos médicos es clave, no solo para enriquecer o incluso ayudar en la construcción automática de recursos terminológicos, sino que también constituye un elemento clave para otras aplicaciones de tecnología del lenguaje médico, incluyendo traducción automática, extracción de información, generación de resúme-

ISSN 1135-5948. DOI 10.26342/2018-61-5

nes y particularmente sistemas automáticos de codificación de documentos clínicos.

Desde la perspectiva de la extracción de términos, la tarea consiste en identificar dos rasgos principales (Kageura y Umino, 1996): (1) Unicidad (*unithood*): grado de cohesión o estabilidad de las palabras de una locución. (2) Termicidad (*termhood*): grado de especificidad del término con respecto a una disciplina concreta.

Por tanto, si deseamos comunicarnos sin confusión ni malentendidos, debemos ponernos de acuerdo sobre qué términos utilizaremos para representar los conceptos y cómo se deben traducir estos términos a diferentes idiomas (Foo, 2012).

La extracción manual de términos es una tarea larga, repetitiva, y tediosa, por ello, corre el riesgo de ser poco sistemática y subjetiva. Además, es muy costosa en términos económicos y está limitada por la información disponible. Por ello, la extracción automática de términos es una tarea crucial en el procesamiento del lenguaje natural, ya que permite que todo el proceso se lleve a cabo de forma mucho más ágil, eficiente, y económica, tanto en tiempo como en dinero.

La extracción automática de términos es una tarea relevante que puede ser útil en una amplia gama de tareas, como el aprendizaje ontológico, la traducción asistida y automática, la construcción de tesauros, la clasificación, la indexación, la recuperación de información, así como en la minería de textos y en la extracción de resúmenes.

Sin embargo, el proceso de extracción automática de términos no es una tarea trivial. El cambio constante en la terminología hace necesario que las herramientas sean capaces de detectar dichos términos nuevos, así como sus posibles variaciones. Las tareas de extracción suelen ser específicas, y por tanto, deben adaptarse a los requerimientos y particularidades propias de cada una de ellas. Según Ananiadou y Nenadić (2006) se pueden distinguir cinco variaciones terminológicas: ortográfica, morfológica, léxica, estructural, acrónimos y abreviaturas. A las que podemos añadir también las producidas por sinonimia y homonimia. Por último, existe una falta de convenciones firmes en la nomenclatura. Se han creado directrices pero no se imponen restricciones, así, junto con los términos denominados "bien formados" existen otros ad-hoc, que son problemáticos para los sistemas automáticos de identificación de términos.

Para hacer frente a estos problemas, los sistemas de extracción automática de términos se suelen clasificar en cuatro grandes grupos no excluyentes: (1) Sistemas basados en **características internas**: suelen atender a la ortografía, como mayúsculas, uso de dígitos, caracteres griegos, etc. (2) Sistemas que aprovechan **pistas morfológicas**: atienden sobre todo a afijos específicos y formantes cultos (principalmente griegos y latinos). (3) Sistemas que aprovechan la información procedente del **análisis sintáctico**: atienden a la estructura gramatical, para extraer términos procedentes de los sintagmas nominales, verbales, y preposicionales. (4) Sistemas basados en **medidas estadísticas** para promover candidatos a términos: basados en frecuencias, *log-likelihood*, *logDice*, *l-value*, *c-value*, etc. Como veremos más adelante, CUTEXT aprovecha tanto la información lingüística (3), como la estadística (4). CUTEXT permite el reconocimiento automático de términos técnicos, de una o varias palabras, a partir de documentos médicos, admitiendo varios formatos de entrada. Este sistema permite a los usuarios sin conocimientos técnicos identificar fácilmente términos médicos detectados en grandes corpus clínicos, y así facilitar la construcción de diccionarios clínicos, índices o glosarios médicos, con un énfasis particular en aplicaciones de reconocimiento de conceptos clínicos. Este artículo resume las principales características de CUTEXT, así como los resultados que se obtienen al aplicarlo a diferentes corpus, incluidos textos biomédicos en inglés, literatura médica española y textos clínicos en español.

En el apartado 2, se revisa el estado del arte, en el 3 se describen las características principales de CUTEXT, en el 4 se muestran los corpus, la colección que se ha utilizado durante el uso real con CUTEXT, y, por último, en el apartado 5 se resumen las conclusiones obtenidas así como las líneas de trabajo futuro.

## *2 Trabajo Relacionado*

Las aproximaciones utilizadas en el estado del arte para la extracción automática de términos se pueden dividir, siguiendo a Krauthammer y Nenadic (2004), en cuatro tipos:

1. **Basadas en diccionario**: utilizan listas de palabras, de *stop-words* (sin con-

tenido semántico), ontologías, glosarios, y tesauros del dominio. Se utilizan como "filtro lingüístico" para eliminar palabras y reconocer términos. Tiene la ventaja de ser simple y eficiente, pero suele ser incompleta, además de no estar disponible en todos los dominios e idiomas.

2. **Basadas en reglas**: En este enfoque se emplean patrones y conocimiento gramatical para la detección de términos. Es uno de los métodos más empleados desde los años 90, sin embargo, posee el mismo inconveniente que el anterior, a saber, no está extendido igualmente en todos los dominios e idiomas.

3. **Basadas en estadística y aprendizaje automático**: Las técnicas estadísticas tratan de determinar lo característico de una palabra (o lema) en un corpus específico con respecto a su frecuencia en un corpus general. Es decir, intentan saber qué términos son sobreutilizados o infrautilizados en el corpus utilizado en comparación con su frecuencia en un corpus de referencia.

4. **Híbridas**: Combinan dos o más de las aproximaciones anteriores. CUTEXT, como veremos, es un enfoque híbrido que combina la aproximación lingüística con la estadística.

En los últimos años se han desarrollado varias herramientas de extracción automática de términos, sin embargo, la mayoría de ellas son dependientes del idioma: en portugués - ExATOlp (Lopes et al., 2009), en español-vasco - Elexbi (Gurrutxaga et al., 2006), en español-alemán - Autoterm (Haller, 2008), en árabe (Boulaknadel, Daille, y Aboutajdine, 2008), en esloveno e inglés - Luiz (Vintar, 2010), en inglés e italiano - KX (Pianta y Tonelli, 2010), o en inglés y alemán (Ramm et al., 2018).

Otros trabajos relacionados se pueden encontrar en (Koza Orellana, 2015), así como en (Barrón-Cedeno et al., 2009).

Algunas herramientas se han adaptado a dominios especificos como por ejemplo, TermExtractor (Sclano y Velardi, 2007), TerMine (Frantzi, Ananiadou, y Mima, 2000) o BioYaTeA (Golik et al., 2013). La herramienta TermSuite, se desarrolló durante el proyecto europeo *Terminology Extraction, Translation Tools and Comparable Corpora* (TTC). Este proyecto se centró en la adquisición automática o semiautomática de terminologías alineadas bilingues para la traducción asistida y automática.

Por último, citamos algunas aplicaciones, en español, en las que se ha utilizado la extracción automática de términos: En Castro et al. (2010) utilizan la extracción automática de términos para la detección de conceptos en notas clínicas y su posterior asociación, o mapeo, a la ontología *Systematized Nomenclature of Medicine — Clinical Terms* (SNOMED–CT). Vivaldi y Rodríguez (2010) utilizan un sistema de extracción de términos en un corpus biomédico, de tal forma que, una vez encontrado un candidato a término, se intenta encontrar una página en la Wikipedia que se corresponda con dicho candidato, después se encuentran todas las categorías de la Wikipedia asociadas a dicha página, y por último, se explora la Wikipedia siguiendo recursivamente todos los enlaces de categorías encontrados, para expandir y enriquecer la frontera del dominio. Por último, Moreno-Sandoval y Campillos-Llanos (2013) utilizan un sistema de extracción de términos para elaborar un corpus compuesto por textos biomédicos en español, árabe, y japonés.

## *3 Descripción de CUTEXT*

Las características principales de CUTEXT son las siguientes:

- Está implementado en java, por lo que es multiplataforma. Se ha probado bajo Windows y Linux.
- Es multilingüe: Se ha probado en inglés, castellano, catalán, y gallego. Se puede adaptar fácilmente a otros idiomas con tan solo cambiar el fichero de texto de configuración de etiquetas léxicas.
- Puede utilizar como etiquetador a TreeTagger[1] o a GeniaTagger[2].
- Los documentos que admite pueden ser en texto plano, o en pdf.
- Permite tanto interfaz gráfica como por consola (modo texto).

[1] `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`. Permite etiquetar textos en más de 23 idiomas.

[2] `http://www.nactem.ac.uk/GENIA/tagger/`. A diferencia de Treetagger, sólo etiqueta textos en inglés, pero a cambio tiene la ventaja de haber sido adaptado específicamente para textos biomédicos

- Permite seleccionar el idioma, el etiquetador, los umbrales de frecuencia y de c-value, y la entrada del documento o documentos.
- La salida se proporciona en texto plano, en formato JSON, y en BioC[3] (Comeau et al., 2013). Se incluyen los tiempos parciales y el tiempo total.

CUTEXT tomó como punto de partida un extractor de alto impacto denominado TerMine (Frantzi, Ananiadou, y Mima, 2000), que permite extraer términos en textos escritos en inglés. Como se ha mencionado anteriormente, y siguiendo a TerMine, contiene dos filtros: uno lingüístico y otro estadístico. El filtro lingüístico se compone de expresiones regulares, y una lista de *stop–words*, y es dependiente del idioma. Por tanto, para añadir un nuevo idioma a CUTEXT, también habría que proporcionarle las expresiones regulares, (el filtro lingüístico), para dicho idioma. Para el inglés el filtro está dividido en tres: El primero tiene en cuenta sólo *nombres*, el segundo *nombres* y *adjetivos*, y el tercero *nombres*, *adjetivos* y *preposiciones*. Por otro lado, para castellano, catalán, y gallego es un poco más sofisticado. Está dividido en dos subfiltros, denominados *cerrado* y *abierto*: El filtro cerrado detecta tres patrones: (1) *nombre* seguido de *adjetivo*, (2) *nombre* seguido de *de* seguido de *adjetivo*, y (3) *nombre*. El filtro abierto está basado en cinco expresiones regulares que contienen *nombres*, *palabras extranjeras*, preposición *de*, *acrónimos*, y *adjetivos*. Al ser dependiente del idioma y del ámbito, se puede cambiar. Hemos visto, que los *verbos* no están incluidos. Sin embargo, para ciertas tareas sería deseable incluirlos. Por ejemplo, en un corpus en el que es más frecuente la expresión *"se descarta cáncer"* que *"cáncer descartado"*, los filtros actuales dejarían pasar el segundo pero no el primero, y sería deseable que ambos pasasen el filtro, por lo que se podría incluir una nueva expresión regular que tuviese en cuenta este *tipo* de verbos. Por otra parte, el filtro estadístico es independiente del idioma, y combina las siguientes tres métricas:
1. La frecuencia y longitud del término candidato.
2. La frecuencia del término candidato como parte de otros términos candidatos más largos.
3. El número de estos términos candidatos más largos.
En concreto, el $C-value$ que se asigna a un término candidato $a$, viene dado por la siguiente expresión:

$$C{-}value(a) = log_2|a|(f(a) - \frac{1}{\#(T_a)} \sum_{b \in T_a} f(b))$$

Donde,
$a$ es el término candidato
$f(a)$ es la frecuencia del término candidato
$T_a$ es el conjunto de términos candidatos que contienen a $a$ (*candidatos mayores*)
$\#(T_a)$ es el número de términos de los *candidatos mayores*
$\sum_{b \in T_a} f(b)$ es la frecuencia total en la que $a$ aparece en el conjunto de los *candidatos mayores*

El diagrama de CUTEXT se muestra en la figura 1.

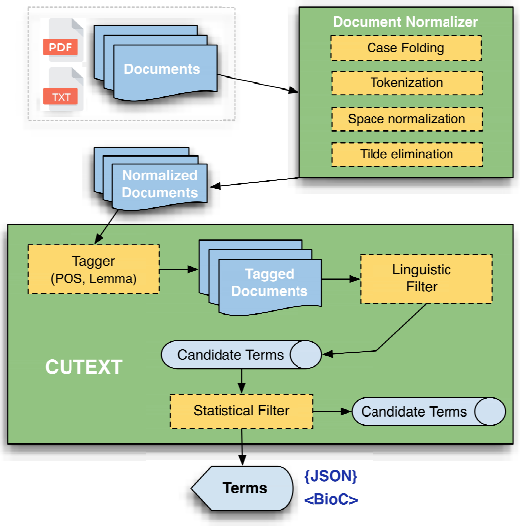


Figura 1: Diagrama de CUTEXT

El proceso denominado *Normalizer* se encarga de normalizar el texto procedente de los diversos documentos, y está separado de CUTEXT, es decir, no está incluido como parte íntegra en él. Básicamente lo que hace este proceso es: (1) convierte todo el documento (o documentos) a minúsculas, (2) separa el texto en tokens[4], (3) asigna un espacio entre

[3] Es un formato interoperable para textos biomédicos, basado en XML. Utilizado principalmente para el intercambio y el almacenamiento de datos de forma sencilla.

[4] Cada signo de puntuación lo considera un token, que se separa del resto. Así si, como es habitual, hay una palabra seguida de una coma – "por tanto," – la palabra y la coma se separan en dos tokens – "por

los tokens, y, por último, (4) elimina las tildes (si existen). Los documentos, ahora normalizados, se entregan a CUTEXT. El etiquetador (*Tagger*), se encarga de asignar a cada palabra su etiqueta léxica y su lema. Esta información es utilizada por el filtro lingüístico (*Linguistic Filter*) para obtener los términos candidatos, a partir de las reglas asignadas a priori. Los términos candidatos, sin ningún tipo de estadística salvo su frecuencia, se almacenan (*Terms*), y se entregan al filtro estadístico (*Statistical Filter*) que asigna el denominado C-Value. Estos términos, son almacenados de nuevo internamente (*Terms*), y también se muestran en los diferentes formatos (*Terms*) al usuario final.

### 3.1 Modo Texto – Opciones

En modo texto (por línea de comandos), CUTEXT ofrece más opciones que en modo gráfico. En concreto, en la tabla 1, se muestran las diferentes opciones que permite, así como el tipo para cada opción, y su valor por defecto.

| Nombre | Tipo | Por defecto |
|---|---|---|
| -help | no aplica | no aplica |
| -displayon | boolean | true |
| -postagger | string | TreeTagger |
| -language | string | Spanish |
| -frecT | integer | 0 |
| -cvalueT | double | 0.0 |
| -bioc | boolean | false |
| -convert | boolean | true |
| -withoutcvalue | boolean | false |
| -incremental | boolean | false |

Tabla 1: Opciones de CUTEXT en modo texto.

Todas las opciones son autoexplicativas, salvo, quizá, *-displayon*, *-convert*, *-withoutcvalue*, e *-incremental*.

***-displayon:*** Se utiliza para mostrar por la salida estándar el proceso de ejecución, así como el tiempo que tarda en cada una de las fases. Por defecto, es *true*, es decir que sí se mostrará. Si se pone a false no presentará nada por la salida estándar.

***-convert:*** Convierte a minúsculas el texto de entrada. Conviene tener en cuenta que en general, TreeTagger etiqueta las palabras escritas en mayúsculas, como nombres propios, por lo que si se selecciona este etiquetador (con la opción *-postagger*), conviene también poner este parámetro a true (que es su valor por defecto, al igual que el valor por defecto de *-postagger* es TreeTagger).

***-withoutcvalue:*** Si se pone a *true* entonces CUTEXT ejecutará solamente el filtro lingüístico. Esto es útil en las aplicaciones donde es importante la rapidez, y no tanto el valor de c-value, (sólo con la frecuencia nos es suficiente), ya que CUTEXT, lógicamente, tardará menos en obtener los términos.

***-incremental:*** Si se pone a *true* entonces CUTEXT ejecuta cada línea del fichero de entrada por separado. Es decir, para cada línea del fichero de entrada se ejecutan todas y cada una de las fases, tratándose, por tanto, cada línea como si fuese el corpus completo.

La versión más rápida de CUTEXT se da cuando los parámetros *-incremental* y *-withoutcvalue* están ambos a *true*.

## *4 Corpus Generados*

CUTEXT se ha probado en diversos corpus, tanto *de juguete* como reales. Ha sido, como es lógico, con los corpus reales donde nos hemos dado cuenta de que para hacerlo práctico teníamos que hacerlo más eficiente, acelerando su ejecución. A partir de la salida generada por CUTEXT, se han obtenido diferentes corpus terminológicos. En las siguientes subsecciones explicaremos los corpus utilizados así como los obtenidos por CUTEXT. En concreto: (1) corpus Genia, (2) corpus de la tarea Biomedical Abbreviation Recognition and Resolution (BARR), (3) corpus de la Base de datos para la Investigación Farmacoepidemiológica en Atención Primaria (BIFAP).

### 4.1 Corpus Genia

El corpus Genia[5], está formado por resúmenes en inglés anotados, tomados de la base de datos MEDLINE de la Biblioteca Nacional de Medicina. Están anotadas un subconjunto de las sustancias y de las estructuras subcelulares de proteínas, basadas en un modelo de datos (denominada ontología GENIA) del dominio biológico, en formato XML (GPML). La version 3.0x consta de 2000 resúmenes. Los resúmenes básicos se seleccionan de los

tanto , " –. Esto evita que el etiquetador considere como token palabras seguidas por comas, como "tanto,".

[5] `http://www.geniaproject.org/`

resultados de búsqueda con las palabras clave (términos MeSH) *Human, Blood Cells and Transcription Factors*. El corpus está anotado con seis niveles de información lingüística y semántica: (1) Anotación léxica, (2) Constituyentes (anotación sintáctica), (3) Anotación de conceptos Genia, (4) Anotación de eventos, (5) Anotación de relaciones, (6) Anotación de correferencia.

En concreto, la anotación de términos cubre la identificación de entidades biológicas físicas, así como otros términos importantes. Para este corpus, CUTEXT generó un recurso terminológico de 38.903 términos. Los términos con los valores más altos de c-value se corresponden con términos biológicos, por lo que el recurso obtenido por CUTEXT puede valer en aquellas tareas en las que se necesiten dichos términos, como validación, mapeo, etc. Como se ha explicado anteriormente, el corpus Genia contiene anotación de conceptos Genia, por lo que se utilizó esta para su análisis comparativo. La medida $F_1$, obtenida por CUTEXT utilizando a TreeTagger como etiquetador léxico fue de un $F_1 = 52,6\,\%$, mientras que al utilizar como etiquetador a GeniaTagger obtuvo una medida ligeramente superior $F_1 = 53,8\,\%$. Esto es lógico, ya que GeniaTagger es un etiquetador de dominio biomédico. Esta medida es bastante superior a la obtenida por la versión web de TerMine[6], que fue de tan sólo de $F_1 = 39,7\,\%$.

## 4.2 Corpus BARR

BARR (Villegas et al., 2018) fue una tarea que se propuso dentro de IBEREVAL 2017[7]. Se trataba de reconocer abreviaturas y su expansión, así como relacionar ambas. Fue particularmente interesante, ya que algunos resúmenes se transcribieron manualmente, algo que se asemeja a las características de preprocesamiento que se encuentran en documentos clínicos. Para llevarla a cabo, se liberaron el corpus BARR anotado manualmente (Gold Standard), y una colección de documentos de resúmenes de artículos médicos escritos en español (la mayor colección unificada existente de la que tenemos noticia), distribuida a través de un acuerdo especial con el editor Elsevier. En concreto, el número total de tokens en los resúmenes (*abstracts*) es de 36.981.968, y el número total de tokens en los títulos de los *abstracts* es de 2.359.516.

Para este corpus, CUTEXT obtuvo 975.963 términos, cubriendo la mayoría de las abreviaturas y de sus formas largas[8]. Por tanto, el corpus terminológico obtenido, que se corresponde con términos biomédicos que incluyen abreviaturas, puede ser útil en tareas no sólo biomédicas sino propias de reconocimiento y extracción de abreviaturas.

## 4.3 Corpus BIFAP

La base de datos de BIFAP[9] está informatizada con registros médicos de Atención Primaria (AP) para la realización de estudios farmacoepidemiológicos, perteneciente a la Agencia Española de Medicamentos y Productos Sanitarios[10] (AEMPS), y cuenta con la colaboración de Comunidades Autónomas y el apoyo de las principales sociedades científicas implicadas. BIFAP incluye la información registrada por 5.752 médicos de familia y pediatras de AP del Sistema Nacional de Salud, integrando información de 7.890.485 historias clínicas anonimizadas. El 27 de marzo de 2015, la AEMPS pone a disposición de investigadores del ámbito público la base de datos BIFAP[11], para la investigación con medicamentos. Es en este ámbito en el que se ha elaborado un acuerdo de colaboración entre BIFAP y el Centro Nacional de Investigaciones Oncológicas (CNIO) dentro del marco del Plan de Impulso de las Tecnologías del Lenguaje (PlanTL).

Para este corpus, CUTEXT ha obtenido 1.440.306 términos. El corpus generado, no sólo es válido para la tarea en la que hemos colaborado[12], sino también en aquellas otras relacionadas con casos clínicos, ya que los términos extraídos pertenecen a este ámbito. Los términos extraídos, no se han podido evaluar, ya que BIFAP no tiene ningún gold-standard a este respecto. Sin embargo, sí los hemos examinado manualmente, encon-

[6] `http://www.nactem.ac.uk/software/termine/`

[7] `http://cabrillo.lsi.uned.es/nlp/IberEval-2017/index.php`

[8] No incluimos aquí la medida $F$, ya que no tiene sentido, porque CUTEXT extrae todo tipo de términos, no solamente abreviaturas y formas largas, y sin embargo el gold-standard está compuesto sólo por aquéllas.

[9] `http://www.bifap.org/`

[10] `https://www.aemps.gob.es/`

[11] `https://www.aemps.gob.es/informa/notasInformativas/laAEMPS/2015/docs/NI-AEMPS_03-2015-jornada-BIFAP-marzo-2015.pdf`

[12] Básicamente, consiste en una tarea de mapeo de un literal nuevo, introducido por un médico, frente a los literales almacenados en su base de datos.

trando que los términos con un alto valor de c-value (en general por encima de 50.0) son términos específicos del dominio (como por ejemplo, *adenocarcinoma de pulmón*, con un c-value de 405.839), mientras que aquellos con un c-value bajo, suelen ser más genéricos o suelen estar mal escritos (como por ejemplo, *ansiedad gerneralizada*).

## 5 Conclusiones y Trabajo Futuro

En este artículo se ha mostrado que la extracción automática de términos es una tarea crucial en el ámbito del procesamiento del lenguaje natural. Se ha puesto de relieve que, actualmente, los principales extractores automáticos de términos son dependientes del idioma y de la plataforma. Por todo ello, hemos presentado una herramienta multilingüe y multiplataforma, denominada CUTEXT, que permite extraer automáticamente los términos de un corpus, y asignarle un valor (denominado c-value) a cada uno de ellos, que determina su fiabilidad.

Hemos mostrado tres corpus de distinto ámbito generados por CUTEXT, que se pueden utilizar en tareas muy diversas: (a) Corpus compuesto por términos biológicos, (b) corpus compuesto por términos biomédicos que incluyen abreviaturas y sus formas largas, (c) corpus terminológico de casos clínicos.

Resumimos, a continuación, los puntos principales, y más relevantes de CUTEXT, no vistos en otros extractores automáticos:

1. Es multiplataforma, se ha testeado bajo diferentes sistemas operativos (Windows y Linux), y dispone de una interfaz gráfica y textual altamente configurable.

2. Es un sistema abiertocapaz de generar recursos para medicina, a partir de grandes corpus heterogéneos.

3. Es multilingüe: procesa textos biomédicos en castellano, inglés, catalán, y gallego, pero se pueden añadir idiomas de una forma sencilla.

4. Se ha comprobado su utilidad en aplicaciones reales, como la desarrollada para BIFAP.

5. Admite textos en diferentes formatos, y es capaz de generar una salida en tres tipos de formatos diferentes: plano, JSON, y BioC.

Debido a que la principal utilidad de un extractor de términos es el mapeo[13], como primera línea de trabajo futuro seguiremos a Krauthammer y Nenadic (2004), que determinan 3 etapas secuenciales para su realización:
**(1) Reconocimiento del término**: Permite diferenciar entre términos y no términos. Esta es la salida que proporciona CUTEXT.
**(2) Clasificación del término**: Consiste en asignar los términos al dominio específico. Es decir, quedarse sólo con los términos pertenecientes al dominio. El objetivo consiste en medir el grado de distintividad de un término en un corpus especializado en contraste con su frecuencia en un corpus general. Las métricas más empleadas son *log-likelihood ratio test*, y *logDice*.
**(3) Emparejamiento del término**: Vincula los términos con conceptos bien definidos de fuentes de datos referentes, como vocabularios controlados o bases de datos.

También tenemos pensado utilizar CUTEXT para procesar nuevos corpus de textos médicos bilingües (por ejemplo, el recurso denominado MeSpEN (Villegas et al., 2018)).

## Agradecimientos

El presente trabajo fue realizado bajo la financiación de la Encomienda MINETAD-CNIO/OTG Sanidad Plan TL y el proyecto H2020 OpenMinted (654021).

## Bibliografía

Ananiadou, S. y G. Nenadić. 2006. Automatic terminology management in biomedicine. En S. Ananiadou y J. McNaught, editores, *Text Mining for Biology and Biomedicine*. Artech House, Inc., páginas 67–98.

Barrón-Cedeno, A., G. Sierra, P. Drouin, y S. Ananiadou. 2009. An improved automatic term recognition method for spanish. En *International Conference on Intelligent Text Processing and Computational Linguistics*, páginas 125–136. Springer.

Boulaknadel, S., B. Daille, y D. Aboutajdine. 2008. A multi-word term extraction program for arabic language. 01.

[13]Denominado en inglés *mapping*. Consiste en vincular los términos con conceptos bien definidos de fuentes de datos referentes, como vocabularios controlados o bases de datos.

Castro, E., A. Iglesias, P. Martinez, y L. Castaño. 2010. Automatic identification of biomedical concepts in spanish-language unstructured clinical texts. páginas 751–757, 01.

Comeau, D. C., R. Islamaj Doğan, P. Ciccarese, K. B. Cohen, M. Krallinger, F. Leitner, Z. Lu, Y. Peng, F. Rinaldi, M. Torii, y others. 2013. Bioc: a minimalist approach to interoperability for biomedical text processing. *Database*, 2013.

Foo, J. 2012. Computational terminology: Exploring bilingual and monolingual term extraction. 04.

Frantzi, K., S. Ananiadou, y H. Mima. 2000. Automatic recognition of multi-word terms. *International Journal of Digital Libraries*, 3(2):117–132.

Golik, W., R. Bossy, Z. Ratkovic, y C. Nedellec. 2013. Improving term extraction with linguistic analysis in the biomedical domain. *Research in Computing Science*, 70:157–172.

Gurrutxaga, A., X. Saralegi, S. Ugartetxea, y I. Alegria. 2006. Elexbi, a basic tool for bilingual term extraction from spanish-basque parallel corpora.

Haller, J. 2008. Autoterm : Term candidate extraction for technical documentation (spanish/german).

Kageura, K. y B. Umino. 1996. Methods of automatic term recognition - a review. *Terminology. Amsterdam, 1996.*, 3(2):259–289.

Koza Orellana, W. 2015. Proposal for automatic extraction of medical term candidates with linguistic information processing description and evaluation of results. *Alfa: Revista de Linguística (São José do Rio Preto)*, 59(1):113–128.

Krauthammer, M. y G. Nenadic. 2004. Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37(6):512 – 526. Named Entity Recognition in Biomedicine.

Lopes, L., P. Fern, R. Vieira, y G. Fedrizzi. 2009. Exatolp – an automatic tool for term extraction from portuguese language corpora.

Moreno-Sandoval, A. y L. Campillos-Llanos. 2013. Design and annotation of multimedica – a multilingual text corpus of the biomedical domain. *Procedia - Social and Behavioral Sciences*, 95:33 – 39. Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013).

Pianta, E. y S. Tonelli. 2010. Kx: A flexible system for keyphrase extraction. páginas 170–173, 01.

Ramm, A., U. Heid, B. Weissbach, C. Loth, y I. Mingers. 2018. Adapting and evaluating a generic term extraction tool. 03.

Sclano, F. y P. Velardi. 2007. Termextractor: a web application to learn the shared terminology of emergent web communities. páginas 287–290, 01.

Villegas, M., S. de la Peña, A. Intxaurrondo, J. Santamaria, y M. Krallinger. 2017. Esfuerzos para fomentar la minería de textos en biomedicina más allá del inglés: el plan estratégico nacional español para las tecnologías del lenguaje. *Procesamiento del Lenguaje Natural*, 59:141–144.

Villegas, M., A. Intxaurrondo, A. Gonzalez, , M. Marimon, y M. Krallinger. 2018. The mespen resource for english-spanish medical machine translation and terminologies: Census of parallel corpora, glossaries and term translations. En *Proceedings of the LREC 2018 Workshop "MultilingualBIO: Multilingual Biomedical Text Processing*, páginas 32–39.

Vintar, p. 2010. Bilingual term recognition revisited the bag-of-equivalents term alignment approach and its evaluation. 16:141–158, 12.

Vivaldi, J. y H. Rodríguez. 2010. Using wikipedia for term extraction in the biomedical domain: first experiences. *Procesamiento del Lenguaje Natural*, 45(0):251–254.

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 57-64 recibido 04-04-2018 revisado 02-05-2018 aceptado 04-05-2018

# Improving the accessibility of biomedical texts by semantic enrichment and definition expansion

## *Mejora de la accesibilidad de textos biomédicos mediante enriquecimiento semántico y expansión de definiciones*

**Pablo Accuosto and Horacio Saggion**
Large-Scale Text Understanding Systems Lab
TALN Research Group, DTIC
Universitat Pompeu Fabra
C/Tànger 122-140, 08018 Barcelona, Spain
{*name.surname*}@upf.edu

**Abstract:** We present work aimed at facilitating the comprehensibility of health-related English-Spanish parallel texts by means of the semantic annotation of biomedical concepts and the automatic expansion of their definitions. In order to overcome the limitations posed by the scarcity of resources available for Spanish, we propose to exploit existing tools targeted at English and then transfer the produced annotations. The evaluations performed show the feasibility of this approach. An enriched set of texts is made available, which can be retrieved, visualized and downloaded through a web interface.
**Keywords:** semantic annotation, definition expansion, biomedical terminology

**Resumen:** Este trabajo busca facilitar la comprensión de textos médicos en un corpus paralelo inglés-español mediante la anotación semántica de conceptos y la expansión automática de definiciones. Considerando la limitación de recursos disponibles para el español, proponemos explotar herramientas dirigidas al inglés para obtener anotaciones que luego se transfieren a los textos en español. Las evaluaciones realizadas muestran la viabilidad de este enfoque. Se hace público un conjunto de textos enriquecidos que se pueden recuperar, visualizar y descargar mediante una interfaz web.
**Palabras clave:** anotación semántica, expansión de definiciones, terminología biomédica

## *1 Introduction*

A vast volume of biomedical knowledge is generated on a daily basis as unstructured text, including scientific articles (Bornmann and Mutz, 2015), patents and medical reports. Natural language processing (NLP) tools have a great deal to contribute to the effective exploitation of this knowledge. In particular, the automatic annotation of biomedical texts with concepts from manually curated thesaurus and knowledge bases, such as the Unified Medical Language System (UMLS) (Bodenreider, 2004), is key to make biomedical knowledge manageable and discoverable. At the same time, linking complex terms in health-related texts to uniquely identified concepts makes it possible to expand their definitions and/or enrich them with information which can improve the text comprehensibility for general audiences. This is particularly relevant as studies show that the possibility of understanding health-related information ("health literacy") predicts a person's health status more accurately than variables such as age, income, level of education and race (MacLeod et al., 2017).

Several resources, tools and methods have been developed to support and/or automate the semantic indexing of biomedical texts, as recently reviewed by (Jovanović and Bagheri, 2017). Most of these resources are, nevertheless, only available for English, with a few exceptions described in the cited review. The lack of tools with similar levels of maturity for Spanish exacerbates an existing imbalance in the access to health-related information for speakers of this language.

ISSN 1135-5948. DOI 10.26342/2018-61-6

The work described in this paper seeks to enhance the access to health information by non-expert population by means of semantic indexing and enrichment of biomedical texts in Spanish. Our hypothesis is that adding a layer of semantic information to biomedical texts can contribute to make them more accessible, as automatically retrieved definitions and related information can be made available to facilitate the comprehensibility of technical terms by non-expert users.

### 1.1 Contributions

Our contributions can be summarized as:

- We explore the possibility of exploiting existing resources and off-the-shelf tools available for English for the annotation of biomedical texts in Spanish;
- We propose a method for transferring automatically-obtained annotations in English texts to their parallel Spanish versions, which we evaluate against a gold standard biomedical corpus. To the extent of our knowledge, our work is the first to compare, for this task, the performance of two semantic similarity functions: one that relies on traditional information retrieval methods based on TF-IDF sparse vectors and another one based on dense vectors that include subword information.
- In order to assess the viability of our approach, we develop a prototype for the annotation of the ScieLO parallel corpus (Neves, Jimeno-Yepes, and Névéol, 2016) as well as an on-line tool that allows the search and visualization of semantically enriched documents in English and Spanish;
- The linguistic resources generated in the context of this work, including the annotated documents in JSON format, are made available for download from the AsisTerm web site.[1]

## *2 Related work*

The automatic or semi-automatic annotation of biomedical texts in English has gathered considerable attention in the past decade and several tools and resources have been developed in this area including cTAKES,[2] MetaMap,[3] NCBO Annotator,[4] and BeCAS,[5] among others. We refer the reader to (Jovanović and Bagheri, 2017), (Hassanzadeh, Nguyen, and Koopman, 2016) and (Groza, Oellrich, and Collier, 2013) for detailed descriptions of these systems and their performance.

The systems developed in the context of the 2013 CLEF-ER challenge for biomedical entity recognition in parallel multilingual corpora (Rebholz-Schuhmann et al., 2013) provide some of the first prototype tools for the annotation of biomedical texts in languages other than English. Among the participating systems there were some targeted at Spanish including the proposed by (Bodnari et al., 2013) and (Attardi, Buzzelli, and Sartiano, 2013), which exploited word alignment information obtained by statistical translation tools in order to transfer annotations from English to Spanish that were then used to train named-entity taggers. In turn, (Berlanga, Nebot, and Jimenez, 2010) introduced the notion of *concept retrieval*, which was based on applying information retrieval methods in order to obtain UMLS concepts relevant to a text and later use them to properly annotate matching text spans. Other initiatives aimed at automatically annotating Spanish biomedical texts include (Carrero, Cortizo, and Gómez, 2008), who proposed to combine machine translation and the MetaMap in order to annotate Spanish texts with UMLS concepts, and (Castro et al., 2010), who developed an automatic system for the recognition of SNOMED CT concepts by computing a similarity function between sentences in clinical notes and SNOMED CT concepts based on the results obtained by querying an Apache Lucene[6] index. (Oronoz et al., 2013) extended the Freeling Spanish analyzer[7] to recognize biomedical entities extracted from available knowledge resources (lists of medical abbreviations and drug names, as well as the SNOMED CT thesaurus) and, more recently, (Pérez, Cuadros, and Rigau, 2018) developed a prototype that uses the UMLS Metathesaurus for biomedical term normalization in order to enrich electronic health records in Spanish.

[1] http://scientmin.taln.upf.edu/scielo/
[2] http://ctakes.apache.org/
[3] https://metamap.nlm.nih.gov/
[4] http://bioontology.org/annotator-service
[5] http://bioinformatics.ua.pt/becas
[6] https://lucene.apache.org/
[7] http://nlp.lsi.upc.edu/freeling/

They evaluated their system by measuring the agreement obtained in parallel English-Spanish corpora annotated with MetaMap (for English) and their prototype (for Spanish). In order to do this, they annotated a set of Spanish health records and their English translations, as well as a manually revised version of the ScieLO corpus. Due to the differences between the proposed approaches and the evaluation methods and datasets–in the cases in which evaluation results are made available–, none of the results of these previous works can be directly compared to ours.

## *3 Semantic annotation of an English-Spanish parallel corpus*

### 3.1 The ScieLO parallel corpus

The ScieLO English-Spanish parallel corpus contains 17,015 metadata entries in Dublin Core XML format from documents included in the SciELO collection, a database of open-access scientific publications. The corpus covers a collection of Spanish health-related scientific journals selected based on their quality. The metadata includes publication information in Spanish (venue, keywords, authors, date) and bi-lingual (English and Spanish) versions of the titles and the abstracts.

### 3.2 Corpus indexing

In order to implement efficient full search and retrieval functionalities in English and Spanish the ScieLO abstracts were converted from XML to JSON and indexed with the Elasticsearch search engine.[8] Basic language processing of the texts (stemming, stop-word removal, relevance scoring of terms) was performed at indexing by means of the standard English and Spanish text analyzers included in Elasticsearch.[9]

### 3.3 Annotation of English abstracts

The semantic annotation of the ScieLO abstracts in English was done by means of the BeCAS annotation services' API.[10] The choice of BeCAS as off-the-shelf annotation system responded mainly to its ease of use and acceptable performance when evaluated against the Medline titles included in the English-Spanish Mantra gold standard (see Section 4.1).

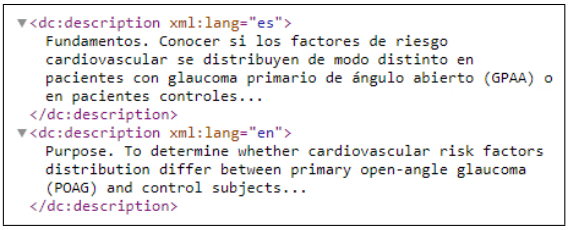


Figure 1: Fragment of a ScieLO document

In the current prototype a filter is applied when calling the BeCAS service in order to retrieve annotations corresponding to the semantic group "DISO" (*Disorders*). We foresee, in future experiments, to expand the coverage of the annotations to the groups *Anatomy*, and *Biological processes*.

We will also analyze the results obtained with other annotation tools with a broader coverage of UMLS types. In particular, the NIH MetaMap, which covers 134 semantic types, in contrast to the 26 UMLS types included in BeCAS.

Even if not an essential element in our proof-of-concept system, the choice of semantic annotator for English would clearly be determining in a real-world case scenario, as the quality and coverage of the annotations of the English text sets an upper bound for the overall performance of the system.

### 3.4 Transfer of annotations to abstracts in Spanish

Once the relevant UMLS concepts are identified in an English abstract by means of the BeCAS service, the spans of texts in the parallel Spanish abstract that best match each of them have to be determined in order to transfer the annotations. Consider, for example, the following fragment from the example included in Fig. 1.

*... risk factors distribution differ between* ***[primary open-angle glaucoma]*** *(****[POAG]****) and control subjects. To assess the strength of this association in* ***[POAG]****.*

BeCAS, in this case, correctly associates the UMLS concept *C0339573* to the three spans of texts in bold. We would like each of these instances to be associated to the corresponding text spans in the Spanish version:

*... los factores de riesgo cardiovascular se distribuyen de modo distinto en pacientes con* ***[glaucoma primario de ángulo abierto]*** *(****[GPAA]****) o*

[8] https://www.elastic.co/

[9] https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-analyzers.html

[10] http://bioinformatics.ua.pt/becas/api

*en pacientes controles. Cuantificar la prevalencia de estos factores en el **[GPAA]**.*

We assume that the instances of the same concepts appear in the same order in the English and Spanish texts. Therefore, we process the Spanish abstract sequentially to find, for each identified concept instance, the text span in Spanish that best matches it. In order to do this, we compute the similarity between each considered text span and all the lexicalizations of the concept available in UMLS.[11] Once a concept instance is associated to a text span, we consider its final offset to continue looking for the next one.[12] It might be the case that, for a given instance of a concept, no matching text span can be identified. In this case, the system will continue with the next annotation retrieved by BeCAS. In the example this might happen, for instance, if the system could not associate the first occurrence of the acronym "GPAA" to the corresponding concept (*primary open-angle glaucoma*). In this case, we would like the system to drop that particular instance of the concept and continue processing the text, looking for a span to which associate the following instance (in this case, the second occurrence of the term "GPAA"). We therefore define a search window for term candidates in Spanish based on the relative position of the annotation in the English text. Defining a search window allows us to keep some level of alignment between the two texts, which minimizes the possibility of associating annotations to the wrong occurrence of a term while, at the same time, prevents the system to keep consuming concept instances in a given position without moving forward.

### 3.4.1 Candidate terms generation

In order to identify spans of text in Spanish as candidates for being annotated, we first split the abstract into sentences, tokenize it and perform part-of-speech (POS) tagging[13] by means of the Stanford CoreNLP library[14] (Manning et al., 2014).

In order to cope with errors produced by the POS tagger and grammatical errors in the source texts, we do not restrict the potential text spans to well-formed noun phrases but instead allow some flexibility in the sequences of POS that can be considered to constitute candidate terms. In fact, we allow any sequence of up to eight tokens beginning with a token tagged as NOUN or PROPN, ending with a token tagged as NOUN, PROPN, ADJ, VERB or NUM, and containing tokens tagged with NOUN, PROPN, ADJ, DASH,[15] CONJ, ADP, DET, ADV, VERB or NUM as potential term candidates. These heuristics are based on the most frequent POS sequences occurring in UMLS terms. In the example above, this rule would produce, as candidate terms: *glaucoma*, *glaucoma primario*, *glaucoma primario de ángulo*, *glaucoma primario de ángulo abierto*.

### 3.4.2 Similarity computation

As mentioned in Section 3.4, a similarity score is computed between candidate terms generated at a given offset in the Spanish text and the UMLS concepts retrieved by BeCAS from the English version. We evaluated two similarity functions: the first one is Elasticsearch's implementation of the BM25 ranking function, which computes the cosine similarity of TF-IDF vectors obtained from the normalized terms of a query (in our case, the candidate text span) and a document (in our case, all the Spanish variants of the concept in UMLS).[16] The second scoring function considered is the cosine similarity between dense vector representations of candidate terms and the Spanish lexicalizations of UMLS concepts. Both for the candidate terms and the lexicalizations of UMLS concepts, their corresponding dense vectors are computed as the average of the normalized embeddings of the words included in them. For the word embeddings we used fastText vectors (Bojanowski et al., 2016) pre-trained with Wikipedia pages in Spanish.[17] Since word representations in fastText are computed as the sum of their character n-gram vectors, embeddings for out-of-vocabulary words can be generated on the

[11] As shown in Table 1, among the sources currently being considered, the one with more concepts variants in Spanish is SNOMED CT, with over 1 million entries.

[12] Overlapping and nested annotations are not considered in the current prototype.

[13] Using a slightly modified version of the universal POS tags: `http://universaldependencies.org/`

[14] `https://stanfordnlp.github.io/CoreNLP/`

[15] The DASH tag was added as the character "-" is frequently used as part of biomedical terms.

[16] `https://www.elastic.co/guide/en/elasticsearch/guide/current/practical-scoring-function.html`

[17] `https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md`

| Source | UMLS code | Entries |
|---|---|---|
| CPT | CPTSP | 2,707 |
| ICPC | ICPCSPA | 723 |
| LOINC Argentina | LNC-ES-AR | 76,586 |
| LOINC Switzerland | LNC-ES-CH | 4,940 |
| LOINC Spain | LNC-ES-ES | 52,641 |
| MedDRA | MDRSPA | 102,097 |
| MeSH | MSHSPA | 70,033 |
| SNOMED CT | SCTSPA | 1,084,815 |
| WHO ART | WHOSPA | 3,106 |

Table 1: Metathesaurus sources in Spanish with number of concept variants

fly. Considering subword information allows the embeddings-based function to correctly assign a high similarity score (0.93) when comparing the concept *C0025179* (lexicalized in the Spanish UMLS as *Metilglucamina* and *Meglumina*) and the candidate term *N-metil glucamina*, which occurs in the Mantra Medline corpus used to evaluate our proposal (see Section 4). This similarity is not captured by the Elasticsearch-based function, which assigns a score of 0 to the candidate term.

## *4 Evaluation*

The Mantra project[18] was one of the first initiatives aimed at the multilingual processing of biomedical texts. In its context, valuable resources were generated, including the Mantra gold-standard parallel corpora for biomedical concept recognition (Mantra GSC) (Kors et al., 2015). The Mantra GSC consists of three parallel corpora: Medline titles, sentences from drug labels provided by the European Medicines Agency (EMEA), and sentences from patents made available by the European Patent Office. The Medline and EMEA corpora include parallel texts in English, French, Dutch, German and Spanish, while the patents corpus is available for English, French and German. In the case of the English-Spanish pairs, both Medline and EMEA corpora include 100 textual parallel units (titles or sentences) annotated with a subset of UMLS concepts from MeSH,[19] SNOMED CT,[20] and MedDRA.[21]

We evaluated the feasibility of the proposed approach for automatically transferring English annotations to Spanish texts by comparing the annotations produced by our system with those included in the Mantra GSC.

For all the evaluations we computed precision (P), recall (R) and F1-score for exact text spans (same boundaries in the gold standard and automatic annotations) as well as for overlapping spans. In order to assess the loss of accuracy when non-exact matching spans where considered, an "overlapping percentage" (OP) was calculated as the relation between the length of the overlapping span and the length of the longest span between the annotation produced by our system and the one in the gold standard. We also considered in our evaluations whether the same gold standard concept CUIs were identified or if different ones where produced by the automatic annotation system. The less restrictive alternatives (overlapping spans and non-matching concepts) were evaluated since they can be of use in specific applications, as argued by (Hassanzadeh, Nguyen, and Koopman, 2016). We report below the results obtained considering only when gold standard CUIs are identified, for the sake of space.[22]

BeCAS does not produce discontinuous annotations and only continuous candidate text spans were considered in the texts in Spanish. Therefore, only continuous annotations in the Mantra corpora were considered for the evaluation and including the first continuous portion of discontinuous gold standard annotations.[23]

The parallel Medline and EMEA corpora included in Mantra are each annotated with 64 different semantic types but BeCAS currently includes only 26.[24] In order to make both sets comparable we considered for the evaluation only Mantra annotations with semantic types recognized by BeCAS.

### 4.1 Evaluation of the annotations produced by BeCAS

We were interested in assessing both the transferring process and the outcomes of the full processing pipeline, including the automatic annotation of the English texts by means of the BeCAS service. We evaluated independently the annotations produced by

[18] `https://sites.google.com/site/mantraeu/`

[19] `https://www.nlm.nih.gov/mesh/`

[20] `https://www.snomed.org/snomed-ct`

[21] `https://www.meddra.org/`

[22] The full results are available at `http://scientmin.taln.upf.edu/scielo/evaluations/Evaluations_AsisTerm.pdf`

[23] Note that 98% of the annotations in the Medline and EMEA corpora are continuous.

[24] `http://bioinformatics.ua.pt/becas/about`

BeCAS against the Mantra English corpora, as it establishes an upper bound for the results of the full pipeline. The F1-score obtained for the BeCAS annotations when considering matching span boundaries was of 0.76 in the case of the Medline corpus and of of 0.67 in the case of EMEA.

## 4.2 Evaluation of the annotation transferring process

Table 2 shows the P, R and F1 results obtained when evaluating the transfer of annotations between the English and the Spanish versions of the Mantra GSC. The F1-scores obtained for the exact matching boundaries for Medline were of 0.60 for the embeddings-based similarity function (FT) and of 0.57 for the Elasticsearch-based one (ES). Different thresholds were considered to decide whether there was a valid match between a candidate term and a concept string when comparing the fastText embeddings. In the case of Medline the best results were obtained with a threshold of 0.7875. In the case of EMEA, the F1-scores obtained were of 0.60 for the embeddings-based similarity function (with a threshold of 0.9250) and of 0.59 for Elasticsearch. It is relevant to note that, even if the embeddings are expected to better capture semantic similarities between candidate text spans and UMLS concepts, considering all the lexicalizations available from multiple UMLS sources for a given concept contribute to mitigate this advantage, which yields to obtaining competitive results with sparse vectors. The choice of similarity function heavily depends, therefore, on the intended usage of the system and implementation decisions.

Due to length restrictions, it is not possible to include in this paper a detailed error analysis. The assumption that the same set of annotations appear in the same order in the English and Spanish versions of the texts and processing them sequentially in the order in which they appear in English, as implemented in our simplified prototype, can explain some of the errors. Others are originated by the proposed similarity functions failing to identify the right candidate term for a given concept. In general, in the case of the Elasticsearch function, this is due to lexical differences between the terms appearing in the texts and the variants of the concepts included in UMLS, as shown in the example

| **Similarity** | **Span** | **P** | **R** | **F1** | **OP** |
|---|---|---|---|---|---|
| *Medline corpus* | | | | | |
| FT | exact | 0,77 | 0,49 | 0,60 | |
| FT | overlap | 0,93 | 0,59 | 0,72 | 0,93 |
| ES | exact | 0,74 | 0,47 | 0,57 | |
| ES | overlap | 0,95 | 0,62 | 0,75 | 0,91 |
| *EMEA corpus* | | | | | |
| FT | exact | 0,89 | 0,45 | 0,60 | |
| FT | overlap | 0,95 | 0,49 | 0,65 | 0,98 |
| ES | exact | 0,77 | 0,48 | 0,59 | |
| ES | overlap | 0,91 | 0,58 | 0,71 | 0,92 |

Table 2: Evaluation of annotations transferred from English to Spanish Mantra GSC

| **Similarity** | **Span** | **P** | **R** | **F1** | **OP** |
|---|---|---|---|---|---|
| *Medline corpus* | | | | | |
| FT | exact | 0,69 | 0,58 | 0,63 | |
| FT | overlap | 0,75 | 0,67 | 0,71 | 0,94 |
| ES | exact | 0,64 | 0,51 | 0,57 | |
| ES | overlap | 0,74 | 0,64 | 0,69 | 0,91 |
| *EMEA corpus* | | | | | |
| FT | exact | 0,67 | 0,49 | 0,57 | |
| FT | overlap | 0,61 | 0,52 | 0,56 | 0,99 |
| ES | exact | 0,56 | 0,49 | 0,52 | |
| ES | overlap | 0,64 | 0,65 | 0,65 | 0,89 |

Table 3: Evaluation of the full pipeline applied to the Mantra GSC

mentioned in section 3.4.2. In the case of the embeddings-based function some errors can be explained by the difficulty to establish a unique, good-for-all, similarity threshold that provides a correct balance between precision and recall so as to identify, in one pass, the best term candidates and their exact boundaries. Performing multiple iterations over the candidate terms–possibly considering different similarity thresholds in each iteration in the case of the embeddings-based score–could contribute to partially mitigate these errors.

## 4.3 Evaluation of the full pipeline

Table 3 shows the results of evaluating the annotations obtained when applying the full processing pipeline, including the automatic annotation of the English texts produced by BeCAS and their transfer to the Spanish versions. When transferring the annotations obtained with BeCAS for the Medline corpus we obtained F1-scores of 0.63 and 0.57 for the fastText and Elasticsearch similarity functions, respectively (consider-

ing exact span boundaries), whereas for the EMEA corpus the F1-scores were of 0.57 and 0.52 for fastText and Elasticsearch functions, respectively. As there were not significant differences between the Medline and EMEA corpora in the evaluation of the annotation transferring process, these lower F1-scores could be explained by the poorer performance observed when annotating EMEA with BeCAS, as mentioned in Section 4.1.

## *5 The AsisTerm prototype*

AsisTerm[25] provides an on-line interface to search and visualize biomedical abstracts from the ScieLO parallel corpus (Neves, Jimeno-Yepes, and Névéol, 2016) in English and Spanish annotated with UMLS concepts.

The annotated abstracts can be downloaded as JSON files including, for each annotation, its starting and ending offsets, the annotated text, and the corresponding UMLS concept, including its CUI, semantic type and group in the Metathesaurus.[26]

### 5.1 Definition expansion

One of the main objectives for associating annotations to the ScieLO abstracts was to make it possible to retrieve additional information that can contribute to facilitate the comprehension of complex terms included in them. When a Spanish abstract is displayed on AsisTerm, its annotations are retrieved and the corresponding text spans are highlighted. When the user clicks on an highlighted text span, all the source-specific identifiers and lexicalizations associated to the concept are displayed, as well as their corresponding definitions, if available. In most cases, UMLS concepts do not have definitions associated in the Metathesaurus.[27] In order to overcome this limitation, we retrieve the definitions (and/or additional information related to the concept) by means of the MedlinePlus Connect API,[28] querying it by the concept's SNOMED CT code.[29] When available (currently for SNOMED, MeSH and MedlinePlus), additional links are included to external pages with source-specific information, such as the concepts' hyperonyms, synonyms, and hyponyms (in the case of SNOMED or MeSH) or related information (in the case of MedlinePlus).

## *6 Conclusions and future work*

In this paper we have presented a prototype aimed at semantically indexing complex terms in biomedical texts as a first step to improve their comprehensibility. As a proof-of-concept experiment, we used these annotations to retrieve and display definitions and related information. We applied our proposal to the annotation of the ScieLO English-Spanish parallel corpus and developed a web-based system to allow searching and visualizing its enriched contents in English and Spanish. We presented a proposal for exploiting existing tools targeted at English and transferring the obtained annotations to Spanish, comparing the performance obtained by means of a classic information retrieval similarity ranking function and the cosine similarity in a continuous vector space. We evaluated both approaches with English-Spanish gold standard corpora in the biomedical domain, obtaining promising results.

In terms of potential extensions to our work, we would like to investigate whether a harmonized combination of annotations obtained from multiple existing tools would significantly improve the accuracy of the results. We would also like to analyze the results obtained with embeddings trained with biomedical texts, which should contribute to obtain vectors better suited for this particular task. Another area to explore is the combination of our annotation transferring proposal with machine translations tools, which would allow to use the system in contexts where no parallel texts are available. Finally, we are also particularly interested in conducting usability tests to assess the degree to which the enriched texts can effectively contribute to improve the comprehension of complex texts by non-expert users.

## *Acknowledgements*

This work is (partly) supported by the Spanish Ministry of Economy and Competitiveness under the María de Maeztu Units of Excellence Programme (MDM-2015-0502) and by the TUNER project (TIN2015-65308-C5-5-R, MINECO / FEDER, UE).

[25] `http://scientmin.taln.upf.edu/scielo/`

[26] `http://ncbi.nlm.nih.gov/books/NBK9679/`

[27] For the subset of UMLS sources that we are currently working with, only 32,338 concepts have definitions in English and 7,154 have definitions in Spanish.

[28] `https://medlineplus.gov/connect/`

[29] MedlinePlus Connect supports queries by SNOMED CT and ICD-10-CM codes.

## *References*


Attardi, G., A. Buzzelli, and D. Sartiano. 2013. Machine translation for entity recognition across languages in biomedical documents. In *CLEF (Working Notes)*.

Berlanga, R., V. Nebot, and E. Jimenez. 2010. Semantic annotation of biomedical texts through concept retrieval. *Procesamiento del Lenguaje Natural*, 45:247–250.

Bodenreider, O. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Bodnari, A., A. Névéol, Ö. Uzuner, P. Zweigenbaum, and P. Szolovits. 2013. Multilingual named-entity recognition from parallel corpora. In *CLEF (Working Notes)*.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Bornmann, L. and R. Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222.

Carrero, F., J. C. Cortizo, and J. M. Gómez. 2008. Building a Spanish MMTx by using automatic translation and biomedical ontologies. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 346–353. Springer.

Castro, E., A. Iglesias, P. Martínez, and L. Castano. 2010. Automatic identification of biomedical concepts in Spanish-language unstructured clinical texts. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 751–757. ACM.

Groza, T., A. Oellrich, and N. Collier. 2013. Using silver and semi-gold standard corpora to compare open named entity recognisers. In *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on*, pages 481–485.

Hassanzadeh, H., A. Nguyen, and B. Koopman. 2016. Evaluation of medical concept annotation systems on clinical records. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 15–24.

Jovanović, J. and E. Bagheri. 2017. Semantic annotation in biomedicine: the current landscape. *Journal of biomedical semantics*, 8(1):44.

Kors, J. A., S. Clematide, S. A. Akhondi, E. M. Van Mulligen, and D. Rebholz-Schuhmann. 2015. A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. *Journal of the American Medical Informatics Association*, 22(5):948–956.

MacLeod, S., S. Musich, S. Gulyas, Y. Cheng, R. Tkatch, D. Cempellin, G. R. Bhattarai, K. Hawkins, and C. S. Yeh. 2017. The impact of inadequate health literacy on patient satisfaction, healthcare utilization, and expenditures among older adults. *Geriatric Nursing*, 38(4):334–341.

Manning, C., M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Neves, M. L., A. Jimeno-Yepes, and A. Névéol. 2016. The ScieLO corpus: a parallel corpus of scientific publications for biomedicine. In *LREC*.

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

Pérez, N., M. Cuadros, and G. Rigau. 2018. Biomedical term normalization of EHRs with UMLS. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*.

Rebholz-Schuhmann, D., S. Clematide, F. Rinaldi, S. Kafkas, E. M. van Mulligen, C. Bui, J. Hellrich, I. Lewin, D. Milward, M. Poprat, et al. 2013. Entity recognition in parallel multi-lingual biomedical corpora: The CLEF-ER laboratory overview. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 353–367. Springer.

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 65-71 recibido 04-04-2018 revisado 02-05-2018 aceptado 04-05-2018

# Plataforma para la extracción automática y codificación de conceptos dentro del ámbito de la Oncohematología (Proyecto COCO)

## *Platform for the automatic extraction and coding of concepts within the scope of Oncohematology (COCO Project)*

**Silvia Sánchez Seda[1], Francisco de Paula Pérez León[1], Jesús Moreno Conde[1], María C. Gutiérrez Ruiz[1], Jesús Martín Sánchez[2], Guillermo Rodríguez[2], José Antonio Pérez Simón[2], Carlos L. Parra Calderón[1]**

[1] Grupo de Innovación Tecnológica del Hospital Universitario Virgen del Rocío de Sevilla
[2] Unidad de Gestión Clínica de Hematología del Hospital Universitario Virgen del Rocío de Sevilla
(jesus.moreno.conde.exts@juntadeandalucia.es)

**Resumen:** El proyecto COCO tiene como objetivo diseñar, desarrollar y validar un sistema de extracción de conocimiento que, a partir de los textos de la Historia de Salud Electrónica, codifique automáticamente los diagnósticos de Oncohematología mediante Tecnologías del Lenguaje basada en un estándar de pipeline interoperable. La necesidad de normalizar el conocimiento de la Historia Clínica constituye un gran desafío. Puesto que la CIE-10 presenta limitaciones para representar esta información, se desarrolló la norma CIE-O-3, para dar soporte a este tipo de patologías. Se propone desarrollar el primer pipeline de Procesamiento del Lenguaje Natural de componentes interoperables, así como, el primer codificador automático CIE-O-3 y CIE-10. Nuestro sistema servirá de apoyo a la decisión, investigación y gestión clínica en este campo.
**Palabras clave:** Recuperación de conocimiento clínico, Codificación Automática, Procesamiento del Lenguaje Natural, Oncohematología, CIE-10, CIE-O-3

**Abstract:** The COCO project aims to design, develop and validate a knowledge extraction system that, based on the texts of the Electronic Health Record, automatically codes Oncohematology diagnostics using Language Technologies based on an interoperable pipeline standard. The need to standardize knowledge of the Electronic Health Record is a major challenge. Since ICD-10 has limitations in representing this information, ICD-O-3 was developed to support this type of pathology. It is proposed to develop the first Natural Language Processing pipeline of interoperable components, as well as the first ICD-O-3 and ICD-10 automatic encoder. Our system will support clinical decision making, research and management in this field.
**Keywords:** Recovery of clinical knowledge, Automatic Coding, Natural Language Processing, Oncohematology, ICD-10, ICD-O-3

## *1 Introducción*

El tratamiento de la información no estructurada en el sector sanitario es un gran desafío, a la par que, muy prometedor. El uso de información estructurada en los hospitales es escaso, y la mayor fuente de información textual son los informes escritos por los médicos después de una consulta o un procedimiento, los cuales se expresan en texto libre. Estos informes narrativos (e.g., informes de consulta, patología, radiología, etc.) contienen información valiosa para el diagnóstico que puede ayudar al pronóstico y a predecir el comportamiento biológico de las enfermedades (Mohanty, 2007). En este sentido, varios estudios han señalado la viabilidad del uso de técnicas de Procesamiento del Lenguaje Natural (PLN) para la estructuración de informes de texto libre (Spasić, 2014) (Buckley, 2012) (Warner, 2011)

ISSN 1135-5948. DOI 10.26342/2018-61-7 

(Martinez, 2011). Evaluaciones de estos sistemas sugieren que pueden proporcionar datos precisos sobre la prestación de servicios y el estado clínico del paciente (Chan, 2010). En particular, las técnicas de PLN pueden ser relevantes para la investigación y soporte a la decisión clínica, al permitir la recuperación de información, así como la respuesta a preguntas en el contexto de una multitud de datos disponibles. Todo ello, eliminando el uso intensivo de recursos para la revisión manual (Tange, 1998).

La gran mayoría de los trabajos citados se han centrado en el idioma inglés, no así en el español, que es el segundo idioma más hablado del mundo hoy en día. En el ámbito de PLN en español aplicado a medicina se encuentran solamente algunos trabajos (Menasalvas, 2016) (Costumero, 2014) (Oronoz, 2013) (Medrano, 2018), mientras que el número especializado en el área de la codificación automática de diagnósticos es más reducido. Existen aplicaciones comerciales como Savanamed (Espinosa, 2016) y MeaningCloud que realizan codificaciones automáticas en el dominio de la medicina en español, pero ninguna de ellas realiza una codificación automática en un dominio más especializado como es la Oncohematología, haciendo uso de la CIE-O-3 ya que es un dominio para el que la CIE-9 o CIE-10 presentan limitaciones.

La CIE-10 es la décima edición de la Clasificación internacional de enfermedades, que se corresponde con la versión española de la ICD (International Statistical Classification of Diseases and Related Health Problems), y determina la clasificación y codificación de las enfermedades y una amplia variedad de signos, síntomas, hallazgos anormales, denuncias, circunstancias sociales y causas externas de daños y/o enfermedad. Sin embargo, esta codificación no facilita códigos de morfología en el índice alfabético de enfermedades ni dispone de un apéndice de morfologías, por lo que, para la adecuada codificación de la morfología de los tumores, se deberá acudir a la clasificación CIE-O. La CIE-O es una clasificación dual, con sistemas de codificación tanto para la topografía como para la morfología. El código topográfico de la CIE-O describe el sitio de origen de las neoplasias. El código de morfología describe el tipo de células del tumor y su actividad biológica; en otras palabras, las características del tumor mismo. En la actualidad, se utiliza la tercera edición de la CIE-O (CIE-O-3).

En la literatura, tan sólo encontramos dos trabajos orientados a codificar de forma automática la Historia de Salud Electrónica **(**HSE) en base a CIE-O (Jouhet, 2012) (Kavuluru, 2013). En (Jouhet, 2012) se clasifican informes de anatomía patológica en base CIE-O-3 utilizando para ellos técnicas de Aprendizaje Automático. En estas técnicas, como algoritmos de clasificación se utilizaron Naïve Bayes y SVM, los cuáles fueron evaluados sobre 5.121 informes de anatomía patológica elaborados por 35 profesionales médicos. Para medir la eficacia del sistema desarrollado, se tuvo en cuenta la capacidad del mismo para atribuir correctamente un código preciso de la CIE-O-3, tanto para los ejes topográficos como morfológicos. El sistema obtuvo una medida-F de 71,5% para topografía y 85,4% para morfología, estos resultados, sugieren que los informes de anatomía patológica podrían ser útiles como fuente de datos para sistemas automatizados con el fin de identificar y notificar nuevos casos de cáncer. Además, se hace evidente la necesidad de trabajos futuros que incluyan técnicas de PLN así como la incorporación de otros tipos de documentos médicos.

Por su parte, en (Kavuluru, 2013) se codificó automáticamente el diagnóstico en base a CIE-O-3 para un total de 56.426 informes de anatomía patológica asociados con casos de cáncer almacenados en el Registro de Cáncer de Kentucky, y procedentes de 35 laboratorios diferentes. En este documento, sólo un código CIE-O-3 se vinculó con cada informe de anatomía patológica, y los métodos de Aprendizaje Automático implementados obtuvieron una medida-F de 90%. Los autores del artículo manifiestan la necesidad de seguir investigando en esta línea con objeto de perfeccionar el sistema y extenderlo a otros dominios. Estos trabajos están centrados en el dominio del inglés y según nuestro conocimiento, no existe ningún trabajo previo que aplique estas técnicas en historias clínicas en español, de ahí la relevancia y el impacto de la innovación propuesta. Esta propuesta supondría desarrollar, en el dominio del español, el primer pipeline de PLN de componentes interoperables basados en el estándar UIMA en castellano para extraer conocimiento de la HSE así como también implementar el primer codificador automático en base a CIE-O-3 y CIE-10, sirviendo de experiencia temprana para el desarrollo de procesos de Compra Pública de Innovación

sobre TL en Sanidad tanto por parte de la Administración Pública Andaluza como de otras Comunidades Autónomas en el Marco del Plan de Impulso de Tecnologías del Lenguaje del MINETAD.

## *2 Objetivo*

El proyecto COCO tiene como objetivo diseñar, desarrollar y validar un sistema de extracción de conocimiento a partir de la información contenida en la HSE. Dicho sistema, mediante técnicas de Aprendizaje Automático, se centrará en la codificación automática de diagnósticos en el dominio de Oncohematología, sirviendo como soporte a la toma de decisiones clínicas, además de proporcionar grandes ventajas tanto a la investigación como a la gestión clínica. Para ello este proyecto persigue:

- Desarrollar un pipeline de componentes interoperables que apliquen técnicas de PLN adaptadas al dominio clínico en español capaz de procesar el texto libre para estructurar la información contenida en la HSE.
- Desarrollar mediante técnicas de Aprendizaje Automático, nuevos modelos de clasificación que permitan la codificación automática del diagnóstico, en base a CIE-0, asociado a la HSE de los pacientes de Oncohematología, mapeando dicha codificación con la clasificación CIE-10 utilizada en el Sistema Sanitario Público de Andalucía (SSPA) a la hora de codificar los diagnósticos asociados a los pacientes.

## *3 Material y Métodos*

Este proyecto propone un estudio compuesto por 6 fases principales que se desarrollarán en 24 meses:

**Fase 1**. Identificación de los sujetos de estudio y análisis preliminar de las HSE.

- Identificación de los sujetos de estudio.
- Análisis preliminar de la HSE.
- Proceso de extracción, transformación y carga (ETL) de las HSE.

**Fase 2.** Desarrollo y ejecución de componentes interoperables que apliquen técnicas de Procesamiento del Lenguaje Natural (PLN) para la extracción de conocimiento de la HSE.

- Creación de un corpus ad-hoc para la extracción del conocimiento de la HSE.
- Desarrollo de módulos interoperables que permitan realizar el workflow completo de PLN para extraer el conocimiento de las HSE.
- Evaluación de la precisión general del sistema para verificar la adecuación de los procesos y herramientas generados.

**Fase 3**. Estudio e identificación de la información estructurada y análisis combinado del conocimiento generado.

- Identificación de información estructurada.
- Integración de datos estructurados identificados en esquema común.

**Fase 4.** Creación, análisis y despliegue de modelos de clasificación para la codificación automática de diagnósticos.

- Análisis descriptivo de datos recopilados.
- Creación de modelos para el Aprendizaje Automático.
- Mapeo entre CIE-O-3 y CIE-10.
- Evaluación de los modelos generados.
- Despliegue de los modelos más prometedores.

**Fase 5.** Integración de la información en un data-warehouse orientado a la investigación clínico y traslacional y gestión clínica.

- Diseño del Data Warehouse.
- Implementación del Data Warehouse.

**Fase 6.** Validación e implementación del sistema en el entorno de la Unidad de Gestión Clínica (UGC) de Hematología del Hospital Universitario Virgen del Rocío de Sevilla (HUVR).

- Despliegue del sistema.
- Evaluación del sistema.

La población principal de sujetos de estudio estará compuesta por datos anonimizados de pacientes que acudan al Servicio de Oncohematología del HUVR desde el año 2013 hasta el fin de 2016. Se incluirán los pacientes con algún episodio en el Servicio de Oncohematología del HUVR y se excluirán aquellos con falta de datos relevantes.

Actualmente, el proyecto COCO se encuentra en las primeras 2 fases del proyecto donde se centran los trabajos en la identificación y creación del corpus de trabajo y en el diseño y desarrollo de la arquitectura de componentes del sistema.

## 3.1 Arquitectura

El diseño del sistema COCO presenta una arquitectura similar a la que se presenta a continuación:

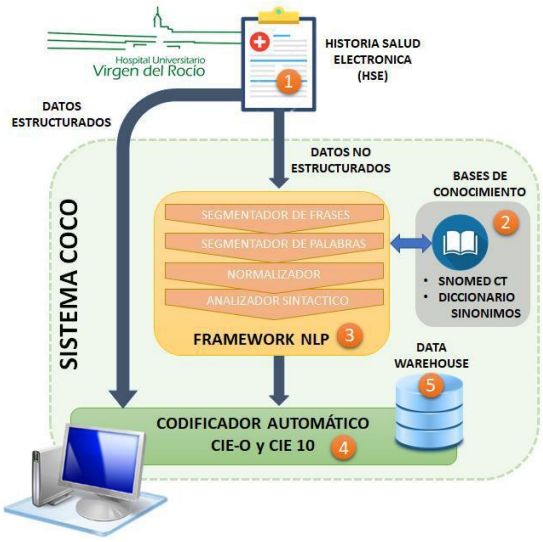


Figura 1: Representación de la arquitectura de componentes del sistema COCO

**(1) HSE:** informes clínicos que contienen la información procedente de los sistemas de Historia Clínica Electrónica del Hospital. Esta información podrá ser información de texto libre no estructurada que pasará por los algoritmos de procesamiento de lenguaje natural, así como códigos e información estructurada que permitirá mejorar la codificación automática de los códigos de diagnósticos de los pacientes

**(2) Bases de conocimiento:** servicio interno que contiene información sobre las terminologías de referencia dentro del ámbito de la oncohematología. Además, nos servirá para identificar los conceptos relevantes en el texto libre extraído de la HSE, así como, información relacionada con los mismos (Ej. Diccionarios de sinónimos, acrónimos, estructuras gramaticales, etc.)

**(3) Framework NLP**: Servicio Web basado en Apache UIMA y dividido en varios módulos encargados de realizar una tarea concreta en el procesamiento del texto (Ej. segmentación, normalización, tokenización, análisis, etc.). Para la realización de este framework se está valorando sustituir Apache UIMA por Python (mediante la librería NLTK) como lenguaje de referencia para el desarrollo de tareas de procesamiento de lenguaje natural. A través de este framework se implementará una interfaz de servicios para la ejecución de tareas de PLN sobre el texto libre contenido en la HSE de los pacientes, garantizando la escalabilidad y reutilización del sistema, a través de un Pipeline de trabajo segmentado.

**(4) Codificador automático**: algoritmos basados en técnicas avanzadas de Data Mining y aprendizaje automático para la identificación de conceptos CIE-O a partir de los conceptos identificados en el Framework NLP y los datos estructurados recogidos en la HSE de los pacientes.

(**5) DataWareHouse**: sistema de almacenamiento y análisis de la información clínica de los pacientes obtenida de implementar los procesos de PLN. Este sistema ofrecerá una interfaz de trabajo sobre la que implementar consultas guiadas sobre los datos registrados con el fin de facilitar la explotación de la información contenida en el sistema.

## 3.2 Corpus

El corpus trabajo generado dentro del marco del proyecto consiste en informes clínicos procedentes del Hospital Virgen del Rocío de Sevilla. De todos ellos, se distinguen informes únicos de alta, informes de consulta y hojas de evolución, todos ellos escritos en texto libre. Los informes fueron cogidos aleatoriamente entre informes del año 2013 al 2016 y fueron completamente anonimizados. El corpus completo fue anotado por dos expertos del dominio, siguiendo las guías de anotación que desarrollamos previamente. Algunos documentos del corpus no fueron incluidos en el corpus final, ya que sirvieron de entrenamiento para los anotadores. La herramienta de anotación usada fue la que ofrece APACHE UIMA y la tarea de anotación consistió en identificar los eventos clínicos relacionados con alguno de los 160 conceptos CIE-O-3 que entraban dentro del alcance del

proyecto, las partículas que expresan negación en la frase y las palabras que están dentro del alcance de cada partícula de negación.

### 3.3 Flujo de procesamiento NLP

La tecnología utilizada para el procesamiento del lenguaje es Apache UIMA, el único estándar reconocido por OASIS (Organización para el Avance de Estándares de Información Estructurada). El desarrollo bajo el estándar Apache UIMA permite la interoperabilidad y los distintos componentes de PLN podrían ser implementados por distintas organizaciones. También se plantea el uso de Python y la librería NLTK, ya que es un lenguaje de fácil uso y multiplataforma.


Figura 2: Representación del pipeline del procesamiento del texto

Se entrenarán 160 clasificadores binarios, uno por cada código CIE-O-3.
Se realizará un proceso de evaluación para el sistema desarrollado de forma que se considera que un registro está correctamente codificado si el clasificador del código que le corresponde lo clasifica como positivo y los clasificadores de los códigos que no le corresponden lo clasifican como negativo. Para este proceso de evaluación se usará un subconjunto de documentos para la fase de entrenamiento del sistema y el resto de documentos para la fase de evaluación de los algoritmos. En la fase de evaluación se analizarán parámetros de calidad como la precisión, la exhaustividad y el valor-F para cada clasificador.

## *4 Resultados*

Como resultado esperado, se prevé desarrollar un nuevo producto que servirá de apoyo tanto a procesos de soporte a la decisión clínica como a tareas de investigación y gestión clínica dentro de la UGC de Hematología del Hospital Universitario Virgen del Rocío. Además, se espera que este sistema se integre en la práctica clínica habitual de la Unidad, pudiendo replicarse en otros dominios y centros del SSPA. Este hecho favorecería importantes mejoras al SSPA, entre las que se destacan:

- Reducción de costes: Se produciría una disminución de los recursos necesarios para llevar a cabo la codificación de la HSE (personas involucradas en el proceso/mes). Además, se reducirían también los costes asociados a la formación de codificadores.
- Disminución de errores. Se reduciría el número de errores que los anotadores humanos cometen al tener que trabajar con miles de posibles códigos al asignar las etiquetas de la CIE-10 a un documento.
- Automatización de procesos de codificación. La automatización del proceso de asignación de códigos CIE-10 a la HSE aumentaría el número de historias clínicas que se codifican.
- Mejora en la capacidad de análisis epidemiológico o estadístico: Se podrían tener datos sobre la incidencia y prevalencia de las distintas patologías en una población dada. Además, se mejoraría el conocimiento profundo y actualizado de la información gestionada por la UGC de Hematología para la dirección de políticas públicas.
- Mejora en la gestión de la Unidad: Facilitaría las tareas de gestión en la UGC de Hematología para el cálculo de sus indicadores anuales. Además, el sistema de gestión sanitaria obtendría de origen la información clínica con los diagnósticos codificados, con todas las ventajas que esto supone.
- Aumento de la investigación: Facilitaría la identificación de cohortes de pacientes para el desarrollo de ensayos clínicos en el HUVR. La estandarización de la información en texto libre contenida en la HSE facilitaría su explotación por

otros servicios del SSPA. Esta información podría ser compartida con otros sistemas sanitarios del SSPA lo que permitiría avanzar en la investigación de las hemopatías malignas, patologías graves con poca frecuencia.

- Mejora de la riqueza semántica de los sistemas de Historia Clínica. La UGC de Hematología contaría con la información asociada a las hemopatías malignas normalizada de forma automática a nivel internacional según el estándar CIE-O-3.
- Promoción del uso de herramientas que faciliten el desarrollo de procesos de PLN en el SSPA. El desarrollo de un pipeline de PLN de componentes interoperables permitiría ejecutar las distintas fases (segmentador de frases, segmentador de palabras, etiquetador gramatical, etc.) de forma independiente en función de las necesidades de las tareas a resolver. Al estar desarrollado bajo el estándar UIMA, los distintos componentes de PLN interoperables podrían ser implementados por distintas organizaciones.
- Proveer al SSPA de recursos lingüísticos reutilizables dentro de la política de Reutilización de la Información del Sector Público (RISP).
- Impulsar el desarrollo de herramientas de soporte a la decisión para la personalización de tratamientos en base a pacientes con diagnósticos similares, nuevos servicios de información que apliquen algoritmos predictivos, etc.

## *5 Discusión*

Entre los trabajos previos, centrados en el ámbito del español, solamente se encuentran algunas aplicaciones comerciales incipientes sin información detallada como Savanamed (Espinosa-Anke, 2016) o MeaningCloud que hacen uso de PLN, del Aprendizaje Automático y otras técnicas para beneficiar al sector de la salud y extraer conocimiento de la HSE. En cuanto a los trabajos orientados a codificar de forma automática la HSE en base a CIE-O-3, no existe ningún trabajo previo para el español. Por tanto, el sistema COCO presenta una clara novedad en el SSPA pues no hay ninguna iniciativa parecida que resuelva el problema planteado. Esto, supone una gran mejora, no sólo desde el punto de vista económico sino desde el punto de vista de la investigación, la práctica clínica y traslacional, la gestión, etc., y servirá de experiencia temprana para el desarrollo de procesos de Compra Pública de Innovación sobre TL en Sanidad, tanto por parte de la Administración Pública Andaluza como de otras Comunidades Autónomas en el Marco del Plan del MINETAD.

## *Agradecimientos*

Esta investigación ha sido financiada en parte por la Plataforma de Innovación en Tecnologías Médicas y Salud (Plataforma ITEMAS, PT13/0006/0036) financiado por el Instituto de Salud Carlos III y por el proyecto COCO (PIN-0121-2017) financiado por la consejería de Salud de la Junta de Andalucía ambos cofinanciados a través de los Fondos Europeos de Desarrollo Regional (FEDER).

## *Bibliografía*

Mohanty, S. K., Piccoli, A. L., Devine, L. J., Patel, A. A., William, G. C., Winters, S. B., y Parwani, A. V. 2007. Synoptic tool for reporting of hematological and lymphoid neoplasms based on World Health Organization classification and College of American Pathologists checklist. *Bmc Cancer*, 7(1), 144.

Spasić, I., Livsey, J., Keane, J. A., y Nenadić, G. 2014. Text mining of cancer-related information: review of current status and future directions. *International journal of medical informatics*, 83(9), 605-623.

Buckley, J. M., Coopey, S. B., Sharko, J., Polubriaginof, F., Drohan, B., Belli, A. K., y Specht, M. C. 2012. The feasibility of using natural language processing to extract clinical information from breast pathology reports. *Journal of pathology informatics*, 3.

Warner, J. L., Anick, P., Hong, P., y Xue, N. 2011. Natural language processing and the oncologic history: is there a match? *Journal of oncology practice*, 7(4), e15-e19.

Martinez, D., y Li, Y. 2011. Information extraction from pathology reports in a hospital setting. In *Proceedings of the 20th ACM international conference on*

*Information and knowledge management*, pages 1877-1882. ACM.

Chan, K. S., Fowles, J. B., y Weiner, J. P. 2010. Electronic health records and the reliability and validity of quality measures: a review of the literature. *Medical Care Research and Review*, 67(5):503-527.

Tange, H. J., Schouten, H. C., Kester, A. D., & Hasman, A. 1998. The granularity of medical narratives and its effect on the speed and completeness of information retrieval. *Journal of the American Medical Informatics Association*, 5(6):571-582.

Menasalvas, E., Rodriguez-Gonzalez, A., Costumero, R., Ambit, H., y Gonzalo, C. 2016. Clinical Narrative Analytics Challenges. In *International Joint Conference on Rough Sets*, pages 23-32. Springer, Cham.

Costumero, R., García-Pedrero, Á., Gonzalo-Martín, C., Menasalvas, E., y Millan, S. 2014. Text analysis and information extraction from Spanish written documents. In *International Conference on Brain Informatics and Health*, pages 188-197. Springer, Cham.

Oronoz, M., Casillas, A., Gojenola, K., y Perez, A. 2013. Automatic annotation of medical records in Spanish with disease, drug and substance names. In *Iberoamerican Congress on Pattern Recognition*, pages 536-543. Springer, Berlin, Heidelberg.

Medrano, I. H., Guijarro, J. T., Belda, C., Ureña, A., Salcedo, I., Espinosa-Anke, L., y Saggion, H. 2018. Savana: Re-using Electronic Health Records with Artificial Intelligence. *International Journal of Interactive Multimedia and Artificial Intelligence*, 4 (Special Issue on Big Data and e-Health).

Espinosa-Anke, L., Tello, J., Pardo, A., Medrano, I., Ureña, A., Salcedo, I., y Saggion, H. 2016. Savana: A Global Information Extraction and Terminology Expansion Framework in the Medical Domain.

Jouhet, V., Defossez, G., Burgun, A., Le Beux, P., Levillain, P., Ingrand, P., y Claveau, V. 2012. Automated classification of free-text pathology reports for registration of incident cases of cancer. *Methods of information in medicine*, 51(3):242.

Kavuluru, R., Hands, I., Durbin, E. B., y Witt, L. 2013. Automatic extraction of ICD-O-3 primary sites from cancer pathology reports. *AMIA Summits on Translational Science Proceedings*, 2013, 112.

# *Análisis del habla*

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 75-82 recibido 04-04-2018 revisado 24-04-2018 aceptado 04-05-2018

# Estudio sobre el impacto del corpus de entrenamiento del modelo de lenguaje en las prestaciones de un reconocedor de habla

## *Study on the impact of the training corpus of the language model on the performance of a speech recognizer*

**Andrés Piñeiro Martín[1], Carmen García-Mateo[1], Laura Docío-Fernández[1], Xosé Luis Regueira[2]**

[1] AtlanTTic Research Center – Escola de Enxeñaría de Telecomunicación – Universidade de Vigo Campus Universitario 36310 Vigo (Spain)

[2] Instituto da Lingua Galega Universidade de Santiago de Compostela Praza da Universidade, 4, 15782 Santiago de Compostela (Spain)

E-mail: apinheiro@gts.uvigo.es, carmen.garcia@uvigo.es, ldocio@gts.uvigo.es, xoseluis.regueira@usc.es

**Resumen:** Dentro del reconocimiento automático del habla, los modelos de lenguaje estadísticos basados en la probabilidad de secuencia de palabras (n-gramas) suponen uno de los dos pilares sobre los que se basa su correcto funcionamiento. En este trabajo se expone el impacto que tienen sobre las prestaciones de reconocimiento a medida que estos modelos se mejoran con más texto de mejor calidad, cuando estos se ajustan a la aplicación final del sistema, y por lo tanto, cuando se reducen el número de palabras fuera de vocabulario (*Out Of Vocabulary - OOV*). El reconocedor con los distintos modelos de lenguaje ha sido aplicado sobre cortes de audio correspondientes a tres marcos experimentales: oralidad formal, habla en noticiarios, y TED talks en gallego. Los resultados obtenidos muestran claramente una mejora sobre los marcos experimentales propuestos.

**Palabras clave:** modelos de lenguaje, reconocimiento automático del habla, palabras fuera de vocabulario

**Abstract:** Within the automatic speech recognition, statistical language models based on the probability of word sequences (n-grams) represent one of the two pillars on which its correct functioning is based. In this paper, the impact they have on the recognition result is exposed as these models are improved with more text of better quality, when these are adjusted to the final application of the system, and therefore, when the number out of vocabulary (*OOV*) words is reduced. The recognizer with the different language models has been applied to audio cuts corresponding to three experimental frames: formal orality, talk on newscasts, and TED talks in Galician. The results obtained clearly show an improvement over the experimental frameworks proposed.

**Keywords:** language models, automatic speech recognition, Out of vocabulary words

## *1 Introducción*

Hoy en día, los modelos de lenguaje estadísticos (*ML*) son usados en numerosas aplicaciones como el reconocimiento del habla, el reconocimiento de la escritura, el reconocimiento óptico de caracteres (*OCR*), en correcciones ortográficas, etc.

Dentro del reconocimiento automático del habla, los modelos de lenguaje usados definen la estructura de lenguaje, es decir, restringen de forma adecuada las secuencias de unidades lingüísticas más probables. Los más utilizados son los que funcionan como modelos de probabilidad de secuencia de palabras (n-gramas), y son estimados a partir del análisis de grandes cantidades de texto. Estos MLs poseen una integración sencilla con el modelado acústico, pero pueden ser muy generales, requiriendo una adaptación a la tarea de reconocimiento de la que se trate.

En este trabajo se busca analizar el efecto de mejorar el corpus de aprendizaje de los modelos de lenguaje sobre las prestaciones del reconocedor de habla. Para ello, se han reunido amplios corpus de texto de distintas fuentes y con distintas características y temáticas. Los MLs entrenados con estos corpus de textos han sido probados reconociendo cortes de audio dentro de tres marcos experimentales formados

ISSN 1135-5948. DOI 10.26342/2018-61-8 

por: cortes de audio de oralidad formal (lecturas literarias y discursos leídos), cortes de audio de telediarios de la TVG, y cortes de audio correspondientes a TED talks en gallego.

A lo largo del trabajo se estudia el efecto de ajustar los MLs en función de la aplicación final del reconocedor, así como la relación entre las palabras fuera de vocabulario de cada uno de los modelos, la WER (*Word Error Rate*) obtenida y la perplejidad.

## *2 Modelos de lenguaje estadísticos*

### 2.1 N-gramas

Actualmente los modelos de lenguaje más usados en el reconocimiento automático del habla son los probabilísticos basados en n-gramas, los cuales permiten hacer una predicción estadística de la próxima palabra en la secuencia de texto, ya que asumen que la n-ésima palabra depende de las *n-1* anteriores (historia).

Estos modelos describen el lenguaje como cadenas de Márkov de orden *n-1*, donde la probabilidad de que aparezca una palabra depende únicamente de la palabra anterior o palabras anteriores (en función del orden). Por lo tanto, la probabilidad $P(w_1,...,w_m)$ de observar la secuencia de palabras $w_1,...,w_m$ se aproxima como:

$$P(w_1,\ldots,w_m)=\prod_{i=1}^{m} P(w_i|w_1,\ldots,w_{i-1})\approx \approx\prod_{i=1}^{m} P(w_i|w_{i-(n-1)},\ldots,w_{i-1}) \quad (1)$$

Esta última probabilidad condicionada puede calcularse a partir recuentos de frecuencias de la siguiente forma:

$$P(w_i|w_{i-(n-1)},\ldots,w_{i-1})= =\frac{C(w_{i-(n-1)},\ldots,w_{i-1},w_i)}{C(w_{i-(n-1)},\ldots,w_{i-1})} \quad (2)$$

donde $C(w_1,...,w_m)$ es el número de veces que la secuencia $w_1,...,w_m$ ha sido vista en el corpus de entrenamiento (Jurafsky y Martin, 2008).

Los modelos basados en n-gramas poseen una formulación sencilla, facilidad en su implementación y consistencia con los datos de entrenamiento (debido a esto son los más utilizados actualmente), pero también presentan limitaciones:

- El valor de *n* debe ser relativamente pequeño, ya que a medida que este crece aparecen problemas computacionales en la estimación de los parámetros. Debido a esto, los modelos más comunes suelen ser trigramas (n=3) o tetragramas (n=4), aunque comienzan a aparecer pentagramas (n=5) u órdenes mayores.
- Los modelos de n-gramas no se acomodan bien a los cambios en el discurso. Es muy complicado obtener un ML que sea representativo del habla que se quiere reconocer cuando las temáticas del audio varían. Por esto, es necesario actualizar o variar el ML en función del tema o temática del audio que se desea reconocer.
- Estos modelos tienen problemas de dispersión. Existen gran cantidad de eventos u oraciones que no son vistos durante el proceso de entrenamiento y obtienen una frecuencia relativa igual a cero. Esto causa que a frases con alta probabilidad que contengan dicho evento, se les asigne probabilidad cero. Para evitar que el modelo se "rompa" cuando aparecen estos ceros, es necesario aplicar a los modelos suavizados (*smoothing*), donde se asignan probabilidades pequeñas a este tipo de eventos en lugar de probabilidades iguales a cero; e interpolados, donde las probabilidades se combinan con probabilidades de órdenes menores para intentar mantener la información de qué palabras aparecen con una mayor frecuencia en el texto. Un modelo interpolado que combine probabilidades de unigrama (1-grama) y bigrama (2-grama) es definido de la siguiente forma:

$$P_{Interpolated}(w_i|w_{i-1})= =\lambda P(w_i|w_{i-1})+(1-\lambda)P(w_i) \quad (3)$$

Donde $0\leqslant\lambda\leqslant 1$ es el peso del interpolado. Para este caso, cuanto mayor sea, más se parecerá al comportamiento de un modelo bigrama; y cuanto menor sea , más se parecerá al comportamiento de un modelo unigrama (Jurafsky y Martin, 2008).

### 2.2 Modelos de lenguaje alternativos

Existen modelos alternativos (Vicente et al. 2015) o híbridos que lidian con los anteriores inconvenientes, donde se combina la eficiencia local del modelo de n-gramas con otros que capturan información sintáctica o de larga distancia (información que no puede extraerse de los eventos del modelo de n-gramas).

Los modelos sintácticos usan gramáticas libres de contexto probabilísticas (Probabilistic

Context Free Grammars, PCFGs) para estudiar cómo se relacionan las palabras del corpus, donde el objetivo es basarse más en la gramática. Las PCFGs representan un mecanismo eficiente para modelar las relaciones de larga distancia entre las diferentes unidades léxicas en una oración. Estos modelos de lenguaje alternativos se utilizan en campos como el reconocimiento sintáctico de formas y lingüística computacional, pero no presentan buenos resultados en tareas complejas con grandes vocabularios como el reconocimiento automático del habla, ya que el costo computacional es elevado.

Los modelos factorizados suponen una solución intermedia: poseen la extensión de los basados en n-gramas y son menos demandantes en recursos que los modelos sintácticos. En estos modelos cada palabra *w* es una colección de características o factores, y es posible incorporar conocimiento lingüístico.

Actualmente, con el aumento de la potencia de cómputo disponible, el uso de redes neuronales profundas se ha extendido numerosos campos de aplicación. Dentro del ASR, los modelos de lenguaje basados en redes neuronales recurrentes (*Recurrent neural network langauge models - RNNLM)* están ganando terreno a los modelos de n-gramas clásicos (Mikolov et al. 2011). Sin embargo, estos modelos no pueden ser utilizados de forma sencilla en la decodificación, por lo que la técnica habitual para su aplicación consiste en hacer una etapa de rescoring sobre una decodificación previa que utiliza los clásicos n-gramas. Existen diversos algoritmos que implementan este rescoring como los presentados en (Xu et al. 2018), (Sundermeyer et al. 2014), o (Chen et al. 2017).

## 2.3 Evaluación de los MLs: Perplejidad

A la hora de evaluar la calidad de un modelo de lenguaje, la perplejidad (*PP*) es la medida típicamente usada.

La perplejidad puede ser considerada como una medida, en promedio, de cuántas palabras diferentes igualmente probables pueden seguir a una palabra determinada, es decir, sobre un conjunto de prueba, es la probabilidad inversa del conjunto, normalizado por el número de palabras. Por lo tanto, para calcularla se necesita tanto un modelo de lenguaje como un texto de prueba. Puede ser calculada de la siguiente forma (Jurafsky y Martin, 2008), para un conjunto de prueba $W = w_1,...,w_m$:

$$PP(W) = \sqrt[m]{\prod_{i=1}^{m} \frac{1}{P(w_i | w_1, \ldots, w_{i-1})}} \qquad (4)$$

Las medidas más bajas de perplejidad representan mejores modelos de lenguaje. Aquí se debe matizar que lo que nos dice la PP es que "modelan mejor el lenguaje" de ese texto de prueba, y no necesariamente que funcionen mejor en los sistemas de reconocimiento de habla.

## *3 Creación de los MLs*

Hoy en día existe una enorme cantidad de texto en formato digital. Si pensamos en el ámbito de Internet, sólo el número libros, artículos, noticias o reportajes que son publicados diariamente y están disponibles al acceso de cualquier usuario es inmenso; pero existen inconvenientes si se quiere usar este texto en el entrenamiento de los modelos de lenguaje. Los principales problemas son:

- La limpieza y el preprocesado de los textos: la variedad de textos disponibles en la red es directamente proporcional a la cantidad de formatos con los que estos están presentes. El unificar formatos y codificaciones a la hora de descargarlos o utilizarlos puede suponer un problema.

- El otro problema principal está relacionado con que los textos utilizados para entrenar los modelos de lenguaje sean representativos del habla que se quiere reconocer. En nuestro caso, con el gallego, no es sencillo conseguir grandes cantidades de texto representativo del idioma, ya que no existen amplios corpus de texto disponibles de forma abierta, el número de diarios o revistas que publican en el idioma son limitados, y los libros que pueden ser encontrados en mayor medida no son representativos de un habla informal o actual del idioma.

Para trabajar sobre los MLs, principalmente para su uso en el reconocimiento de automático de voz y en la traducción automática, la herramienta más utilizada sea probablemente SRI Language Modeling (SRILM), y ha sido la utilizada en nuestro estudio. Esta permite realizar el *n-gram count*, filtrados, interpolados con órdenes inferiores o añadir *smoothing* para evitar romper nuestros modelos de lenguaje. Existen otras opciones como IRST Language Modeling Toolkit (IRSTLM), para trabajar con grandes MLs, o KenLM, pero menos utilizadas.

## *4 Marco experimental*

### 4.1 Descripción del sistema ASR

Actualmente nuestro reconocedor funciona con cortes de audio a una frecuencia de muestro ($f_s$) de 16 KHz y en formato *wav (WAVE form audio file format)*. Previo paso al reconocedor, se realiza una segmentación y una parametrización del corte de audio. En esta extracción de parámetros se obtiene los MFCC (*Mel Frequency Cepstral Coefficients*), los cuales son coeficientes para la representación del habla basados en la percepción auditiva humana.

Con respecto al modelado acústico, este utiliza una red neuronal profunda de 5 capas ocultas, con 1024 neuronas y función de activación RELU (Peddinti, Povey y Khudanpur, 2015). La red ha sido entrenada con material correspondiente a TC-STAR (Docío, Cardenal y García, 2006), con 79 horas de horas de habla en castellano; y Transcrigal (García et al., 2004), con 30 horas de habla en gallego.

Por último, el sistema usa modelos de lenguaje de trigramas entrenados con la herramienta SRILM. En el siguiente apartado se explica en mayor detalle cómo se han entrenado.

### 4.2 Entrenamiento de los MLs

El primer paso para la creación de los modelos de lenguaje ha sido la obtención de texto. Este texto es descargado de Internet y guardado de forma unificada en distintos archivos en función de su origen y características.

Antes del entrenamiento de los modelos, se realiza un primer filtrado con el objetivo de eliminar caracteres extraños, palabras en otros idiomas, errores del proceso de descarga y también de unificar formatos. En este paso la totalidad del texto es divida en oraciones separadas por saltos de línea. Una vez dividido, los signos de puntuación son eliminados (salvo el guión) ya que no aportan información al modelo que utilizará nuestro reconocedor.

El segundo filtrado realizado tiene el objetivo de reducir el tamaño del modelo en función de la probabilidad de que aparezcan las palabras, es decir, se utilizarán sólo las palabras del vocabulario creado que aparecen con una cierta probabilidad.

Una vez se ha preprocesado y filtrado el texto, es posible realizar el proceso de entrenamiento. Los modelos han sido entrenados usando la SRI Language Modeling Toolkit. Se han usado modelos de orden 3, es decir, trigramas. También se ha aplicado un discounting modificado de Kneser-Ney de Chen y Goodman, junto con una interpolación que hace que las estimaciones de probabilidad de orden 3 se interpolen con estimaciones de orden inferior (Stolcke, 2002).

Cada uno de estos modelos entrenados ha sido combinado con un modelo base de trigramas de forma equiprobable. Por último, se realiza la transcripción fonética con la herramienta COTOVIA (Campillo y Rodríguez, 2005) y se crea el grafo que utilizará el reconocedor con las herramientas proporcionadas por KALDI (Povey et al., 2011).

A continuación se explican cada uno de los modelos entrenados, y en la Tabla 1, se observan los principales parámetros de estos modelos. En ella se presenta el número de palabras en vocabulario, el tamaño del texto de entrenamiento, la media de palabras OOV y la perplejidad media sobre los cortes de audio analizados:

| ML | Nº palabras en vocabulario | Tamaño del texto de entrenamiento (en millones de palabras) | Porcentaje medio de palabras OOV sobre cortes analizados | Perplejidad media |
|---|---|---|---|---|
| **DOG** | 210.000 | 65 | 9,7 % | 789 |
| **DUVI** | 65.000 | 2,3 | 11,4 % | 616 |
| **Wikipedia** | 450.000 | 29 | 6,1 % | 703 |
| **BEPUB** | 400.000 | 317 | 6,8 % | 546 |
| **GEV** | 550.000 | 81 | 3,9 % | 728 |
| **CORGA** | 420.000 | 35 | 3,2 % | 582 |

Tabla 1: Datos sobre los MLs utilizados.

- **DOG**: Modelo entrenado únicamente con texto extraído del DOG (Diario Oficial de Galicia). Se trata de un ML de tamaño medio con texto representativo de un habla muy formal. Aunque este ML posee un tamaño considerable, presenta múltiples tecnicismos o formalismos en su texto, por lo que no se esperan grandes resultados.
- **DUVI**: Modelo entrenado con textos del DUVI (Diario de la Universidad de Vigo). Se trata de un ML pequeño, pero que posee un texto muy limpio y representativo, incluyendo multitud de palabras y expresiones actuales.
- **BEPUB**: Modelo entrenado con un conjunto de 5.000 novelas en castellano traducidas al gallego usando un traductor automático (Alegría et al., 2006). Se debe tener en cuenta que los resultados obtenidos pueden tener errores sistemáticos debidos a la traducción castellano-gallego. El potencial de este modelo consiste, más que en su vocabulario, el cual es cierto que puede no

presentar palabras o expresiones actuales, en la elevada confiabilidad que pueden presentar los n-gramas gracias a la gran cantidad texto sobre el que están entrenados.

- **Wikipedia**: Modelo entrenado con el material en gallego disponible en la Wikipedia. Se trata de un texto que puede reducir el número de palabras fuera de vocabulario, ya que presenta material actualizado y muy variado. Su principal desventaja reside en la falta de texto representativo del habla actual o de diálogos de habla espontánea, ya que está formado en su mayoría por definiciones.
- **Ghoxe + Eroski + Vieiros (GEV)**: Modelo entrenado con texto del diario digital Galicia Hoxe, publicado en Santiago de Compostela; con texto de la página de noticias de Eroski, la cual contiene texto de noticias variadas sobre temas de actualidad; y de Vieiros, otro diario digital editado íntegramente en gallego. Se trata de un corpus extenso, con material limpio y representativo, ya que posee noticias de actualidad de temática variada.
- **CORGA**: Modelo entrenado con textos del Corpus de Referencia del Gallego Actual (CORGA), integrado por distintos textos representativos correspondientes a libros, diarios, revistas, obras de teatro, material audiovisual y blogs. Se trata de un corpus de tamaño medio con texto muy cuidado y muy representativo.

A lo largo del estudio, tras haber realizado las primeras pruebas con los modelos que se acaban de describir, el siguiente paso (fase 2) fue realizar combinaciones de modelos. Las mezclas realizadas han sido las siguientes:

- **Mezcla 1**: uniendo todo el texto de CORGA, GEV y DUVI, y volviendo a entrenar un nuevo ML.
- **Mezcla 2**: combinando directamente los modelos de lenguaje ya entrenados de CORGA, GEV y DUVI en un nuevo ML. De esta forma se busca analizar las diferencias de resultados entre entrenar nuevos modelos con la totalidad del texto o mezclar los modelos ya entrenados.
- **Mezcla 3**: combinando los MLs ya entrenados de CORGA y BEPUB.
- **Mezcla 4**: combinando los MLs ya entrenados de CORGA, GEV, DUVI y BEPUB.

Los modelos DOG y Wikipedia han sido descartados en estas mezclas debido a los malos resultados obtenidos en la fase 1.

En la Tabla 2 se muestra, para cada una de las combinaciones de MLs, el número de palabras en vocabulario, el porcentaje medio de palabras fuera de vocabulario y la perplejidad media sobre los cortes de audio analizados:

| ML | Nº palabras en vocabulario | Porcentaje medio de palabras OOV sobre cortes analizados | Perplejidad media |
|---|---|---|---|
| **Mezcla 1** | 720.000 | 2,5 % | 627 |
| **Mezcla 2** | 730.000 | 2,5 % | 608 |
| **Mezcla 3** | 630.000 | 2,8 % | 627 |
| **Mezcla 4** | 900.000 | 2,3 % | 674 |

Tabla 2: Datos sobre combinaciones de MLs.

## 4.3 Corpus de análisis.

Las pruebas han sido realizadas sobre tres corpus con características diferentes.

### 4.3.1 Primer Corpus: Oralidad formal

Corpus con cortes de audio de oralidad formal correspondientes a lecturas literarias y discursos leídos y orales.

Se trata de 30 cortes con una duración media de 3:50 minutos por corte y una duración total de aproximadamente 115 minutos (cerca de 2 horas).

### 4.3.2 Segundo Corpus: Habla en noticiarios

Corpus con cortes de audio correspondientes a telediarios de la TVG (Televisión de Galicia). Presentan mezclas de habla espontánea y habla planeada y leída, pero con temas y vocabulario más actuales que en el primer corpus.

Está compuesto por 10 audios con una duración media de 34 minutos por corte y una duración total de 340 minutos (5 horas y 40 minutos).

### 4.3.3 Tercer Corpus: Habla en TED Talks

Corpus con cortes de audio correspondientes a charlas TED Talks en gallego. Estos presentan habla planeada pero no leída, y parte de habla espontánea.

Se trata de 10 cortes con una duración media de 16 minutos por audio y una duración total de 163 minutos (2 horas y 43 minutos).

## *5 Resultados experimentales*

Los siguientes resultados muestran la WER media obtenida con cada uno de los MLs. También se muestra el porcentaje medio de palabras fuera de vocabulario para cada uno de los marcos experimentales en función del ML analizado.

Para los tres marcos experimentales el proceso de análisis y extracción de los resultados será el siguiente: como primer paso (fase 1), se calcula la WER tras reconocimiento y se obtiene el número de palabras fuera de vocabulario usando los MLs “simples”; y como

segundo paso (fase 2), se analiza el efecto de combinar los MLs que mejores resultados han obtenido.

## 5.1 Resultados con MLs simples – fase 1

En la Tabla 3 se muestra la WER media obtenida en cada uno de los corpus de análisis en función del ML usado, así como el intervalo de confianza (IC) del 95% y la desviación típica.

Los mejores resultados los obtienen los MLs GEV y CORGA. Este último funciona especialmente bien dentro del Corpus 1, donde se consigue reducir en más de un 11 % la WER obtenida por el modelo DOG; mientras que GEV obtiene los mejores resultados en los Corpus 2 y 3, pero reduciendo la WER en menor medida (aproximadamente un 3 % en el Corpus 2 y 4 % en el Corpus 3, ambos con respecto al ML DOG).

| WER (%) | | DOG | DUVI | WIKI PEDIA | BEPUB | GEV | CORGA |
|---|---|---|---|---|---|---|---|
| Corpus 1 | Valor medio | 28,58 | 28,55 | 23,85 | 21,01 | 20,94 | **17,36** |
| | IC 95% | ± 1,98 | ± 2,16 | ± 1,83 | ± 2,01 | ± 1,84 | ± 1,84 |
| | Desv. típica | 5,50 | 6,05 | 5,1 | 5,63 | 5,14 | 5,14 |
| Corpus 2 | Valor medio | 25,08 | 24,77 | 24,1 | 24,5 | **22,13** | 23,5 |
| | IC 95% | ± 2,34 | ± 2,3 | ± 2,34 | ± 2,17 | ± 2,19 | ± 2,17 |
| | Desv. típica | 3,27 | 3,22 | 3,27 | 3,03 | 3,06 | 3,03 |
| Corpus 3 | Valor medio | 27,89 | 26,71 | 25,64 | 24,5 | **23,57** | 24,02 |
| | IC 95% | ± 4,76 | ± 4,46 | ± 5,09 | ± 5 | ± 4,96 | ± 5,32 |
| | Desv. típica | 6,66 | 6,23 | 7,11 | 6,99 | 6,94 | 7,44 |

Tabla 3: Resultados para cada ML.

## 5.2 Resultados con combinaciones de MLs – fase 2

Si nos fijamos en los resultados obtenidos con las mezclas de modelos presentados en la Tabla 4, para el Corpus 1 se observa que no se consigue reducir la WER media con ninguna combinación, pero sí que se consigue reducir el intervalo de confianza de los resultados. Los mejores resultados de combinaciones para el Corpus 2 son muy similares a los obtenidos con los modelos simples. Sólo para el Corpus 3 se consiguen mejorar ligeramente los resultados obtenidos, por lo que se puede concluir que las combinaciones de modelos no logran reducir de forma significativa la WER media obtenida, y sólo consiguen reducir ligeramente el intervalo de confianza y la desviación típica en los Corpus 1 y 2. La principal ventaja de los modelos combinados reside en que no sería necesario ir cambiando de modelo en función de la temática del audio, ya que presentan unos resultados más robustos en media frente a los tres corpus de estudio que cualquiera de los modelos simples.

| WER (%) | | Mezcla 1 | Mezcla 2 | Mezcla 3 | Mezcla 4 |
|---|---|---|---|---|---|
| Corpus 1 | Valor medio | 17,61 | **17,51** | 17,55 | 18,14 |
| | IC 95% | ± 1,76 | ± 1,71 | ± 1,78 | ± 1,77 |
| | Desv. típica | 4,92 | 4,78 | 4,98 | 4,95 |
| Corpus 2 | Valor medio | 22,3 | **22,22** | 24,14 | 23,57 |
| | IC 95% | ± 2,17 | ± 2,15 | ± 2,28 | ± 2,27 |
| | Desv. típica | 3,03 | 3,01 | 3,19 | 3,18 |
| Corpus 3 | Valor medio | 23,35 | **23,14** | 23,84 | 23,6 |
| | IC 95% | ± 5,32 | ± 5,26 | ± 5,19 | ± 5,21 |
| | Desv. típica | 7,44 | 7,26 | 7,26 | 7,28 |

Tabla 4: Resultados para las combinaciones de MLs.

Es interesante comparar los resultados obtenidos por la Mezcla 1 y la Mezcla 2. En estos se aprecia que la hora de combinar modelos, los valores más bajos de WER se obtienen cuando mezclan los modelos previamente entrenados por separado.

En la Figura 1 se muestra en perspectiva la evolución de la WER media para cada ML en función de cada uno de los Corpus estudiados. En ella podemos observar que la reducción de la WER conseguida en el Corpus 1 es mucho más evidente que para el resto de corpus:

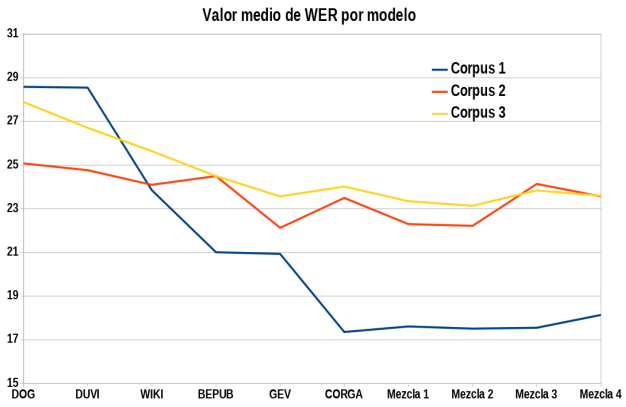


Figura 1: Evolución de la WER media para cada uno de los marcos experimentales.

Fijándonos en el porcentaje medio de palabras OOV, en la Figura 2 se observa su evolución para cada uno de los marcos experimentales. Como era de esperar, los valores más bajos de palabras OOV se obtienen con el modelo que combina más MLs (Mezcla 4). También se aprecia que los valores finales a los que se llega son similares en los tres marcos experimentales, mientras que claramente en el Corpus 1 es en el que la reducción de palabras OOV es mayor.

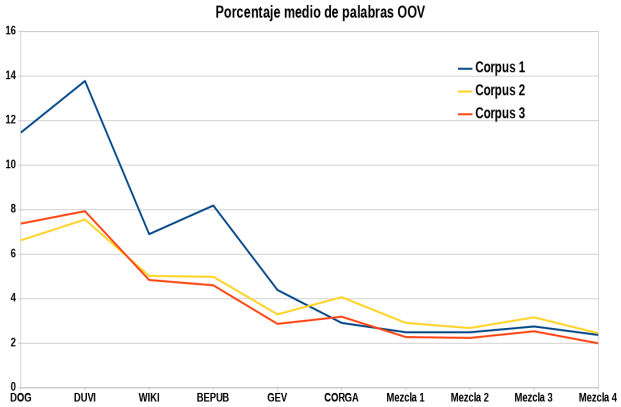


Figura 2: Evolución de las palabras OOV para cada uno de los marcos experimentales.

## 6 *Discusión*

Si analizamos los resultados obtenidos comparando cómo se consigue reducir el porcentaje de palabras OOV y la WER media obtenida para cada marco experimental, se obtiene lo representado en la Figura 3 (para el Corpus 2). En ella se observa una clara correlación entre el porcentaje de palabras OOV y el valor de WER obtenido, habiendo una progresión muy similar para las dos líneas.

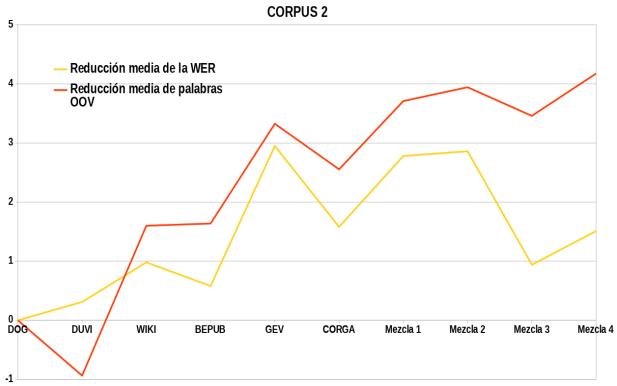


Figura 3: Comparación con DOG de la reducción de la WER y de las palabras OOV para el Corpus 2.

Por otra parte aunque para los tres corpus se observa esta correlación, también se concluye de los resultados obtenidos que reducir el porcentaje de palabras OOV no siempre es sinónimo de reducir la WER obtenida (véanse Figuras 1 y 2).

El comportamiento disimilar entre el Corpus 1 y los otros dos puede obedecer al diferente carácter de las muestras lingüísticas. El Corpus 1 está constituido mayoritariamente por textos leídos (lengua escrita), mientras que el corpus 2 presenta un número elevado de locutores, mezcla heterogénea de tipos de habla (lengua leída, declaraciones de diferentes hablantes, situaciones de ruido, mezcla de gallego y castellano, entre otras). El corpus 3, aunque más homogéneo, está constituido por discurso oral, y cabe hipotetizar sobre si la falta de respuesta de la WER a la suma de MLs (que si tienen efecto en la reducción de OOV, ver Figura 2) está relacionada con que los MLs son fundamentalmente modelos de lengua escrita.

Un análisis más detallado de los errores de reconocimiento puede llevar a reducir todavía más los porcentajes de WER. Una parte de los errores típicos está relacionado con la incorporación de modelos de gallego y castellano, necesarios para el reconocimiento de corpus en que las dos lenguas están presentes (como sucede en el Corpus 2), pero no en los restantes (Corpus 1 y 3). Mas debe tenerse en cuenta que una parte de los errores, al menos en los corpus de lengua oral (no leída), debe asumirse como inevitables, ya que obedecen a dudas y errores de pronunciación de los hablantes, desviaciones de formas, etc., que pueden aparecer marcados como errores, pero en los que el reconocedor en realidad acierta.

## 7 *Conclusiones y líneas futuras*

La importancia de la calidad de los corpus de entrenamiento con los que se crean los modelos de lenguaje ha quedado reflejada en los resultados presentados.

Se ha conseguido una bajada media de la WER de aproximadamente un 11 % para el corpus de entrenamiento 1, de un 3 % para el corpus 2 y de un 4,75 % para el corpus 3. Por otra parte, se ha comprobado cómo las combinaciones de modelos no logran mejorar de forma significativa los resultados.

También se ha observado que los resultados obtenidos para cada ML dependen claramente del marco experimental sobre el que se está trabajando, es decir, es complicado crear un ML que funcione de forma correcta cuando se varía la temática y características del audio. Como solución a esto, las combinaciones de MLs

obtienen buenos resultados en media frente a los tres corpus de test.

Y por último, la clara correlación (aunque no estrictamente directa) entre el porcentaje de palabras OOV y la WER obtenida evidencia la necesidad de mejorar o adaptar los vocabularios de nuestros modelos.

Como líneas futuras, en recientes estudios se comienza a utilizar las redes neuronales para el entrenamiento de modelos de lenguaje. Aunque se sigue utilizando el entrenamiento "clásico" basado en la predicción estadística como primer paso, se realiza un rescoring de los modelos usando redes neuronales. Los resultados de este rescoring muestran mejoras sobre los resultados experimentales.

Otras posibles líneas de investigación serían aumentar el orden de los n-gramas con los que se entrenan los modelos de lenguaje, puesto que la tecnología actual ya permite trabajar con tetragramas (4-gramas) o incluso pentagramas (5-gramas); o disponer de modelos basados en el habla para poder utilizarlos de forma opcional según el carácter del corpus a reconocer.

## *Agradecimientos*

El trabajo realizado está enmarcado en el proyecto del Plan Nacional TraceThem TEC2015-65345-P y en la red gallega TecAnDaLi ED431D 2016/011 financiada por la Xunta de Galicia. Asimismo se beneficia de las ayudas de la Xunta de Galicia de Grupos de Referencia Competitiva GRC2014/024 y Agrupación Estratéxica Consolidada de Galicia acreditación 2016-2019 y a la Unión Europa a través de los fondos FEDER. Se agradece al Instituto Ramón Piñeiro de la Xunta de Galicia el acuerdo de colaboración para la utilización del material del CORGA y su participación en el etiquetado de los corpus 2 y 3.

## *Referencias*

Peddinti, Vijayaditya, D. Povey y S. Khudanpur. 2015. A time delay natural network architecture for efficient modeling of long temporal context. En *Proceedings of INTERSPEECH*.

Stolcke, Andreas. 2002. SRILM An extensible language modeling toolkit. En *Proceedings of the International Conference on Statistical Language Processing*. Denver, Colorado.

García, Carmen, J. Tirado, L. Docío y A. Cardenal. 2004. Transcrigal: A bilingual system for automatic indexing of broadcast news. *IV International Conference on Language Resources and Evaluation*.

Docío, Laura, A. Cardenal y C. García. 2006. TC-STAR 2006 automatic speech recognition evaluation: The uvigo system. En *Proc. Of TC-STAR Workshop on Speech-to-Speech Translation*. ELRA, París, France.

Jurafsky, Daniel, y J.H. Martin. 2008. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.

Vicente, Marta, C. Barros, F. Peregrino, F. Agulló y E. Lloret. 2015. La generación de lenguaje natural: análisis del estado actual. Computación y Sistemas. Volumen: 9, n.º 4.

Povey, Daniel, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlícek, Y. Quian, P. Schwarz, J. Silovský, G. Stemmer y K. Veselý. 2011. The Kaldi Speech Recognition Toolkit. En *ASRU Workshop*.

Campillo, Francisco y E. Rodríguez. 2005. Evaluación del modelado acústico y prosódico del sistema de conversión texto-voz Cotovía. En *Procesamiento del Lenguaje Natural. Volumen 35, páginas 5-12*.

Alegría, Iñaki, I. Arantzabal, M. Forcada, X. Gómez, L. Padró, J.R. Pichel y J. Waliño. 2006. OpenTrad: Traducción automática de código abierto para las lenguas del estado Español. En *Procesamiento del Lenguaje Natural*. Volumen: 37, páginas 356-358.

Mikolov, Tomas, S. Kombrink, A. Deoras, L. Bruget y J. Cernocky. 2011. Rnnlm-recurrent neuronal network language modeling toolkit. En *Proc. of ASRU Workshop*.

Xu, Hainan, T. Chen, D. Gao, Y. Wang, K. Li, N. Goel, Y. Carmiel, D. Povey y S. Khudanpur. 2018. A pruned rnnlm lattice-rescoring algorithm for automatic speech recognition. En *ICASSP*.

Sundermeyer, Martin, Z. Tüske, R. Schlüter y H. Ney. 2014. Lattice decoding and rescoring with long-span neural network language models. En *Fifteenth Annual Conference of the International Speech Communication Association*.

Chen, Xie, X. Liu, A. Ragni, Y. Wang y M. Gales. 2017. Future word contexts in neuroal network language models. ArXiv preprint arXiv:170805592.

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 83-89 recibido 04-04-2018 revisado 29-04-2018 aceptado 04-05-2018

# Bi-modal annoyance level detection from speech and text

## *Detección del nivel de enfado mediante un sistema bi-modal basado en habla y texto*

**Raquel Justo, Jon Irastorza, Saioa Pérez, M. Inés Torres**
Universidad del País Vasco UPV/EHU. Sarriena s/n. 48940 Leioa. Spain
{raquel.justo,manes.torres}@ehu.eus

**Abstract:** The main goal of this work is the identification of emotional hints from speech. Machine learning researchers have analysed sets of acoustic parameters as potential cues for the identification of discrete emotional categories or, alternatively, of the dimensions of emotions. However, the semantic information gathered in the text message associated to its utterance can also provide valuable information that can be helpful for emotion detection. In this work this information is included within the acoustic information leading to a better system performance. Moreover, it is noticeable the use of a corpus that include spontaneous emotions gathered in a realistic environment. It is well known that emotion expression depends not only on cultural factors but also on the individual and on the specific situation. Thus, the conclusions extracted from the present work can be more easily extrapolated to a real system than those obtained from a classical corpus with simulated emotions.
**Keywords:** speech processing, semantic information, emotion detection on speech, annoyance tracking, machine learning

**Resumen:** El principal objetivo de este trabajo es la detección de cambios emocionales a partir del habla. Diferentes trabajos basados en aprendizaje automático han analizado conjuntos de parámetros acústicos como potenciales indicadores en la identificación de categorías emocionales discretas o en la identificación dimensional de las emociones. Sin embargo, la información semántica recogida en el mensaje textual asociado a cada intervención, puede proporcionar información valiosa para la detección de emociones. En este trabajo se combina la información textual y acústica dando lugar a mejoras en el rendimiento del sistema. Es importante recalcar por otra parte, el uso de un corpus que incluye emociones espontáneas recogidas en un entorno realista. Es bien sabido que la expresión de la emoción depende no solo de factores culturales si no también de factores individuales y de situaciones particulares. Por lo tanto, las conclusiones extraídas en este trabajo se pueden extrapolar más fácilmente a un sistema real que aquellas obtenidas a partir de un corpus clásico en el que se simula el estado emocional.
**Palabras clave:** procesamiento del habla, información semántica, reconocimiento emocional en el habla, rastreo del enfado, aprendizaje automático

### *1 Introduction*

The detection of emotional status has been widely studied in the last decade within the machine learning framework. The goal of researchers is to be able to recognise emotional information from the analysis on voice, language, face, gestures or ECG (Devillers, Vidrascu, y Lamel, 2005). One of the main important challenges that need to be faced in this area is the need of supervised data, i.e. corpora including human data annotated with emotional labels (Devillers, Vidrascu, y Lamel, 2005) (Vidrascu y Devillers, 2005) and this is not a straightforward task due to the subjectivity of emotion perception by humans (Devillers, Vidrascu, y Lamel, 2005) (Eskimez et al., 2016). Many works considered corpora that consist of data from professional actors simulating the emotions to be analyzed. However, it usually leads to poor

ISSN 1135-5948. DOI 10.26342/2018-61-9

results due to many factors like the differences among the real situations the detection system has to deal with and the emotional status picked up in the corpus. Moreover, the selection of valuable data including spontaneous emotions depends on the goals of the involved research and it is difficult to find an appropriate corpus that matches the specific goal of each task.

Focussing on emotion identification from speech and language a wide range of potential applications and research objectives can be found (Valstar et al., 2014) (Wang et al., 2015) (Clavel y Callejas, 2016). Some examples are early detection of Alzheimer's disease (Meilán et al., 2014), the detection of valency onsets in medical emergency calls (Vidrascu y Devillers, 2005) or in Stock Exchange Customer Service Centres (Devillers, Vidrascu, y Lamel, 2005). Emotion recognition from speech signals relies on a number of short-term features such as pitch, additional excitation signals due to the non-linear air flow in the vocal tract, vocal tract features such as formants (Wang et al., 2015) (Ververidis y Kotropoulos, 2006), prosodic features (Ben-David et al., 2016) such as pitch loudness, speaking rate, rhythm, voice quality and articulation (Vidrascu y Devillers, 2005) (Girard y Cohn, 2016), latency to speak, pauses (Justo et al., 2014) (Esposito et al., 2016), features derived from energy (Kim y Clements, 2015) as well as feature combinations, etc. Regarding methodology, statistical analysis of feature distributions has been traditionally carried out. Classical classifiers such as the Bayesian or SVM have been proposed for the identification of emotional characteristics from speech signals. The model of continuous affective dimensions is also an emerging challenge when dealing with continuous rating of emotion labelled during real interaction (Mencattini et al., 2016). In this approach recurrent neural networks have been proposed to integrate contextual information and then predict emotion in continuous time to just deal with arousal and valence (Wollmer et al., 2008) (Ringeval et al., 2015).

When regarding text, there are numerous works dealing with sentiment analysis whose application domains range from business to security considering well-being, politics or software engineering (Cambria, 2016). However, there are few works considering the recognition of specific emotions such as joy, love or anger (Medeiros y van der Wal, 2017; Gilbert y Karahalios, 2010; Marsden y Campbell, 2012). Moreover, it seems reasonable to think that the combination of acoustic and textual information might lead to improve emotion recognition systems performance. However, although there are plenty of research articles on audio-visual emotion recognition, only a few research works have been carried out on multimodal emotion recognition using textual clues with visual and audio modality (Eyben et al., 2010; Poria et al., 2016).

In this work we deal with a problem proposed by a Spanish company providing customer assistant services through the telephone (Justo et al., 2014). They want to automatically detect annoyance rates during customer calls for further analysis, which is a novel and challenging goal. Their motivation is to verify if the policies applied by operators to deal with annoyed and angry customers lead to shifts in customer behavior. Thus an automatic procedure to detect those shifts will allow the company to evaluate their policies through the analysis of the recorded audios. Moreover, they were interested in providing this information to the operators during the conversation. As a consequence this work is aimed at detecting different levels of annoyance during real phone-calls to Spanish complain services. Mainly, we wanted to analyse the effect of including textual information into the annoyance detection system based on acoustic signals.

The paper is organised as follows, Sec. 2 describes the previous work carried out to solve the presented problem with the specific dataset we are dealing with. In Sec. 3 the annotation procedure in terms of speech signal and text is described and Sec. 4 details the experiments carried out and the obtained results. Finally, Sec. 5 summarises the concluding remarks and future work.

## *2 Dataset and previous work*

The Spanish call center company involved in this work offers customer assistance for several phone, tv and internet service providers. The customer complaint services of these companies receive a certain number of phone-calls from angry or annoyed customers. But the way of expressing annoyance is not the same for all the customers. Some of them are furious and shout; others speak

quickly with frequent and very short micropauses but do not shout (Justo et al., 2014), others seems to be more fed-up than angry; others feel impotent after a number of service failures and calls to the customer service. The dataset for this study consisted of seven conversations between customers and the call-centre operators that were identified and selected by experienced operators. All the selected customers were very angry with the service provider because of unsolved and repeated service failures that caused serious troubles to them. In a second step each recording was named according to the particular way the customer expresses his annoyance degree. Thus, call-center operators qualified the seven subjects in conversations as follows: *Disappointed, Angry (2 records), Extremely angry, Fed-up, Impotent and annoyed in disagreement*. All these feelings correspond to the different ways the customer in the study expressed their annoyance with the service provided. More specifically they correspond to the way the human operators perceived customer feelings. The duration of the conversations was 42s, 42s, 35s, 16m20s, 1m08s, 1m02s and 1m35s respectively, resulting in a total of 22.1 minutes.

In a previous work (Irastorza y Torres, 2016) the different records were manually annotated. Two members of the research group acted as expert annotators. They first identified customer speech segments, agent speech segments and overlapping segments. Only intelligible customer speech segments were considered for the experiments. In a second step, annotators were asked to identify the changes in the degree of perceived emotion in each recording using zero for neutral or very low, one for medium and two for high degree. They were asked to mark time steps where they perceived a change in the degree of expression and then label each segment with the corresponding perceived level. The annotator agreement was high in the identification of the time steps where they perceived changes in the degree of expression. Then, just one of the two annotations was chosen to fix segments bounds. However, the procedure resulted in a significant level of disagreements when regarding the label given to each step. Thus, most frequent disagreements were considered as new levels in the proposed scale of annoyance expression. The resulting set of categories consists of five degrees defined as follows: *very low* agreed by annotators, *low*, which corresponds to a low-medium disagreement, *medium* agreed by annotators, *high*, which corresponds to a medium-high disagreement and *very high* agreed by annotators. Less frequent disagreements were not considered. The right side of Table 1 (SPEECH-BASED) shows the final number of segments identified for each audio file and annoyance level.

An automatic classification was carried out in (Irastorza y Torres, 2016) using acoustic parameters extracted from the audio files. The acoustic signal was divided into 20 ms overlapping windows (frames) from which a set of features was extracted. The classification procedure was carried out over those frames. A combination of Intensity and intensity-based suprasegmental features along with LPC coefficients achieved the highest frame classification accuracies for all the expressions of annoyance analysed. The obtained results validated the annotation procedure and also showed that shifts in customer annoyance rates could be potentially tracked during phone calls.

## *3 Text annotation process*

The transcription of the utterances provides additional information related to the semantics, language style, among others, that cannot be found in acoustics and might help in the detection of annoyance or other emotional categories. Thus, in this work the text associated with the utterances was considered and annotated to be included in the classification process.

### 3.1 Transcription

The audio files described in the previous section were transcribed in order to use text as an additional information source. The transcriptions were carried out making use of *Praat* (Boersma y Weenink, 2016) software tool. The segments obtained from the aforementioned annotation, in terms of time steps, were also employed here and only those intelligible customer speech segments were transcribed.

In this case, given that there is no any ambiguity in the transcription task, only one member of the research group listened to all the audios, segment-to-segment, and provided the corresponding transcriptions.

### 3.2 Annotation

Once the transcription was obtained, the segments were annotated with an emotional label extracted from the text. In order to obtain a label that only considers textual information, the labelling was carried out by an annotator that was not involved in the transcription process. In this way, the annotator did not listened to the audio previously and was only focused on text. Two members of the research group carried out the annotation independently. The same levels of anger employed in the acoustic annotation was also considered here: zero for very low, one for medium and neutral and two for high degree of anger. Given the ambiguity associated to this annotation procedure a method was also needed to deal with disagreement among the two annotators. Thus, the set of categories consisting of five degrees defined as *very low* agreed by annotators, *low*, which corresponds to a low-medium disagreement, *medium* agreed by annotators, *high*, which corresponds to a medium-high disagreement and *very high* agreed by annotators was also employed here.

An analysis of the labeled segments showed that one of the two annotators always used a higher level of anger than the other one. Therefore, one/zero or two/one labels appeared more frequently than zero/one and one/two labels. It is noteworthy, that segments labeled as zero/two and two/zero (strong disagreement) were very unfrequent. The left side of Table 1 (TEXT-BASED) shows the number of segments for each audio file.

### 3.3 Analyzing the annoyance

People's behaviour can be very different when regarding the way in which they express annoyance. Some people vary the pitch or intensity of their utterance very easily when they are upset, while others keep these variables unaltered and emphasize the meaning of their message. Thus, both issues have to be considered to identify annoyance rates. For instance, in the audio *Very angry* there is a speech segment where the speaker says: *"sorry, sorry but you are the impolite"*. Regarding the acoustic annotation it was labeled with a low annoyance rate due to an unaltered acoustic signal, whereas the text annotation indicated a high annoyance rate, because its meaning clearly denotes that the user is upset.

Table 1 shows that significant differences can be observed in text-based and speech-based annotations. Mainly, it seems that higher anger rates are associated to speech-based annotations: there are 57 very high segments annotated in speech versus 13 in text, while at the other end, there are 5 very low segments annotated in speech versus 78 in text. An example of this behaviour would be a segment of the audio "Angry 2" where the speaker says: *"all things you will tell me, they have already told me"*. This fragment, was annotated as high in speech annotation and low in textual annotation. It seems, according to the obtained results, that annoyed people tend to vary the features of their utterances (pitch, intensity, etc.) more easily while keeping the meaning of their messages unaltered. It might be because changes in the acoustic signal occur in an spontaneous and involuntary way and it does not happen when regarding the content of the message. Thus, only when the annoyance is kept in a longer period of time annoyance signals appear in text messages. The combination between speech annotator information and text annotator information, provides the chance to complement the information provided from isolated sources, leading to nuanced results, in cases of discrepancy.

A bi-modal annoyance recognizer, combining acoustic and text, was developed in this work and the results provided by its evaluation are given in section 4.

## *4 Experimental Results*

The experiments carried out aimed at analyzing the validity of the assumptions made in the annotation procedure and also the performance when combining acoustic features and textual information. We used the Naive Bayes Classifier (NB) and the Support Vector Machine (SVM) which had already proven their efficiency classifying emotional hints (Irastorza y Torres, 2016). To this end we shuffled the set of frames of each audio file and then split this set into a training and a test set that included 70 % and 30 % of frames, respectively. We used the frame classification accuracy as evaluation metric.

Two series of classification experiments were carried out in order to include textual information in different ways. Firstly, we chose a combination of acoustic (Linear Predic-

| | TEXT-BASED | | | | | SPEECH-BASED | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Activation level category | | | | | | | | | | |
| | v_low | low | medium | high | v_high | v_low | low | medium | high | v_high | Total |
| Disappointed | 1 | 1 | 5 | 2 | 0 | 0 | 6 | 1 | 2 | 0 | 9 |
| Angry 1 | 0 | 4 | 0 | 1 | 1 | 0 | 0 | 4 | 2 | 0 | 6 |
| Angry 2 | 0 | 1 | 2 | 4 | 1 | 1 | 2 | 1 | 4 | 0 | 8 |
| Very Angry | 65 | 61 | 53 | 35 | 9 | 0 | 95 | 26 | 45 | 57 | 223 |
| Fed-up | 6 | 2 | 4 | 1 | 2 | 0 | 3 | 11 | 1 | 0 | 15 |
| Impotent | 6 | 3 | 0 | 0 | 0 | 4 | 4 | 1 | 0 | 0 | 9 |
| All | 78 | 72 | 64 | 43 | 13 | 5 | 110 | 44 | 54 | 57 | 270 |

Table 1: Number of segments for each audio file.

tion Coefficients, LPC) and textual (labels provided in the text annotations) features. The labels obtained from the speech annotations were employed to train the classifiers. In a second stage only LPC acoustic features were considered, but in this case the labels provided to train the classifiers were those obtained from the text annotation. We aimed at analysing the behaviour of the selected sets of features and annotation schemes when classifying frames into the categories that represent the customer annoyance degree in each particular call. Moreover, speaker dependent and speaker independent experiments were also considered. In speaker dependent experiments only frames obtained from a specific speaker (an audio file) were involved in the classification process (training and test).

### 4.1 First series of experiments

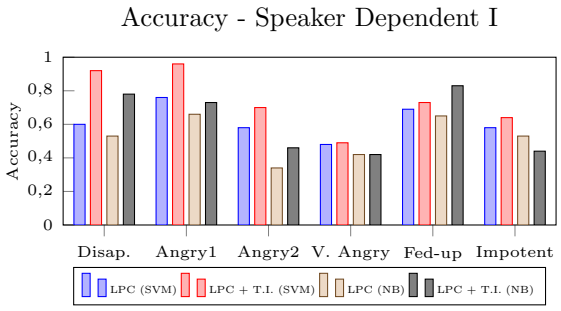


Figure 1: Comparison of SVM and NB frame classification approaches. Bar graphs show the frame identification accuracies based on the sets of features selected for the first series of experiments. The two bar graphs on the left side of each audio correspond to results obtained by SVM classifier whereas the ones on the right side correspond to the results obtained by NB.

In this first set of frame classification experiments, Figure 1 confirms that Linear Prediction Coefficients plus textual information classification outperforms Linear Prediction Coefficients classification in both SVM and NB models for speaker dependent. For instance, we can see an accuracy improvement up to 0.3 in the *Disappointed* audio when combining acoustic parameters plus textual information using SVM classifier. Equally, we carried out speaker independent experiments and the results also showed better performace using acoustic features plus textual information, improving the accuracy from 0.46 in to 0.52. Looking at the results it seems that there is information within text, that is missing in the acoustic signal, that could be useful for detecting annoyance rate.

### 4.2 Second series of experiments

Then a second set of frame classification experiments was carried out using both SVM and Naive Bayes models. This series was aimed at evaluating speaker dependent/independent model using acoustic parameters with textual labeling.

Figure 2 shows that the classification based on acoustic features (LPC) along with the use of labels based on the text annotation provides lower accuracy values. This loss of accuracy reinforces the idea that cognitive processing diverges depending on which senses are involved in the annotation process. On the other hand, speaker independent results showed also worse performance, since accuracy result did not achieve more than 0.31.

## *5 Conclusions and future work*

In conclusion, two main ideas resulted from the experiments. On the one hand, the possibility of joining textual and acoustic information in order to predict annoyance rates was explored and validated. The experiments show that the inclusion of labels extracted from text as a feature improve the classification accuracy. However, using acoustic fea-

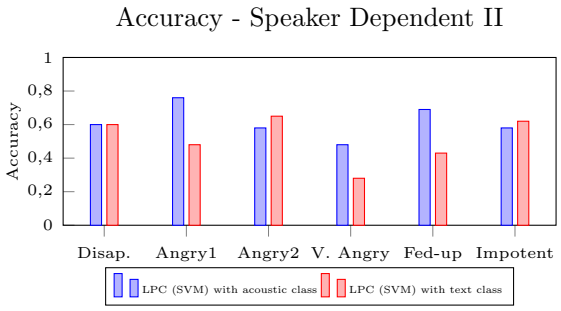


Figure 2: Comparison of SVM and NB frame classification approaches. Bar graphs show the frame identification accuracies based on the sets of features selected for the first second of experiments. The bar graph on the left side of each audio correspond to results obtained by SVM classifier using the class based on the text whereas the right graph correspond to the results obtained using the class based on the speech.

tures with text-based annotation does not provide a good system performance. Acoustic features are linked to acoustic signal, and that is why categories based on semantic analysis are meaningless.

Moreover, the use of a corpus that include spontaneous emotions gathered in a realistic environment leads to an easy extrapolation of the obtained results to a real system.

For further work we propose to explore alternative ways of integrating textual information along with deep learning based classification paradigms.

## *References*

Ben-David, B. M., N. Multani, V. Shakuf, F. Rudzicz, y P. H. H. M. van Lieshout. 2016. Prosody and semantics are separate but not separable channels in the perception of emotional speech: Test for rating of emotions in speech. *Journal of Speech, Language, and Hearing Research*, 59(1):72–89.

Boersma, P. y D. Weenink. 2016. Praat: doing phonetics by computer. Software tool, University of Amsterdam. version 6. 0.15.

Cambria, E. 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107, Mar.

Clavel, C. y Z. Callejas. 2016. Sentiment analysis: From opinion mining to human-agent interaction. *IEEE Transactions on Affective Computing*, 7(1):74–93, Jan.

Devillers, L., L. Vidrascu, y L. Lamel. 2005. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407 – 422. Emotion and Brain.

Eskimez, S. E., K. Imade, N. Yang, M. Sturge-Apple, Z. Duan, y W. Heinzelman. 2016. Emotion classification: how does an automated system compare to naive human coders? En *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, páginas 2274–2278, March.

Esposito, A., A. M. Esposito, L. Likforman-Sulem, M. N. Maldonato, y A. Vinciarelli, 2016. *Recent Advances in Nonlinear Speech Processing*, capítulo On the Significance of Speech Pauses in Depressive Disorders: Results on Read and Spontaneous Narratives, páginas 73–82. Springer International Publishing, Cham.

Eyben, F., M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, y R. Cowie. 2010. On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues. *Journal on Multimodal User Interfaces*, 3(1):7–19, Mar.

Gilbert, E. y K. Karahalios. 2010. Widespread worry and the stock market. En *ICWSM 2010 - Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, páginas 58–65.

Girard, J. M. y J. F. Cohn. 2016. Automated audiovisual depression analysis. *Current Opinion in Psychology*, 4:75 – 79.

Irastorza, J. y M. I. Torres. 2016. Analyzing the expression of annoyance during phone calls to complaint services. En *2016 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, páginas 000103–000106, Oct.

Justo, R., O. Horno, M. Serras, y M. I. Torres. 2014. Tracking emotional hints in spoken interaction. En *Proc. of VIII Jornadas en Tecnología del Habla and IV Iberian SLTech Workshop (IberSpeech 2014)*, páginas 216–226.

Kim, J. C. y M. A. Clements. 2015. Multimodal affect classification at various temporal lengths. *IEEE Transactions on Affective Computing*, 6(4):371–384, Oct.

Marsden, P. V. y K. E. Campbell. 2012. Reflections on conceptualizing and measuring tie strength. *Social Forces*, 91(1):17–23.

Medeiros, L. y C. N. van der Wal. 2017. An agent-based model predicting group emotion and misbehaviours in stranded passengers. En E. Oliveira J. Gama Z. Vale, y H. Lopes Cardoso, editores, *Progress in Artificial Intelligence*, páginas 28–40, Cham. Springer International Publishing.

Meilán, J. J. G., F. Martínez-Sácnhez, J. Carro, D. E. López, L. Millian-Morell, y J. M. Arana. 2014. Speech in alzheimer?s disease: Can temporal and acoustic parameters discriminate dementia? *Dementia and Geriatric Cognitive Disorders*, 37(5-6):327–334.

Mencattini, A., E. Martinelli, F. Ringeval, B. Schuller, y C. D. Natlae. 2016. Continuous estimation of emotions in speech by dynamic cooperative speaker models. *IEEE Transactions on Affective Computing*, PP(99):1–1.

Poria, S., I. Chaturvedi, E. Cambria, y A. Hussain. 2016. Convolutional mkl based multimodal emotion recognition and sentiment analysis. En *2016 IEEE 16th International Conference on Data Mining (ICDM)*, páginas 439–448, Dec.

Ringeval, F., F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, y B. Schuller. 2015. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, 66:22 – 30.

Valstar, M., B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, y M. Pantic. 2014. Avec 2014: 3d dimensional affect and depression recognition challenge. En *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, AVEC '14, páginas 3–10, New York, NY, USA. ACM.

Ververidis, D. y C. Kotropoulos. 2006. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162 – 1181.

Vidrascu, L. y L. Devillers. 2005. Detection of Real-Life Emotions in Call Centers. En *Proceedings of INTERSPEECH'05: the 6th Annual Conference of the International Speech Communication Association*, páginas 1841–1844, Lisbon, Portugal. ISCA.

Wang, K., N. An, B. N. Li, Y. Zhang, y L. Li. 2015. Speech emotion recognition using fourier parameters. *IEEE Transactions on Affective Computing*, 6(1):69–75, Jan.

Wollmer, M., F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, y R. Cowie. 2008. Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. páginas 597–600, 9.

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 91-98 recibido 05-04-2018 revisado 30-04-2018 aceptado 04-05-2018

# Análisis de errores de pronunciación y fluidez en la lectura oral en un corpus de habla leída de aprendices españoles de inglés como lengua extranjera

## *Analysis of pronunciation errors and oral reading fluency in a read corpus of Spanish learners of English as a foreign language*

**Patricia Elhazaz Walsh**
Universidad CEU San Pablo, Urbanización Montepríncipe, 28925 Alcorcón, Madrid
pelhazaz@ceu.es

**Resumen:** El objetivo del presente estudio ha sido recopilar un corpus de aprendices de habla leída para analizar los errores de pronunciación y medir la fluidez en la lectura oral (FLO) en estudiantes españoles de inglés como lengua extranjera. El corpus está compuesto de 6 horas de habla (42.230 palabras) producida por 117 estudiantes españoles. Las muestras de habla fueron segmentadas y transcritas ortográficamente. Se midió la FLO calculando el número de palabras correctas por minuto (pcpm). Los análisis estadísticos llevados a cabo muestran diferencias significativas entre grupos en medidas de FLO (***, p<0,001). Además, se hizo una clasificación y transcripción fonética de 3.522 errores de pronunciación anotando los distintos tipos de error (inserción, sustitución y elisión). Presentamos los errores de pronunciación más frecuentes que podrían ser corregidos en intervenciones educativas para mejorar la fluidez en la lectura oral y la pronunciación de estudiantes españoles en inglés como lengua extranjera.
**Palabras clave:** Corpus de habla no nativa, inglés como LE, adquisición de pronunciación en una L2/LE, fluidez en la lectura oral

**Abstract:** This study was aimed at compiling a learner corpus of read speech to analyze pronunciation errors and measure oral reading fluency (ORF) in Spanish learners of English as a foreign language (EFL). The corpus features 6 hours of read speech (42,230 words) recorded from 117 Spanish students. The speech data was segmented, and transcribed at the orthographic level using Transcriber software. Oral reading fluency was measured by calculating the number of words read correctly per minute (wcpm). Analyses of variance revealed significant differences between groups on ORF scores (***, p<0.001). Additionally, 3,522 pronunciation errors were transcribed at the phonetic level and classified following an encoding system that identified the types of errors (insertion, substitution and deletion). We report on the most frequent types of pronunciation errors that could be addressed in instructional interventions to help Spanish students improve their oral reading fluency and pronunciation in English as a foreign language.
**Keywords:** Non-native speech corpora, English as a FL, L2/FL pronunciation acquisition, oral reading fluency

## *1 Introducción*

La fluidez en la lectura oral (FLO) ha sido definida como la habilidad para leer un texto de manera precisa, con una velocidad de habla natural y con la expresión adecuada (National Reading Panel, 2000). La mayor parte de los estudios están de acuerdo en que la fluidez en la lectura oral está compuesta por la precisión (número de palabras leídas de manera correcta por minuto), la velocidad (número de palabras leídas por minuto) y la expresividad. De estos tres componentes, la velocidad y la precisión en el reconocimiento de palabras han sido los más estudiados como indicadores principales de fluidez en la lectura oral de palabras (Breznitz, 2006). A pesar de que la FLO no mide la

ISSN 1135-5948. DOI 10.26342/2018-61-10 

comprensión lectora de manera directa, la correlación entre ambas ha sido establecida en numerosos estudios (Perfetti, 1985; Fuchs et al., 2001; Jenkins et al., 2003).

Por otra parte, distintos estudios indican que la FLO en una segunda lengua (L2) o lengua extranjera (LE) es generalmente más lenta en desarrollarse que en una lengua materna (L1) debido, en gran medida, a la falta de automaticidad en las destrezas de reconocimiento de palabras (National Reading Panel, 2000). Una de las mayores dificultades a las que se enfrentan los lectores de una L2 o LE es la diferencia en la profundidad de los códigos alfabéticos entre su lengua materna y la lengua meta. Dicha diferencia puede llevar al lector a aplicar de manera incorrecta las reglas de conversión grafema-fonema, lo cual puede interferir con la descodificación precisa de palabras.

Según la hipótesis de la profundidad ortográfica, los hablantes de una lengua transparente como el español pueden experimentar dificultades para descodificar palabras en una lengua opaca como el inglés, debido a la naturaleza irregular de las correspondencias entre grafía y sonido en esta lengua (Basetti, 2009).

Este estudio tiene como objetivo proporcionar datos sobre el estado de la fluidez en la lectura oral (expresado en número de palabras correctas leídas por minuto, pcpm) y estudiar los errores de pronunciación más frecuentes producidos por alumnos españoles en la lectura del inglés como LE. Para ello, se diseñó y recogió un corpus de habla leída compuesto por 360 sesiones de lectura de un minuto de duración. Los datos obtenidos fueron empleados para entrenar un reconocedor de habla infantil integrado en un sistema de evaluación automática de la fluidez en la lectura oral (Bolaños et al., 2012).

Este artículo se estructura de la siguiente manera: la sección 2 contiene la descripción la metodología utilizada para la creación del corpus. El apartado 3 presenta los resultados de la fluidez en la lectura oral, y los errores de pronunciación más frecuentes. Por último, se discuten los resultados y conclusiones en el apartado 4.

# *2 Metodología*

## 2.1 Participantes

El grupo objeto de estudio está compuesto por 117 estudiantes españoles de 5º y 6º de Educación Primaria (E.P.) y 1º de Educación Secundaria (E.S.) Las grabaciones fueron recogidas en dos colegios bilingües concertados de la Comunidad de Madrid. En ambos colegios existía un modelo bilingüe de enseñanza de inglés implantado desde los primeros cursos de Educación Infantil. En la tabla 1 se muestra la distribución de los participantes del estudio.

| Curso | Edad | Chicos | Chicas | Total |
|---|---|---|---|---|
| 5º curso E. P. | 10-11 | 17 | 19 | 36 |
| 6º curso E. P. | 11-12 | 23 | 18 | 41 |
| 1º curso E. S. | 12-13 | 26 | 14 | 40 |
| Total | | 66 | 51 | 117 |

Tabla 1: Descripción de los participantes en el estudio

## 2.2 Instrumentos y procedimiento

Los textos utilizados para crear nuestro corpus corresponden a la prueba de fluidez en la lectura oral de DIBELS (The Dynamic Indicators of Basic Early Literacy Skills) diseñada para medir la precisión y fluidez en la lectura de textos (Good y Kaminsky, 2001). DIBELS es un conjunto de pruebas estandarizadas elaboradas por investigadores de la Universidad de Oregón para evaluar los cinco componentes básicos de la destreza lectora. Dicha prueba consiste evaluar la lectura en voz alta durante un minuto utilizando historias que no hayan sido leídas antes por los estudiantes.

Después de analizar tanto la gramática como el vocabulario de las historias de DIBELS y compararlos con la programación curricular de inglés como lengua extranjera para cada uno de los cursos estudiados, decidimos utilizar las historias correspondientes al primer y segundo curso de Educación Primaria en Estados Unidos con nuestros alumnos españoles de último ciclo de Primaria y primer curso de Educación Secundaria. Se utilizó un total de 48 historias distintas, compuestas por un total de 42.230 palabras y 1.462 palabras únicas.

### 2.2.1 Grabaciones

Las grabaciones se recogieron con un micrófono Sennheisser con cancelador de ruido

a una calidad de 16 KHz (16 bits por muestra) y los archivos de audio se almacenaron en tres formatos: MP3, WAV y RAW. Se recopiló un total de 6 horas de habla. Cada alumno leyó una media de tres historias distintas de un minuto de duración.

### 2.2.2 Transcripción

Todos los textos grabados fueron transcritos ortográficamente y alineados con la señal sonora con el programa de software Transcriber (Barras et al. 2001). Se utilizó la ortografía convencional para la representación de las palabras en la transcripción de nuestro corpus oral. Se anotaron todas las pausas llenas y las repeticiones o reparaciones. Las pausas llenas son aquellos elementos utilizados en los momentos en que el hablante duda en el momento de la producción del habla (por ej. *ehh*). Todas las transcripciones ortográficas anotadas están disponibles en formato XML. En la tabla 2 se muestra el total de palabras transcritas por curso.

| Curso | nº palabras transcritas |
|---|---|
| 5º curso E. P. | 13.333 |
| 6º curso E. P. | 14.089 |
| 1º curso E. S. | 14.808 |
| Total | 42.230 |

Tabla 2: Total de palabras transcritas por curso

### 2.2.3 Anotación de errores

Cada una de las transcripciones ortográficas fue analizada y los errores en la lectura anotados para medir el número correcto de palabras leídas por minuto (pcpm).

Dicho valor se obtiene sumando el número total de palabras leídas por minuto y restando las palabras leídas de manera errónea, las pausas llenas y las reparaciones. La tasa de lectura expresada en palabras correctas leídas por minuto se ha convertido en un estándar ampliamente aceptado en la comunidad científica para medir la fluidez en la lectura oral (Good et al., 2001). A continuación, se llevó a cabo una clasificación y análisis fonológico de los errores en la lectura. Para ello, se creó una tipología de errores de pronunciación. Los errores en la lectura se clasificaron analizando los procesos fonológicos subyacentes al error. Dichos procesos se originan en las dificultades que experimentan los aprendices en la pronunciación de la LE dando lugar a realizaciones incorrectas que a su vez nos pueden ayudar a explicar la causa del error. Tradicionalmente se distinguen tres procesos fonológicos: inserción, omisión y sustitución. Llisterri (2002, p.98) distingue tres tipos de errores de pronunciación: errores que impiden la comunicación, errores que dificultan la comunicación y errores que no dificultan la comunicación. Por otra parte, Collins y Mees (2008) establecen la siguiente clasificación de errores de pronunciación en inglés como L2 o LE según los problemas de inteligibilidad que producen en los oyentes:

- Categoría 1: Errores que producen problemas de inteligibilidad como la confusión de contrastes fonémicos cruciales entre vocales (/iː- ɪ/, /e-æ/, /ɜː-ɑː/ y /ɒ-ʌ/), problemas con la articulación de grupos consonánticos (/sp/, /sk/, /pt/), confusión de contrastes fonémicos cruciales entre consonantes (/b-v/, /ʃ-s/), elisión de /h/ o sustitución por /x/ y cambio de lugar de la sílaba tónica.
- Categoría 2: Errores que producen dificultad de comprensión o irritabilidad como la confusión en la de las fricativas dentales (/θ/-/t/), contrastes vocálicos menos significativos (/ uː- ʊ/; ɒ- ɔː/) y la pronunciación incorrecta de formas débiles.
- Categoría 3: Errores que producen pocas reacciones y pueden pasar desapercibidos como errores de entonación y ausencia de consonantes silábicas.

Los errores de pronunciación corregidos en este estudio son los correspondientes a las dos primeras categorías, aquellos que impiden o dificultan la comprensión del mensaje. Un total de 3.522 errores de pronunciación fueron identificados, clasificados y transcritos en el nivel fonológico. Para ello, comparamos cada uno de los errores de pronunciación con la pronunciación canónica establecida en el diccionario de pronunciación de Cambridge (Jones, 2011). Un anotador externo bilingüe verificó un subconjunto de los errores (10%) para evaluar la precisión de las transcripciones. Para ello se llevó a cabo una sesión de entrenamiento con el objetivo de llegar a un acuerdo satisfactorio entre anotadores. La tabla 3 muestra la distribución de errores transcritos por curso.

| Curso | |
|---|---|
| 5º E. Primaria | 1.058 |
| 6º E. Primaria | 1.201 |
| 1º E. Secundaria | 1.263 |
| Total | 3.522 |

Tabla 3: Distribución por curso de errores de pronunciación

La tipología de errores creada para este estudio consta de los siguientes elementos:

*2.2.3.1 Vocales*

a) Sustitución vocálica: Cambios de timbre o tensión en la pronunciación de las vocales, p. ej., la palabra *again* pronunciada como [**a**ˈgen], en vez de [**ə**ˈgen].
b) Inserción vocálica: p. ej., la inserción de una vocal epentética a principio de palabra ante grupo consonántico en la palabra *still;* [**e**stɪl] en lugar de [stɪl].
c) Elisión vocálica: p.ej.; *wanted* pronunciado como [wɑntd] en lugar de [ˈwɑntɪd].

*2.2.3.2 Consonantes*

a) Sustitución consonántica: Cambios en el punto y/o modo de articulación: p.ej., la palabra just pronunciada como [**x**ʌst] en lugar de [**dʒ**ʌst].
b) Inserción consonántica: la inserción del fonema /l/ en la palabra *could*; [kʊld] en lugar de [kʊd].
c) Elisión consonántica: la eliminación del fonema /t/ en posición final en la palabra *asked*; [ɑːsk] en lugar de [ɑːskt].
d) Metátesis: el cambio de posición de los fonemas /k/ y /tʃ/ en la palabra *kitchen*; [ˈ**tʃɪkɪ**n] en lugar de [ˈ**kɪtʃ**ɪn].

*2.2.3.3 Prosodia*

a) Errores de acento de intensidad: el cambio de acento de intensidad en la palabra *upset* de la segunda sílaba a la primera; [ˈ**ʌp**set] en lugar de [ʌpˈ**set**]

## *3 Resultados*

### 3.1 Fluidez en la lectura oral

En este apartado se presentan los resultados sobre la fluidez en la lectura oral expresado en palabras correctas por minuto (pcpm) de la muestra total de alumnos. La tabla 4 recoge la estadística descriptiva de los datos (media, desviación estándar, varianza).

| | N | Mínimo | Máximo | Media | D. estándar | Varianza |
|---|---|---|---|---|---|---|
| Fluidez | 124 | 61 | 191 | 114,06 | 24,208 | 586,032 |
| Curso | 124 | 1 | 3 | 2,07 | ,777 | ,604 |
| N válido (por curso) | 124 | | | | | |

Tabla 4: Estadística descriptiva de la fluidez en la lectura oral por curso

Los resultados del análisis de la varianza (ANOVA) muestran que la mejora de la fluidez en la lectura oral en el progreso de los alumnos es estadísticamente significativa (***, p<0,001). La tabla 5 muestra la media de palabras correctas leídas por minuto para cada curso.

| Curso | N | Media | Desviación estándar | Error estándar |
|---|---|---|---|---|
| 5º E. P. | 33 | 103,98 | 21,128 | 3,678 |
| 6º E. P. | 49 | 113,63 | 22,926 | 3,275 |
| 1º E. S. | 42 | 122,49 | 25,288 | 3,902 |
| Total | 124 | 114,06 | 24,208 | 2,174 |

Tabla 5: Media de palabras correctas leídas por minuto según el curso.

La fluidez lectora oral media para alumnos de 5º de E.P. es de 103,8 pcpm, 113,63 pcpm para alumnos de 6º, y 122,49 pcpm para alumnos de 1º E.S. Por otra parte, los resultados del análisis de la varianza (ANOVA) sobre el número de errores por curso muestran que el descenso en el número de errores cometidos no es estadísticamente significativo.

La media de errores cometidos en un minuto de lectura es 9,60 en 5º de E.P., 8,27 en alumnos de 6º y 7,49 en alumnos de 1º E.S.

Los datos parecen indicar que la precisión en la lectura (número de errores) es más lenta en desarrollarse que la automaticidad en la lectura (velocidad).

Esto podría deberse a que los errores de pronunciación en la lectura tienden a fosilizarse con el paso del tiempo en la adquisición del inglés como LE en un contexto formal.

## 3.2 Errores de pronunciación

En total se registraron 3.522 errores en el habla de los 117 participantes. Los dos tipos de error más frecuentes fueron los de sustitución e inserción vocálica. En la tabla 6 se muestran los porcentajes por tipo de error.

| Tipo de error | Nº | % |
|---|---|---|
| Sustitución vocálica | 1.593 | 45,23 |
| Inserción vocálica | 557 | 15,81 |
| Sustitución consonántica | 531 | 15,08 |
| Elisión consonántica | 297 | 8,43 |
| Inserción consonántica | 250 | 7,10 |
| Elisión vocálica | 159 | 4,51 |
| Acento de intensidad | 112 | 3,18 |
| Metátesis | 11 | 0,1 |

Tabla 6: Tipos de error

Los errores de sustitución vocálica fueron los más frecuentes en el habla de los participantes a los que se atribuye un total de un 45,23% de los errores en el corpus. Este resultado se puede explicar en parte por la diferencia entre los sistemas vocálicos del castellano y del inglés. En primer lugar, el castellano presenta cinco fonemas vocálicos (Quilis, 1993) (/a/, /e/, /i/, /o/, /u/), mientras que el inglés cuenta con doce vocales (/ɪ/, /e/, /æ/, /ʌ/, /ɒ/, /ʊ/, /ə/, /iː/, /ɜː/, /ɑː/, /ɔː/, /uː/) (Finch y Ortiz Lira, 1982).

Dada la diferencia entre ambos sistemas, los hispanohablantes se enfrentan a la tarea no sólo de aprender siete nuevas vocales que no existen en su inventario fonético, sino que tienen que ser capaces de distinguir las vocales existentes en ambos inventarios.

En segundo lugar, a diferencia del español, las vocales inglesas pueden clasificarse en vocales cortas y largas en función de su cantidad vocálica. Los hispanohablantes pueden tener dificultades en la percepción y producción de esta distinción debido a que la cantidad no es fonológicamente relevante en el sistema vocálico español.

Por último, la diferencia en la profundidad de los códigos alfabéticos entre ambas lenguas supone una dificultad añadida. En español existe una clara correspondencia entre los cinco fonemas vocálicos del castellano y las cinco grafías vocálicas que los representan. Sin embargo, el inglés cuenta con doce vocales simples que se que se corresponden con setenta representaciones ortográficas comunes y otras grafías de menor frecuencia (Finch y Ortiz Lira, 1982). Esta diferencia puede llevar al lector a aplicar de manera incorrecta las reglas de conversión grafema-fonema.

Los participantes de nuestro estudio tuvieron más problemas con la producción de vocales largas y centrales inexistentes en castellano /iː/, /ɜː/, /ʊ/, /ʌ/, y /ə/ y con los diptongos /ei/, /əʊ/ y /ai/. Al analizar los fonemas vocálicos elegidos por los estudiantes en estas sustituciones, se podría pensar que un gran número de errores de pronunciación parecen ser causados por interferencia del plano escrito o de la ortografía. Por ejemplo, la vocal /ə/, inexistente en el inventario fonético del castellano, es producida por los participantes con la vocal española correspondiente a la grafía en cada caso; como /o/ en ***p**olice*, como /a/ en ***a**bout*, /u/ en *moist**u**re*, /e/ en *ev**e**n* y en *aft**er.***

Además, el análisis de los datos muestra una frecuencia de error más elevada en aquellas palabras cuya ortografía es más opaca, o alejada de las reglas de conversión grafema fonema en español. Así pues, para la vocal corta /i/ las tasa de error es de 0,98 en palabras en las que el fonema está representado por la grafía *i* como en *sit,* pero mucho más elevado en los siguientes casos: 25,68 para la grafía *e* (***e**nough*), 31,25 en la grafía *ui* (*b**ui**ld*) y 33,33 para la grafía *ai* (*mount**ai**n*).

En segundo lugar, encontramos los errores de inserción vocálica (15,81%). Un ejemplo muy frecuente de este tipo de error es la inserción de una vocal en la pronunciación del morfema *-ed* de pasado y participio a pesar de que dicho morfema tiene tres realizaciones distintas: /Vd/, /d/, y /t/. Así pues, encontramos la palabra *called* en la que *-ed* corresponde al fonema /d/ (/kɔːld/) pronunciado como /kɔːled/. Otro tipo de inserción vocálica muy frecuente es la prótesis vocálica, en la que el hablante añade una vocal epentética ante un grupo consonántico a principio de palabra como en la palabra *stay* pronunciada como /esˈteɪ/ en lugar de /steɪ /.

La inserción de consonantes se encontró en un gran número de casos en la pronunciación de palabras que contienen letras mudas, aquellas presentes en la grafía, pero sin ninguna correspondencia fonética en el habla

como la grafía *l* en *walk* (/wɔ:k /) o la *l* en *could* (/kud/).

En cuanto a la elisión de consonantes (8,43%) la mayor parte de los participantes cometieron errores en la producción de grupos consonánticos a final de palabra como el pasado y participio de verbos regulares (*asked*; /ask/ en vez de /askt/), el sufijo *-s* marca de plural o de la tercera persona del presente simple (*twins*; /twin/ en vez de /twins/).

Por último, los errores menos frecuentes fueron los de acento de intensidad (3,18%) y los de metátesis (0,31%). En la tabla 7 presentamos una lista de las 20 palabras con mayor tasa de error en el estudio, donde las dificultades más repetidas se registraron en las sustituciones vocálicas, la pronunciación del morfema de pasado *-ed,* y la pronunciación de grupos consonánticos a principio de palabra.

| Palabra | Tasa de error % | Tipo de error |
|---|---|---|
| 1. nearby | 100 | Sustitución vocálica |
| 2. bushes | 91 | Sustitución vocálica |
| 3. chocolate | 86 | Sustitución vocálica |
| 4. moisture | 76 | Sustitución vocálica |
| 5. punished | 76 | Inserción vocálica/ elisión consonántica |
| 6. watched | 74 | Inserción vocálica/ elisión consonántica |
| 7. suit | 70 | Sustitución vocálica |
| 8. helped | 68 | Inserción vocálica/ elisión consonántica |
| 9. opened | 66 | Inserción vocálica/ elisión consonántica |
| 10. picked | 65 | Inserción vocálica/ elisión consonántica |
| 11. walked | 62 | Inserción vocálica/ elisión consonántica |
| 12. gathering | 61 | Sustitución consonántica |
| 13. stop | 61 | Inserción vocálica ante grupo consonántico |
| 14. stay | 58 | Inserción vocálica ante grupo consonántico |
| 15. upset | 58 | Acento de intensidad |
| 16. started | 55 | Inserción vocálica/ elisión consonántica |
| 17. still | 54 | Inserción vocálica ante grupo consonántico |
| 18. called | 52 | Inserción vocálica/ elisión consonántica |
| 19. decided | 51 | Inserción vocálica/ elisión consonántica |
| 20. aunt | 51 | Sustitución vocálica |

Tabla 7: Palabras con mayor tasa de error

## *4 Discusión y conclusiones*

El presente estudio ha analizado la adquisición de la fluidez en la lectura oral y las dificultades en la descodificación al leer en voz alta en alumnos de Educación Primaria y Secundaria de la Comunidad de Madrid.

La opacidad del sistema de escritura inglés, la exposición temprana de los alumnos al medio escrito en un contexto formal y la falta de instrucción explícita en el sistema de correspondencias grafo-fonológicas de la lengua meta pueden ocasionar errores en la producción oral.

En el contexto educativo español, en el que la lectura supone una gran parte del input lingüístico disponible para los estudiantes, un rendimiento poco fluido en la lectura puede resultar perjudicial para un adecuado rendimiento académico.

Nuestro estudio aporta datos sobre la fluidez en la lectura oral en inglés como lengua extranjera, utilizando medidas cuantitativas para medir la misma, con el fin de establecer unos parámetros que sirvan para evaluar la misma en distintos cursos de Educación Primaria y Secundaria en España. Dichos datos pueden ser utilizados para orientar una intervención educativa en las primeras etapas del proceso lector que incida en el desarrollo de las destrezas de descodificación.

Una mejora en la identificación automática de palabras podría tener como consecuencia un mejor desarrollo de la fluidez y de la comprensión lectoras en la lengua meta.

Asimismo, los resultados obtenidos con la creación y análisis del presente corpus de habla fueron empleados en un estudio previo para adaptar los modelos acústicos de FLORA (Fluent Oral Reading Assessment of Children's Speech), un reconocedor de habla infantil integrado en un sistema de evaluación automática de la fluidez en la lectura oral en hablantes de inglés como L1. Dicho estudio demostró que la clasificación automática de la fluidez en la lectura oral en términos de palabras correctas por minuto leídas por niños españoles en la lectura de textos en inglés puede ser estimada a un nivel de precisión similar al consenso alcanzado entre anotadores humanos (Bolaños et al., 2012).

Nuestros resultados podrían por lo tanto servir para desarrollar herramientas de enseñanza de la pronunciación asistida por ordenador (EPAO) y tutores de lectura que

incorporen sistemas de reconocimiento del habla para la mejora de la enseñanza del inglés en estudiantes españoles de inglés como lengua extranjera.

Por último, nuestros resultados pueden ser extendidos y mejorados de varias maneras; ampliando el corpus no sólo en cuanto al número de palabras, sino también incrementando el rango de edad de los participantes, mejorando el consenso entre anotadores y llevando a cabo un análisis más detallado de los errores de pronunciación estudiados.

## *Bibliografía*


Barras, C., E. Geoffrois, Z. Wu, y M. Liberman. 2001. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1-2), 5-22.

Bassetti, B. 2009. Orthographic Input and Second Language Phonology. En: Thorsten Piske & Martha Young-Scholten (eds.): *Input Matters in SLA*. (Multilingual Matters), 191-206

Bolaños, D., Cole, R.A., Walsh, P.E., y Ward, W.H. 2012. Automatic assessment of oral reading fluency for Spanish speaking ELs. WOCCI.

Breznitz, Z. 2006. *Fluency in reading: Synchronization of process*. Mahwah, NJ: Lawrence Erlbaum Associates.

Collins, B. y Mees M. 2008. *Practical Phonetics and Phonology: A Resource Book for Students. 2nd edn*. Abingdon: Routledge

Finch, D.F. y H. Ortiz Lira. 1982. *A Course in English Phonetics for Spanish Speakers*. London: Heinemann.

Fuchs, L., Fuchs, D., Hosp, M., y Jenkins, J. 2001. Oral reading fluency as an indicator or reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, 5, 239–256.

Good, R. H., Simmons, D. C., & Kame'enui, E. J. 2001. The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading*, 5(3), 257-288.

Good R., Kaminski R., Smith S., Laimon D, & Dill S. 2001. Dynamic indicators of basic early literacy skills. 5th ed. Eugene, OR: University of Oregon.

Götz, S. 2013. Fluency in Native and Nonnative English Speech. Amsterdam/Philadelphia: John Benjamins. *Studies in Corpus Linguistics*, 53.

Jenkins, J., Fuchs, L., van den Broek, P., Espin, C., y Deno, S. 2003. Accuracy and fluency in list and context reading of skilled and RD groups: Absolute and relative performance levels. *Learning Disabilities Research and Practice*, 18, 237-245.

Jones, D. 2011. *English Pronouncing Dictionary 18th Edition*. Ed. By P. Roach and J. Hartman. Cambridge: CUP.

LaBerge D, Samuels S. 1974. *Toward a theory of automatic information processing in reading*. Cognitive Psychology.

Lems, K. 2006. Reading fluency and comprehension in adult English language learners. En T. Rasinski, C. Blachowicz, & K. Lems, (Eds.)

Lems, K. 2012. The effect of L1 orthography on the oral reading of adult English language learners. *Writing Systems Research*, iFirst, 1–11.

Lennon, P. 1990. Investigating Fluency in EFL: A Quantitative Approach. *Language Learning* 40(3), 387–417.

Llisterri, J. 2003. La enseñanza de la pronunciación, Cervantes. *Revista del Instituto Cervantes en Italia,* 4, 1, páginas 91-114.

National Reading Panel. 2000. Report of the National Reading Panel: Teaching children to read. Report of the subgroups. Washington, DC: US Department of Health and Human Services, National Institutes of Health.

Perfetti C. 1985. *Reading ability*. New York, NY: Oxford University Press.

Quilis, A. 1993. *Tratado de fonología y fonética españolas*. Madrid: Gredos (Biblioteca Románica Hispánica, Manuales, 74).

Rasinski, T.V. 2006. A brief history of reading fluency. En S. Samuels & A. Farstrup (Eds.) What research has to say about fluency

instruction (p. 70-93). Newark, DE: International Reading Association.

# *Aprendizaje automático en PLN*

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 101-108 recibido 01-03-2018 revisado 30-04-2018 aceptado 03-05-2018

# Legibilidad del texto, métricas de complejidad y la importancia de las palabras

## *Text readability, complexity metrics and the importance of words*

**Rocío López-Anguita, Arturo Montejo-Ráez**
**Fernando J. Martínez-Santiago**, **Manuel Carlos Díaz-Galiano**
Centro de Estudios Avanzados en TIC
Universidad de Jaén
{rlanguit, amontejo, dofer, mcdiaz}@ujaen.es

**Resumen:** El presente trabajo expone un estudio sobre la determinación de la edad recomendada de lectura sobre un conjunto de textos infantiles. Se ha evaluado el mismo con 12 medidas de complejidad propuestas por distintos autores. Usando estas medidas como características, hemos modelado los textos y aplicado una validación cruzada con varios clasificadores automáticos. Los resultados se han comparado con otras formas de representación de los textos, como vectores de palabras y vectores TF.IDF. Nuestras conclusiones indican que el rasgo más determinante para la determinación de la edad de lectura recomendada no radica tanto en factores como la complejidad sintáctica o léxica, sino en el uso de determinado vocabulario.
**Palabras clave:** Legibilidad, complejidad textual, modelado del lenguaje

**Abstract:** This article describes our study on the identification of the recommended age for readers in texts written for children. They have been evaluated over 12 complexity metrics proposed by different authors. By using these metrics as features, we have trained several automatic classifiers and cross-validated their performances to detect recommended reader level. The results have been compared with the classification performance obtained from other document models, like word embeddings and TF.IDF vectors. Our conclusiones are that the most relevant facet to identify the recommended reader age is not on lexical or syntactical complexities, but strongly related with the vocabulary involved.
**Keywords:** Readability, text complexity, language modelling

### *1 Introducción*

Conocer cómo de adecuado es un texto para un persona es algo que no está resuelto. Mucho se ha investigado en la legibilidad de textos según el lector. En educación primaria podemos encontrar desde la primera fórmula de Spache (1953) hasta la visión más holística de Larson y Marsh (2014). En el caso de trastornos del lenguaje y la adecuación de los textos a personas con dificultades cognitivas también hay un extenso trabajo realizado, como veremos.

Determinar la legibilidad de un texto no es una tarea sencilla, pues cada lector presenta destrezas diferentes (Cain, Oakhill, y Bryant, 2004) o, incluso, limitaciones como la dislexia (Rello et al., 2013) o el autismo. Entendemos por legibilidad la facilidad de lectura de comprensión de un texto. Si no tenemos en consideración aspectos de forma, color, maquetado y tipografía (Ripoll, 2015), la legibilidad se determina por los rasgos lingüísticos, que suelen agruparse en aquellos relacionadas con la gramática (o lo que es lo mismo, la sintaxis) y aquellos relacionados con el léxico (es decir, el vocabulario) (Alliende González, 1994). Son múltiples los autores que han establecido métricas para la legibilidad en ambos ámbitos. En concreto, este trabajo visita 12 de las métricas más utilizadas para la legibilidad léxica y sintáctica y estudia su idoneidad como características para la determinación de la edad recomendada de lectura en textos de educación primaria.

El uso de algoritmos de aprendizaje automático para determinar la edad recomendada para un texto a partir de estas medidas de complejidad se ha comparado median-

ISSN 1135-5948. DOI 10.26342/2018-61-11

te validación cruzada con otras dos formas de modelar el lenguaje: mediante vectores de palabras (Mikolov et al., 2013) y mediante el modelo de espacio vectorial clásico (Salton, Wong, y Yang, 1975). Los resultados muestran que el último modelo, con una selección de características adecuada y determinados algoritmos de aprendizaje, resulta muy apropiado para esta tarea, destacando de esta forma la relevancia del vocabulario en la comprensión lectora. Los datos utilizados para la experimentación han sido extraídos de la web y consisten en diversas lecturas recomendadas para distintos niveles en Educación Primaria.

El artículo está organizado como sigue: en la Sección 2 se introducen las medidas de complejidad utilizadas para el modelado de los textos; la Sección 3 describe el conjunto de textos que hemos utilizado como base para los experimentos, que se detallan en la Sección 4. En la Sección 5 analizamos los resultados obtenidos y en la Sección 6 se cierra el trabajo con las conclusiones del mismo.

## *2 Medidas de Complejidad*

En este apartado, vamos hacer un recorrido por las diferentes métricas de complejidad que se han propuesto por diversos autores. Si bien algunas de estas medidas proporcionan directamente la edad recomendada, como la medida de García López (2001), basada, a su vez, en la de Flesch (1948), otras ofrecen índices de más difícil interpretación, como la complejidad léxica de Anula (2008), el índice de complejidad de oraciones o la profundidad del árbol de dependencia de (Saggion et al., 2015), entre otras.

**Complejidad del léxico.** Esta medida de complejidad fue propuesta por Anula (2008) para medir la complejidad léxica de un texto, determinada por la frecuencia de uso y la densidad léxica. Se considera que a mayor densidad léxica (mayor número de palabras diferentes por textos) mayor dificultad para la comprensión.

**Legibilidad de Spaulding.** La legibilidad de Spaulding, comúnmente conocido como el índice SSR, fue propuesta por Spaulding (1956). Se centra en medir el vocabulario y la estructura de oraciones para predecir la dificultad relativa de legibilidad de un texto.

**Complejidad de oraciones.** El índice de complejidad de oraciones fue propuesto por Anula (2008). Mide el número de palabras por oración, obteniendo así el índice de longitud oracional, y el número de frases complejas que hay por oración, a partir de un índice de frases complejas.

**Índice de legibilidad automatizado.** Senter y Smith (1967) nos proponen uno de los índice más utilizados debido a su facilidad de cálculo. Mide la dificultad de un texto a partir del número medio de caracteres (letras y números) por palabra y del número medio de palabras por oración.

**Altura del árbol de dependencia.** Esta medida fue propuesta por Saggion et al. (2015). Es una métrica muy útil para capturar la complejidad sintáctica: las oraciones largas pueden ser sintácticamente complejas o contener una gran cantidad de modificadores (adjetivos, adverbios o frases adverbiales). Estos últimos no aumentan la complejidad sintáctica y no dan lugar a árboles muy profundos, mientras que los primeros tienen una fuerte tendencia a producir árboles profundos.

**Marcas de Puntuación.** Esta medida fue también propuesta por Saggion et al. (2015). En la complejidad de un texto, el número promedio de signos de puntuación se utiliza como uno de los indicadores de complejidad del mismo.

**Lecturabilidad de Fernández-Huerta.** Blanco Pérez y Gutiérrez Couto (2002) y Ramírez-Puerta et al. (2013) nos proponen esta medida de complejidad como una adaptación al español de la prueba de legibilidad de Flesch (Flesch (1948)). Parte de que en español las palabras en promedio tienen más sílabas y las oraciones también son más largas. Mide el promedio de sílabas por palabra y el promedio de palabras por oración que hay en el texto.

**Legibilidad de Flesch-Szigrist (IFSZ).** Los trabajos De Granada Barrio-Cantalejo et al. (2008) y Ramírez-Puerta et al. (2013) nos proponen el índice de legibilidad de Flesch-Szigristzt como una modificación de la fórmula de Flesch (Flesch, 1948) adaptada al

castellano. El índice de legibilidad IFSZ es considerado de referencia para la lengua española. Mide el número de sílabas por palabra y el número de palabras por oración que hay en el texto.

**Comprensibilidad de Gutiérrez.** Fue creada para el castellano (Rodríguez, 1980) y consiste en una fórmula matemática, generada por métodos de regresión múltiple, que incluye ciertas características lingüísticas del material cuya dificultad se pretende evaluar. Se centra en medir el promedio de letras por palabra y el promedio de palabras por oración.

**Legibilidad $\mu$.** La Legibilidad $\mu$ propone una fórmula para calcular la facilidad lectora de un texto. Proporciona un índice comprendido entre 0 y 100 y fue desarrollada por Muñoz (2006). Esta medida se centra en medir el número de palabras, la media del número de letras por palabra y su varianza.

**Edad mínima de comprensibilidad.** En el trabajo de García López (2001) podemos encontrar otra fórmula para medir la edad necesaria para entender un texto. Es, de nuevo, una adaptación al castellano de la fórmula original de Flesch (Flesch (1948)) para el inglés. Mide el promedio de sílabas por palabra y el promedio de palabras por oración para obtener la edad mínima necesaria para entender un texto.

**SOL.** Contreras et al. (1999) nos propone la métrica SOL como una adaptación al español de la fórmula SMOG propuesta por Mc Laughlin (1969). Mide la legibilidad de un texto mediante el nivel de grado, que viene a ser el número de años de escolaridad necesarios para entender el texto.

## *3 Descripción del corpus*

Nuestro objetivo es evaluar la idoneidad de un sistemas de clasificación automático que utilice estas métricas para determinar la edad recomendada de lectura para un texto.

El corpus que utilizamos en este trabajo está compuesto por 300 textos de lecturas en español, orientados a alumnos de Educación Primaria. Dichos textos están clasificados por el curso a los que van dirigidos. En consecuencia, tenemos 6 grupos (1º, 2º , 3º , 4º , 5º y 6º de primaria) y cada grupo consta de 50 textos.

Los textos han sido obtenidos, principalmente, de un trabajo realizado por un grupo de profesores de distintos centros de Sevilla, coordinados por María José Moya Bellido y Antonio Ruiz y Martín (Inspectores de Educación del Servicio de Sevilla), en el curso escolar 2011/2012[1].

Además hemos completado nuestro corpus con otras lecturas y con los primeros capítulos de algunos cuentos, obtenidos de distintos sitios web, donde encontramos recursos educativos accesibles y gratuitos, como por ejemplo "Orientación Andújar"[2] o "Rincón de lecturas"[3].

En la Tabla 1, podemos observar las características más relevantes del corpus compilado.

## *4 Experimentos*

Como se ha indicado anteriormente, hemos evaluado la aportación de estas métricas de complejidad en una tarea de determinación de la edad de lectura recomendada mediante aprendizaje automático. Además, hemos estudiado la dependencia de las distintas métricas con el nivel de Educación Primaria mediante un análisis de los valores Chi-cuadrado que obtenemos al estudiar los pares de distribuciones <*valores de la característica, nivel*>

Para evaluar los sistemas hemos configurado distintos conjuntos de datos con diferente granularidad en los niveles de Educación Primaria considerados, tal y como pasamos a detallar a continuación.

### 4.1 Conjuntos de datos

Para llevar a cabo los experimentos hemos definido los dos siguientes conjuntos de datos:

- `123456`: A cada curso de primaria le asignamos un nivel (de 1º a 6º de educación primaria).
- `110022`: Tomamos dos grupos, donde los cursos 1º y 2º pertenecen al nivel 1, 5º y 6º pertenecen al nivel 2 y los cursos 3º y 4º no son considerados.

La razón de esta configuración de las clases (es decir, los niveles), es para propiciar

[1] `http://sosprofes.es`

[2] `https://www.orientacionandujar.es/`

[3] `http://rincondelecturas.com/`

| Curso | 1º | 2º | 3º | 4º | 5º | 6º | Total |
|---|---|---|---|---|---|---|---|
| Nº Textos | 50 | 50 | 50 | 50 | 50 | 50 | 300 |
| Tamaño Vocabulario: palabras distintas en el texto | 1966 | 2648 | 2799 | 3759 | 4979 | 5333 | 21484 |
| Longitud Media en palabras | 150.54 | 204.6 | 222.22 | 370.26 | 489.16 | 514.26 | 1978.04 |
| Longitud Mín en palabras | 36 | 41 | 31 | 61 | 38 | 73 | 280 |
| Longitud Máx en palabras | 454 | 692 | 1508 | 1640 | 1797 | 1572 | 7663 |
| Media de palabras raras | 49.58 | 64.88 | 69.82 | 114.98 | 147.3 | 161.04 | 607.6 |
| Nº palabras por oración | 12.68 | 16.43 | 16.14 | 22.52 | 20.27 | 22.05 | 110.09 |
| Nº sílabas por palabra | 1.88 | 1.98 | 1.95 | 1.96 | 1.98 | 1.995 | 11.75 |
| Media de oraciones complejas | 9.20 | 12.34 | 12.65 | 19.11 | 17.14 | 17.52 | 87.96 |
| Nº total palabras | 7533 | 10202 | 11081 | 18478 | 24429 | 27038 | 98761 |

Tabla 1: Caracterización del Corpus

mayor separabilidad y dilucidar de forma clara qué características son las más significativas cuando de determinar el nivel de lectura se trata. De esta forma, la segunda configuración debería ser más separable a nivel de clases.

### 4.2 Evaluación mediante aprendizaje supervisado

La hipótesis es que, dado que las métricas enumeradas anteriormente capturan distintos aspectos de la complejidad del texto, deberían ser válidas como características en un modelo que represente cada documento en este proceso de clasificación del nivel. Para ello vamos a aplicar validación cruzada sobre los distintos conjuntos (`123456` y `110022`) y algoritmos (LinearSVC, Multilayer Perceptron, Random Forest y Naive Bayes). Adicionalmente, vamos a comparar estos resultados con los obtenidos con otros modelos del texto pero no asociados a la complejidad, como los vectores de palabras (*word embeddings* mediante Word2Vec) o el modelo de espacio vectorial clásico con TF.IDF.

Analizaremos, de esta forma, la pertinencia o no de estas medidas de complejidad de manera indirecta a partir de la evaluación de los clasificadores entrenados con ellas. Obtenemos, por validación cruzada y macropromediado, las medidas de exactitud (*accuracy*), F1, cobertura (*recall*) y precisión (*precision*) de cada algoritmo para cada combinación de clases.

Una vez obtenidos dichos valores para cada clasificador, realizamos una prueba de Chi-cuadrado a las muestras y volvemos aplicar la validación cruzada para ver cuáles son las medidas de complejidad más relevantes.

Por último, también probamos con las representaciones de Word2Vec y TF.IDF.

Para todo esto hemos usado la herramienta *Freeling* (Padró y Stanilovsky, 2012) para procesar los textos y las bibliotecas de Python *SciKit-Learn* (Pedregosa et al., 2011) para aprendizaje automático y *Gensim* (Rehurek y Sojka, 2011) para trabajar con vectores de palabras.

### 4.3 Experimentos

Hemos representado los textos según cuatro modelos diferentes:

1. Los textos son representados por las medidas de complejidad indicadas anteriormente, sobre las que hemos aplicado una normalización L2 de los valores.
2. Word2Vec siguiendo el método de aglomeración propuesto por Montejo-Ráez y Díaz-Galiano (2016).
3. TF.IDF.
4. Medidas de Complejidad + Word2Vec + TF.IDF.

## *5 Análisis de resultados*

Como era de esperar, independientemente de los algoritmos y las representaciones utilizadas para los textos, obtenemos el peor resultado cuando consideramos el conjunto de datos `123456`. Esto se debe a que no hay tanta diferencia entre ellos, puesto que son cursos consecutivos y van aumentando su dificultad gradualmente. El mejor resultado lo obtenemos con el conjunto 110022, puesto que la diferencia entre los cursos considerados de primer ciclo de primaria y tercer ciclo de primaria es más significativa. En cualquier caso,

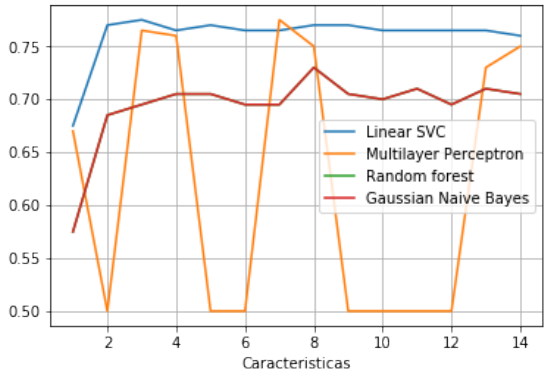


Figura 1: Resultados filtrando por Chi-cuadrado las medidas de complejidad

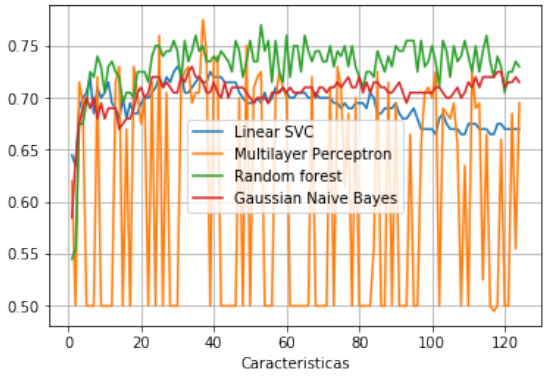


Figura 2: Resultados filtrando por Chi-cuadrado en word2Vec

es importante remarcar que para el objeto de nuestra investigación nos interesa sobre todo analizar las diferencias entre modelos de representación de los textos bajo las mismas condiciones de dificultad de clasificación (clases) y algoritmos usados.

## 5.1 Resultados con medidas de complejidad

En la Tabla 2 presentamos la estadística del aprendizaje automático supervisado con las medidas de complejidad para las distintas configuraciones de clases.

En la Tabla 2, podemos observar que el máximo se alcanza con un *accuracy* de 0.76 con Multilayer Perceptron. En cualquier caso, los resultados arrojados por los distintos clasificadores son comparables. Las diferencias entre ellos pueden responder más a cuestiones de ajuste de los parámetros del clasificador que a su conveniencia o no en esta tarea. El uso de distintos algoritmos no es sino para comprobar si nuestra hipótesis se mantiene independientemente del tipo de algoritmos de clasificación utilizado.

Cuando aplicamos la Chi-cuadrado y después la validación cruzada obtenemos la Figura 1. Podemos observar que el máximo se alcanza con una *accuracy* de 0.775 con tres características con el clasificador Linear SVC.

## 5.2 Resultados con Word2Vec

Los resultados los podemos observar en la Tabla 3 para las dos tareas de clasificación y los distintos algoritmos utilizados.

En la Tabla 3, podemos observar que el máximo se alcanza con una *accuracy* de 0.73 con el clasificador Random Forest. Cuando le aplicamos la Chi-cuadrado antes de la validación cruzada para seleccionar características obtenemos la Figura 2. Siendo el máximo 0.775 y se alcanza con 37 características en el clasificador Multilayer Perceptron.

Como es de esperar en el uso de una red neuronal como es el perceptrón multicapa, cuando el número de características es grande el clasificador puede requerir de una mayor cantidad de ejemplos para llegar a estimar un buen modelo. Esta sería la razón para su buen comportamiento al filtrar características por Chi-cuadrado (a 37) y al usar medidas de complejidad (12) frente a vectores de palabras o vectores TF.IDF que cuentan con centenares de características.

## 5.3 Resultados con TF.IDF

Podemos observar los resultados obtenidos en la Tabla 4, de nuevo, para las dos tareas de clasificación y los distintos algoritmos utilizados.

En la tabla 4, podemos observar que el máximo se alcanza con una *accuracy* de 0.76 con el clasificador SVC Linear. Cuando le aplicamos la Chi-cuadrado al conjunto 110022 con los mismos clasificadores sobre la representación de TF.IDF, obtenemos la Figura 3. Siendo el máximo 0.92 y se alcanza con 199 características en el clasificador Gaussian Naive Bayes.

## 5.4 Resultados con medidas de complejidad + Word2Vec + TF.IDF

Los resultados basados en las medidas de complejidad junto a Word2Vec y TF.IDF se presentan en la Tabla 5.

| clases | algoritmo | accuracy | F1 | precision | recall |
|---|---|---|---|---|---|
| **123456** | Linear SVC | 0.26 | 0.19 | 0.07 | 0.18 |
| **123456** | Multilayer Perceptron | 0.27 | 0.23 | 0.07 | 0.20 |
| **123456** | Random Forest | 0.29 | 0.29 | 0.34 | 0.33 |
| **123456** | Gaussian Naive Bayes | 0.26 | 0.22 | 0.22 | 0.25 |
| **110022** | Linear SVC | **0.76** | 0.74 | 0.59 | 0.64 |
| **110022** | Multilayer Perceptron | **0.76** | **0.75** | 0.25 | 0.5 |
| **110022** | Random Forest | 0.72 | 0.69 | 0.79 | 0.74 |
| **110022** | Gaussian Naive Bayes | 0.7 | 0.68 | 0.76 | 0.72 |

Tabla 2: Resultados con Medidas de Complejidad

| clases | algoritmo | accuracy | F1 | precision | recall |
|---|---|---|---|---|---|
| **123456** | Linear SVC | 0.30 | 0.28 | 0.32 | 0.30 |
| **123456** | Multilayer Perceptron | 0.17 | 0.05 | 0.03 | 0.17 |
| **123456** | Random Forest | 0.26 | 0.25 | 0.32 | 0.26 |
| **123456** | Gaussian Naive Bayes | 0.30 | 0.27 | 0.30 | 0.30 |
| **110022** | Linear SVC | 0.72 | **0.70** | 0.74 | 0.72 |
| **110022** | Multilayer Perceptron | 0.50 | 0.33 | 0.25 | 0.50 |
| **110022** | Random Forest | **0.73** | 0.69 | 0.76 | 0.77 |
| **110022** | Gaussian Naive Bayes | 0.71 | 0.69 | 0.74 | 0.71 |

Tabla 3: Resultados con Word2Vec

| clases | algoritmo | accuracy | F1 | precision | recall |
|---|---|---|---|---|---|
| **123456** | SVC Linear | 0.27 | 0.25 | 0.27 | 0.27 |
| **123456** | Multilayer Perceptron | 0.18 | 0.08 | 0.08 | 0.18 |
| **123456** | Random Forest | 0.27 | 0.31 | 0.34 | 0.31 |
| **123456** | Gaussian Naive Bayes | 0.19 | 0.17 | 0.21 | 0.19 |
| **110022** | SVC Linear | **0.76** | **0.74** | 0.79 | 0.76 |
| **110022** | Multilayer Perceptron | 0.50 | 0.33 | 0.25 | 0.50 |
| **110022** | Random Forest | 0.71 | **0.74** | 0.81 | 0.74 |
| **110022** | Gaussian Naive Bayes | 0.61 | 0.57 | 0.64 | 0.61 |

Tabla 4: Resultados con TF.IDF

| clases | algoritmo | accuracy | F1 | precision | recall |
|---|---|---|---|---|---|
| **123456** | Linear SVC | 0.31 | 0.27 | 0.29 | 0.31 |
| **123456** | Multilayer Perceptron | 0.21 | 0.14 | 0.12 | 0.21 |
| **123456** | Random Forest | 0.28 | 0.28 | 0.32 | 0.28 |
| **123456** | Gaussian Naive Bayes | 0.16 | 0.12 | 0.15 | 0.16 |
| **110022** | Linear SVC | **0.77** | **0.74** | 0.78 | 0.77 |
| **110022** | Multilayer Perceptron | 0.68 | 0.64 | 0.76 | 0.68 |
| **110022** | Random Forest | 0.75 | 0.73 | 0.80 | 0.75 |
| **110022** | Gaussian Naive Bayes | 0.55 | 0.44 | 0.56 | 0.55 |

Tabla 5: Resultados con Medidas de Complejidad + Word2Vec + TF.IDF

En la Tabla 5, podemos observar que el máximo se alcanza con una *accuracy* de 0.77 con el clasificador SVC Linear. Cuando le aplicamos la Chi-cuadrado al conjunto 110022 con los mismos clasificadores sobre las medidas de complejidad y las representaciones de Word2Vec y de TF.IDF, obtenemos la siguiente gráfica que se muestra en la Figura 4. Siendo el máximo 0.90 y se alcanza con 209 características en el clasificador Multilayer Perceptron.

En resumen, obtenemos resultados muy

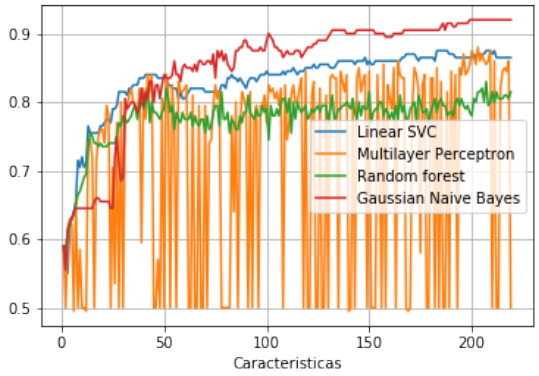


Figura 3: Resultados filtrando por Chi-cuadrado en TF.IDF

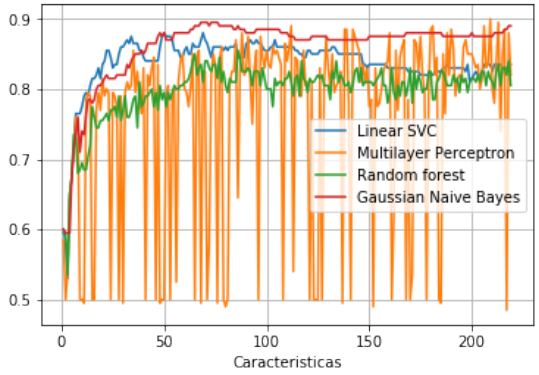


Figura 4: Resultados filtrando por Chi-cuadrado las medidas de complejidad + Word2Vec + TF.IDF

parecidos para las medidas de complejidad y Word2vec. Sin embargo, se mejoran los resultados si consideramos la unión de las medidas de complejidad con Word2Vec y TF.IDF. En cualquier caso, el mejor resultado es el obtenido TF.IDF con filtrado de términos mediante Chi-cuadrado. Los clasificadores Gaussian Naive Bayes y Multilayer Perceptron son sensibles al número de características, y una reducción de las mismas llegar a arrojar resultados destacados.

## 6 Conclusiones

Hemos visto cómo los indicadores clásicos de complejidad textual sobre el español pueden ayudarnos a resolver una tarea de clasificación de textos en distintos niveles o cursos del ciclo formativo de primaria. En cualquier caso, los modelos de representación basados en contenido resultan mejores para esta tarea, si bien las medidas de complejidad pueden complementarles. Destacan los resultados utilizando un modelo clásico de representación de los textos como TF.IDF, frente a representaciones muy utilizadas en la actualidad como los vectores de palabras. Parece no haber duda de la relevancia del vocabulario a la hora de determinar qué es más recomendable que lean.

Nuestro estudio, de una forma empírica, corrobora el de Stahl (2003), al reflejar la fuerte influencia de la riqueza de vocabulario del lector en la comprensión lectora, más allá de los símbolos y la gramática. No obstante, consideramos que las medidas de complejidad pueden ser una forma conveniente para modelar el lenguaje natural en determinadas aplicaciones, como la detección de autoría, la selección de textos para personas con dificultades asociadas a trastornos del lenguaje (autismo, parálisis cerebral...), o la detección temprana de deterioros cognitivos, como el Alzheimer. Otra línea futura de trabajo es el analizar no sólo el lenguaje formal, sino también el informal en entornos como los medios sociales en Internet.

## Agradecimientos

Este trabajo ha sido parcialmente financiado por el Gobierno de España a través del proyecto REDES (TIN2015-65136-C2-1-R).

## Bibliografía

Alliende González, F. 1994. La legibilidad de los textos. *Santiago de Chile: Andrés Bello*, 24.

Anula, A. 2008. Lecturas adaptadas a la enseñanza del español como l2: variables lingüísticas para la determinación del nivel de legibilidad. *La evaluación en el aprendizaje y la enseñanza del español como LE L*, 2:162–170.

Blanco Pérez, A. y U. Gutiérrez Couto. 2002. Legibilidad de las páginas web sobre salud dirigidas a pacientes y lectores de la población general. *Revista española de salud pública*, 76(4):321–331.

Cain, K., J. Oakhill, y P. Bryant. 2004. Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of educational psychology*, 96(1):31.

Contreras, A., R. Garcia-Alonso, M. Echenique, y F. Daye-Contreras. 1999. The sol

formulas for converting smog readability scores between health education materials written in spanish, english, and french. *Journal of health communication*, 4(1):21–29.

De Granada Barrio-Cantalejo, D. S., P. Simón-Lorda, M. Melguizo, I. Escalona, M. Marijuán, P. Hernándo, y others. 2008. Validación de la escala inflesz para evaluar la legibilidad de los textos dirigidos a pacientes.

Flesch, R. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

García López, J. 2001. Legibilidad de los folletos informativos. *Pharmaceutical Care España*, 3(1):49–56.

Larson, J. y J. Marsh. 2014. *Making literacy real: Theories and practices for learning and teaching*. Sage.

Mc Laughlin, G. H. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, y J. Dean. 2013. Distributed representations of words and phrases and their compositionality. En *Advances in neural information processing systems*, páginas 3111–3119.

Montejo-Ráez, A. y M. C. Díaz-Galiano. 2016. Participación de sinai en tass 2016. En *TASS@ SEPLN*, páginas 41–45.

Muñoz, M. 2006. Legibilidad y variabilidad de los textos. *Boletín de Investigación Educacional, Pontificia Universidad Católica de Chile, 21*, 2:13–26.

Padró, L. y E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. En *LREC2012*.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, y others. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Ramírez-Puerta, M., R. Fernández-Fernández, J. Frías-Pareja, M. Yuste-Ossorio, S. Narbona-Galdó, y L. Peñas-Maldonado. 2013. Análisis de legibilidad de consentimientos informados en cuidados intensivos. *Medicina Intensiva*, 37(8):503–509.

Rehurek, R. y P. Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Rello, L., R. Baeza-Yates, S. Bott, y H. Saggion. 2013. Simplify or help?: text simplification strategies for people with dyslexia. En *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, página 15. ACM.

Ripoll, J. C. 2015. Font legibility in first year primary students/legibilidad de distintos tipos de letra en alumnos de primero de primaria. *Infancia y Aprendizaje*, 38(3):600–616.

Rodríguez, T. 1980. Determinación de la comprensibilidad de materiales de lectura por medio de variables lingüísticas. *Lectura y vida*, 1(1):29–32.

Saggion, H., S. Štajner, S. Bott, S. Mille, L. Rello, y B. Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):14.

Salton, G., A. Wong, y C.-S. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Senter, R. y E. A. Smith. 1967. Automated readability index. Informe técnico, CINCINNATI UNIV OH.

Spache, G. 1953. A new readability formula for primary-grade reading materials. *The Elementary School Journal*, 53(7):410–413.

Spaulding, S. 1956. A spanish readability formula. *The Modern Language Journal*, 40(8):433–441.

Stahl, S. A. 2003. Vocabulary and readability: How knowing word meanings affects comprehension. *Topics in Language Disorders*, 23(3):241–247.

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 109-116 recibido 30-03-2018 revisado 30-04-2018 aceptado 04-05-2018

# Clasificación automatizada de marcadores discursivos

## *Automatic categorization of discourse markers*

**Hernán Robledo, Rogelio Nazar**
Instituto de Literatura y Ciencias del Lenguaje
Pontificia Universidad Católica de Valparaíso
hernan.robledo.n@mail.pucv.cl, rogelio.nazar@pucv.cl

**Resumen:** Presentamos un método de clasificación de marcadores del discurso. A partir de una taxonomía generada inductivamente en un trabajo anterior, desde un corpus paralelo de gran tamaño y utilizando una técnica de *clustering*, proponemos ahora un sistema que permite clasificar un marcador discursivo no incluido en esa taxonomía en alguna de las categorías emergentes. Está basado en el cálculo de la similitud estadística entre el nuevo marcador y las categorías. Destacamos la naturaleza cuantitativa del enfoque, que permite la reproducción del experimento en otras lenguas. Además, el sistema propuesto es un clasificador multicategoría, y esto es importante ya que representa un primer acercamiento al estudio de la polifuncionalidad de los marcadores del discurso desde un enfoque empírico e inductivo.
**Palabras clave:** marcadores del discurso, métodos cuantitativos, métodos inductivos, clasificación multicategorial

**Abstract:** We present a method for the categorization of discourse markers. Starting from the result of a previous research, in which we generated a taxonomy of discourse markers by inductive methods from parallel corpus, we propose now a method to classify new discourse markers in one or more of the categories discovered in our previous research. The method is based on the statistical similarity between a new marker and the emerging categories. We highlight the quantitative nature of the approach, because it will allow to replicate experiments in other languages. Furthermore, ours is a multi-label classification method, which is important because it represents a first approach to the study of the polyfunctionality of discourse markers from an empirical and inductive point of view.
**Keywords:** discourse markers, inductive methods, quantitative methods, multilabel categorization

### *1 Introducción*

Los llamados marcadores del discurso (MDs) corresponden a un amplio y heterogéneo conjunto de unidades lingüísticas (por ejemplo, *sin embargo*, *no obstante*, *es decir*, *por lo tanto*, *en consecuencia*, *a todo esto*, *por una parte*, *en primer lugar*, *claramente*, entre muchas otras) de uso muy habitual en las lenguas naturales, tanto en la escritura como en la oralidad.

Los MDs han sido estudiados en una gran variedad de lenguas, como español, inglés, francés, alemán, chino, italiano, japonés, portugués, ruso y muchas otras –incluso, en lenguas de señas– y se han explorado en una diversidad de géneros discursivos y contextos de interacción, como narraciones, discursos políticos, discursos periodísticos, salas de clases, programas radiales, etc. (Maschler y Schiffrin, 2015). Así, si por un lado, los MDs muestran este uso tan extendido entre las lenguas naturales, en la teoría lingüística, en tanto, su delimitación como objeto de estudio, su denominación categorial, sus propiedades funcionales (y/o formales) y, en consecuencia, su clasificación han sido campo de mucha controversia (Fischer, 2006; Loureda y Acín, 2010; Maschler y Schiffrin, 2015).

La categoría de unidades que funcionan como MDs está constituida por elementos que provienen de distintas categorías gramaticales, como conjunciones y locuciones

ISSN 1135-5948. DOI 10.26342/2018-61-12

conjuntivas, adverbios y locuciones adverbiales, interjecciones, preposiciones, expresiones performativas, sintagmas preposicionales, entre otras. Además de esta heterogeneidad de orígenes, los MDs operan en distintos niveles: conectan oraciones, cumplen funciones en el texto y operan a nivel interpersonal (Brinton, 1996; Jucker y Ziv, 1998; Aijmer, 2002). Muchos de ellos, además, son polifuncionales, fenómeno que ha sido descrito desde distintas posturas teóricas (Schiffrin, 1987; Wierzbicka, 2003; Aijmer, Foolen, y Vandenbergen, 2006; Fischer, 2006; Fraser, 2006, por ejemplo) y que refiere al hecho de que un MD puede cumplir distintas funciones pragmáticas en el discurso (ver sección 2.3).

El análisis de estos distintos usos ha sido la motivación de la presente investigación. En una investigación previa (Robledo, Nazar, y Renau, 2017) desarrollamos un método para la inducción automática de taxonomías de MDs a partir de corpus paralelos, utilizando una técnica de *clustering* (conglomerados). Por ejemplo, uno de esos *clusters* nos presenta elementos que denominamos conectores contraargumentativos, tales como *sin embargo*, *por el contrario* o *en vez de ello* (ver sección 2.2).

Lo que presentamos en este artículo es cómo, a partir de esa taxonomía ya generada, desarrollar un sistema de clasificación de cualquier MD, naturalmente sin necesidad de que esté ya incluido en esa taxonomía. Así, un elemento no incluido en esa taxonomía, como *empero*, es detectado como similar –desde el punto de vista distribucional– al *cluster* de los conectores contraargumentativos. En otras palabras, aprovechamos el resultado de un estudio exploratorio, que nos proporcionó *clusters* de MDs que cumplen la misma función y que nosotros etiquetamos con nombres de categorías, para utilizarlo como material de entrenamiento para un sistema de clasificación automática de nuevos marcadores.

Destacamos cuatro aspectos como los más relevantes de esta investigación: 1) que las categorías de MDs no provienen de la literatura, sino que se llega a ellas como categorías emergentes a partir del *clustering* inicial; 2) que ahora podemos clasificar un MD en más de una categoría, ya que algunos de ellos son polifuncionales; 3) que más allá de las consecuencias teóricas de la investigación, encontramos diversas aplicaciones prácticas en lexicografía, *parsing* discursivo y redacción asistida por computador, entre otras; y 4) que este estudio está basado íntegramente en análisis cuantitativo, lo que facilita la reproducción de los resultados del experimento en otras lenguas.

El artículo se organiza de la siguiente forma: en la Sección 2 presentamos un breve estado de la cuestión en el estudio de los MDs. Allí (Subsección 2.2) explicamos también los resultados de nuestra propia investigación anterior y describimos la taxonomía inductiva de MDs. La Sección 3 presenta la metodología para convertir la taxonomía de MDs en un sistema de clasificación. De los resultados (Sección 4) destacamos por un lado una alta precisión en un experimento de detección de errores en la taxonomía (93 % precisión y 78 % cobertura con 145 detecciones en 805 ensayos) y, por otro lado, en la clasificación de MDs (96 % precisión y 98 % cobertura en 619 ensayos). Las conclusiones del artículo y la discusión sobre los próximos pasos a seguir se encuentran en la Sección 5. Un sitio web acompaña al artículo, con documentación sobre avances y demostradores:

http://www.tecling.com/emad

## 2 Marco teórico

En principio, los marcadores del discurso son palabras de índole gramatical y no léxica, puesto que se utilizan para formar las construcciones gramaticales y no para representar de manera inmediata la realidad (Cuartero, 2002).

Si bien los MDs suelen mantener los comportamientos de las clases gramaticales de las que provienen, por ejemplo, las propiedades sintácticas (Fraser, 1999; Martín Zorraquino y Portolés, 1999), ni estos rasgos ni los semánticos son siempre suficientes o necesarios, por cuanto los MDs tienen un alcance operativo a nivel de enunciado y no a nivel oracional (Fraser, 1990; Lenk, 1997; Waltereit, 2006). Aceptamos que los MDs configuran una categoría pragmática, como muchos autores señalan (Fraser, 1996; Fraser, 1999; Pons, 1998; Martín Zorraquino y Portolés, 1999), que cumplen un papel fundamental en el procesamiento de la coherencia, la cohesión, la adecuación y la eficacia del discurso (Blakemore, 1987; Fraser, 1990; Montolío, 2001; Bazzanella, 2006; Maschler y Schiffrin, 2015) y que no aportan significado léxico a las proposiciones (Fraser, 1996; Lenk, 1997; Bazzanella, 2006). Desde el punto de vista

del enfoque procedimental (Sperber y Wilson, 1986; Blakemore, 1987), los MDs son concebidos como guías de las inferencias del interlocutor. En este sentido, los MDs funcionarían como señales metadiscursivas que señalan la estructura y la organización del discurso para beneficio del interlocutor.

La tarea de clasificarlos debe hacerse, entonces, según criterios funcionales y así han procedido muchos autores, por ejemplo, en el ámbito hispánico (Casado, 1993; Martín Zorraquino y Portolés, 1999; Pons, 2000; Montolío, 2001; Fuentes, 2009, entre otros). Cada uno de ellos, sin embargo, considera distintos elementos, conceptos y propiedades para categorizar los MDs.

## 2.1 Clasificaciones automáticas de MDs

Una clasificación automática puede suponer la introducción de un instrumento de medida objetivo como medio para superar las discusiones que conlleva la subjetividad inherente al método introspectivo, comúnmente usado en las clasificaciones manuales de MDs. Sin embargo, el principal obstáculo para una clasificación automática es la falta de consenso entre los especialistas sobre cuáles son las propiedades delimitadoras de la clase de los MDs. Alonso, Castellón, y Padró (2002) han atribuido esta falta de consenso a la preeminencia de las aproximaciones de tipo deductivo, con un sesgo importante por una teoría subyacente.

En la dédaca de 1990, se llevaron a cabo las primeras propuestas de mecanismos formales para detectar y sistematizar los MDs (Knott y Dale, 1995; Knott, 1996; Marcu, 1997). Ya en el nuevo siglo, Hutchinson (2004) utilizó métodos de aprendizaje automático supervisado para caracterizar conectores discursivos. Si bien obtuvo resultados de alta presición con respecto a un *gold standard*, el aprendizaje de los modelos dependió de instancias anotadas manualmente, lo que requiere de importante trabajo manual antes de la aplicabilidad del método y conlleva un sesgo relativo a los anotadores.

Un enfoque para solucionar este problema es el que adoptaron Alonso et al. (2002), quienes presentan la construcción de un lexicón computacional de marcadores discursivos, proponiendo la utilización de una técnica de *clustering* para agrupar instancias de uso de conectores extraídas de un gran corpus. El resultado es que los *clusters* obtenidos contienen, principalmente, instancias en las que los conectores tienen un comportamiento sintáctico similar. Si bien esta propuesta soluciona, en parte, los problemas anteriores, el hecho de que la selección de los atributos se haya hecho a partir de información sintáctica, semántica y retórica de un léxico de conectores codificado a mano, implica que las categorías aglomeradas sean apenas corroboradas con datos de corpus y no provengan de ellos de manera emergente.

Más recientemente, Muller et al. (2016) obtuvieron automáticamente *clusters* de conectores fundados empíricamente, basándose en la significación de la asociación entre conectores y pares de predicados verbales en contexto. Tal como en el caso anterior, los conectores se extrajeron de una lista codificada previamente.

## 2.2 La clasificación inductiva

En nuestro trabajo previo (Robledo, Nazar, y Renau, 2017), presentamos una propuesta de inducción automática de taxonomías de MDs a partir de un corpus paralelo. Propusimos un método para extraer ocurrencias parentéticas (desagregadas de la oración mediante signos de puntuación) de MDs desde un corpus paralelo español-inglés de 1.1 mil millones de tokens para inducir automáticamente categorías de MDs según la similitud entre los elementos, sin recurrir a ningún tipo de anotación previa.

El procedimiento involucró la alineación de los MDs en ambas lenguas aplicando estadísticas de coocurrencia sobre los segmentos alineados del corpus paralelo. Esto proporcionó un listado de MDs equivalentes en las dos lenguas, a partir del cual pudimos obtener los grupos de MDs con un misma función en español. Por ejemplo, si se ha visto en el corpus paralelo una asociación estadística entre las veces en que distintos traductores traducen *además* por *furthermore* y las veces en que se traduce *furthermore* por *asimismo*, suponemos que *además* y *asimismo* cumplen una misma función en el discurso.

Estas asociaciones semántico-pragmáticas fueron plasmadas en una matriz binaria (Tabla 1) que asigna un 0 o un 1 en función de si el MD aparece o no en su lista de equivalentes funcionales.

Sobre esta matriz binaria aplicamos un método de *clustering* aglomerativo con el que

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| a continuación | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| a su vez | - | **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| a veces | - | - | **1** | 0 | 0 | 0 | 0 | 0 | 0 |
| actualmente | - | - | - | **1** | 0 | 0 | 0 | 0 | 0 |
| además | - | - | - | - | **1** | 0 | 0 | 0 | **1** |
| ahora | - | - | - | - | - | **1** | 0 | 0 | 0 |
| ahora bien | - | - | - | - | - | - | **1** | 0 | 0 |
| al menos | - | - | - | - | - | - | - | **1** | 0 |
| asimismo | - | - | - | - | - | - | - | - | **1** |
| ... | - | - | - | - | - | - | - | - | ... |

Tabla 1: Fragmento de muestra de la matriz binaria en la que se basa el experimento. Es una matriz simétrica, por tanto, las columnas corresponden a los mismos marcadores de las filas y los datos debajo de la diagonal principal son redundantes

obtuvimos un total de 100 *clusters* de MDs en español según la similitud entre los elementos. A modo de ilustración, la Figura 1 muestra un ejemplo de resultado de la inducción de taxonomías de MDs en forma de dendrograma. Esta imagen fue realizada con una pequeña muestra aleatoria de 62 MDs para favorecer la legibilidad. El resultado total es un dendrograma mucho más complejo.

### 2.3 Polifuncionalidad de los MDs

A pesar del interés que puede presentar una clasificación inductiva por medio del método de *clustering*, encontramos una limitación importante que es que solo permite clasificar cada elemento en una sola categoría, de manera que este método de clasificación no permite dar cuenta de la polifuncionalidad de los MDs.

La polifuncionalidad es un fenómeno que se observa en muchos MDs y refiere a que un MD puede expresar distintos significados pragmáticos (Wierzbicka, 2003). De ahí que se investigue la posibilidad de identificar un significado nuclear, básico y estable, de un MD, del cual procedan matices eventuales y significados más contingentes, específicos del contexto de emisión (Aijmer, Foolen, y Vandenbergen, 2006). Así, por ejemplo, un MD como *es decir* puede funcionar como un reformulador explicativo (Martín Zorraquino y Portolés, 1999; Fuentes, 2009, por ejemplo), cumpliendo una función similar a la de *o sea* o *esto es*, pero, también puede funcionar, en ciertos contextos, como un conector consecutivo, cumpliendo una función similar a la de *por lo tanto* o *en consecuencia*.

Con la presente investigación es posible abordar la limitación anterior, es decir, dar cuenta, en alguna medida, de la polifuncionalidad de los MDs.

## *3 Metodología*

Tal como explicamos en la introducción, en este artículo planeamos utilizar una taxonomía ya generada para utilizarla luego como un sistema de clasificación de MDs. Esto lo conseguimos utilizando la matriz $M$ cuyo fragmento se expuso en la Tabla 1. Destacamos que en esta etapa de clasificación utilizamos las mismas métricas que para la generación del *clustering*, esto es, mismas características para generación de los vectores y misma distancia.

### 3.1 Primer paso: elaboración de una matriz $M$ de asociación entre MDs

Esta matriz traduce cada MD como un vector binario, lo que permite computar cálculos de similitud que reflejan el grado de asociación distribucional entre dos MDs. Dado un input $i$, que sería un determinado MD que se debe clasificar, procedemos a convertirlo en un vector binario utilizando los mismos pasos que en el caso de los marcadores que ya están en la matriz. El algoritmo 1 expone el pseudocódigo del proceso de vectorización.

**Algorithm 1** Vectorización de MDs.

**Require:** un marcador castellano $i$
1: Buscar $i$ en el corpus paralelo
2: Generar un conjunto $E$ de equivalentes de $i$ en inglés con el corpus paralelo
3: **for each** $j \in E$ **do**
4: Agregar a conjunto $S$ los equivalentes en castellano de $j$
5: **end for**
6: Generar vector binario $\vec{i}$ para $i$ registrando con valor 1 los elementos de $S$

El marcador discursivo se puede decir que es el elemento ideal para ser investigado en el corpus paralelo porque es independiente del contenido de los textos, por lo que se los encuentra en abundancia, y calcular la coocurrencia de los marcadores en los segmentos alineados es una tarea que no presenta dificultad. La ecuación 1 muestra el cálculo de coocurrencia, para un marcador en castellano $i$ y un candidato a equivalente en inglés $j$.

$$cooc(i,j) = \frac{f(i,j)}{\sqrt{f(i)}.\sqrt{f(j)}} \quad (1)$$

En términos simples, lo que esta medida hace es oponer las veces en que aparecen juntos con las veces en que aparecen separados, y esto da la pauta del grado de asociación de los elementos. Normalmente, un MD en castellano tiene múltiples equivalentes en inglés

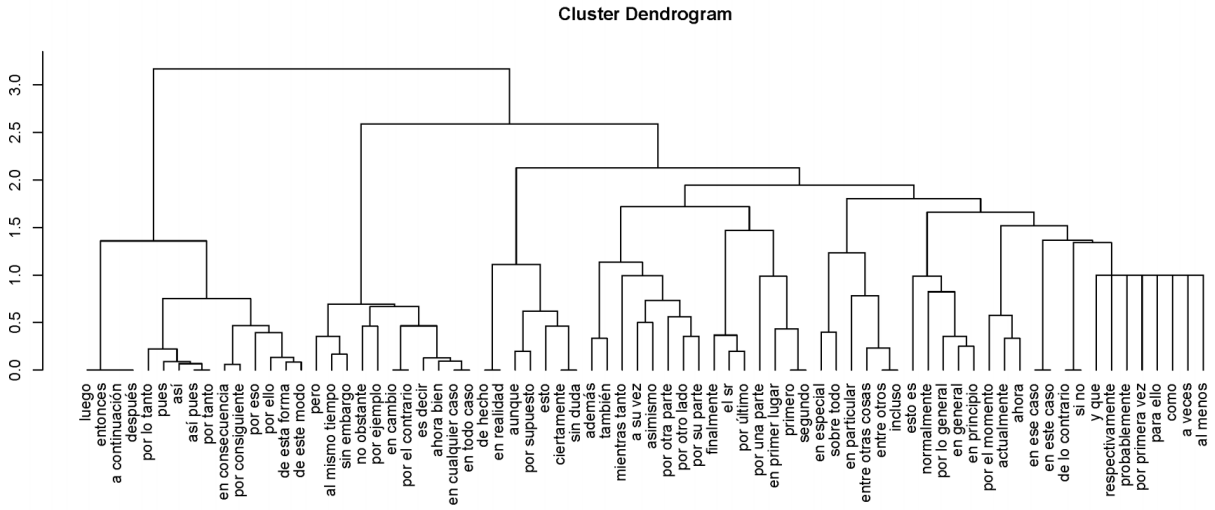


Figura 1: Ejemplo de resultado de inducción automática con una pequeña muestra aleatoria de 62 candidatos a MDs. El dendrograma total contiene 758 MDs, pero la taxonomía final que quedó después de una revisión manual alberga 619 de ellos. Esta figura no tuvo revisión manual y por eso muestra errores, como "el sr"

(o en cualquier otra lengua). Por tanto, de manera general seleccionamos los primeros $n$ equivalentes más probables ($n \approx 3$). Con los equivalentes en inglés (línea 4) hacemos exactamente lo mismo pero al revés, para obtener los equivalentes de nuevo en castellano. La Tabla 2 muestra un fragmento del resultado de esta operación para el caso del marcador contraargumentativo *sin embargo*. Se encuentran equivalentes, y esos equivalentes a su vez arrojan sus propios equivalentes, generando una asociación de los elementos en castellano, que son los que figuran en la tercera columna.

| **Castellano** | **Inglés** | **Castellano** |
|---|---|---|
| sin embargo | however | sin embargo<br>no obstante<br>ahora bien<br>con todo<br>pero |
| | nevertheless | no obstante<br>sin embargo<br>con todo<br>a pesar de ello |

Tabla 2: Algunos equivalentes en primer y segundo grado (inglés/castellano) de la entrada *sin embargo*

## 3.2 Segundo paso: utilización de la matriz $M$ para la clasificación de marcadores

Teniendo los elementos descritos en 3.1, procedimos a diseñar un sistema de clasificación, que resumimos en forma de pseudocódigo en el algoritmo 2.

**Algorithm 2** Clasificación de MDs.

**Require:** (marcador castellano $i$; taxonomía de marcadores $T$; matriz binaria $M$; umbral de similitud $u \approx 0{,}4$; hash table $S$)

1: **if** $i \notin M$ **then**
2: $\quad \vec{i}$ = vectorizar $i$
3: **end if**
4: **for each** $\vec{j} \in M$ **do**
5: $\quad$ **if** $i \neq j \wedge (Jaccard(\vec{i}, \vec{j}) > u)$ **then**
6: $\quad\quad S[T(j)] + +$
7: $\quad$ **end if**
8: **end for**
9: $T(i) = \arg\max(S)$

Esto es, además del elemento a clasificar ($i$), necesitamos la matriz $M$ y la taxonomía $T$, que contiene un total de 619 MDs en castellano clasificados en 19 categorías y que hemos revisado manualmente antes de comenzar este proceso para descartar errores. Las categorías de los *clusters* también fueron asignadas de manera manual con nombres que consideramos descriptivos del contenido, tales como "Refuerzo argumentativo" (ej.: *básicamente, en el fondo, en esencia, fundamentalmente, ...*); "Conclusivos" (ej.: *en definitiva, finalmente, para acabar, ...*), etc. Así, si $x$ = *en definitiva*, entonces $T(x)$ = "Conclusivo".

El algoritmo 2, entonces, compara el vector de $i$ con cada vector $j$ de la matriz $M$ (línea 4 del pseudocódigo). Esa comparación se realiza utilizando el coeficiente de *Jaccard*

(2), una medida apropiada para la comparación de vectores binarios.

$$Jaccard(\vec{i},\vec{j}) = \frac{|\vec{i} \cap \vec{j}|}{|\vec{i} \cup \vec{j}|} \quad (2)$$

Si la comparación arroja un resultado superior a un umbral arbitrario $u$ (línea 5 del pseudocódigo), entonces se obtiene la categoría de $j$ en la taxonomía $T$ y sumamos 1 a ese valor de la estructura de datos (*hash table*) $S$ (línea 6). Finalmente, seleccionamos la categoría de $S$ con el valor más alto.

### 3.3 Tercer paso: detección de polifuncionalidad

Para la detección de polifuncionalidad de los MDs hemos procedido de manera similar a la clasificación monocategorial, pero con una clasificación de más de una categoría. Es decir que se procede a seleccionar las $n$ categorías de $S$ con el valor más alto. Para ello, en la ecuación 3 se establece una función $P(x)$ con el criterio para el filtrado de una categoría $c$.

$$P(c) = \begin{cases} 1 & S[c] > w \wedge \frac{S[c]}{S[T(i)]} > z \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Los umbrales $w$ y $z$ son arbitrarios. Así, por ejemplo, si $w = 2 \wedge z = 0{,}1$, entonces solo aceptamos una categoría con un valor superior a 2, y si ese valor representa una proporción del 10 % del valor de la categoría ganadora.

## *4 Resultados*

Evaluamos primero la capacidad del algoritmo para distinguir entre lo que es y lo que no es un MD (Sección 4.1). Esto es porque si el algoritmo no puede proporcionar una categoría para un elemento $i$, se interpreta entonces que $i$ en realidad no es un MD, por tanto es una operación de depuración automática de la taxonomía. Medimos esto a través de la introducción deliberada de errores en la taxonomía. En tanto, para la evaluación de los resultados de la clasificación de MDs (Sección 4.2), procedimos a reclasificar cada uno de los 619 marcadores que están en la taxonomía $T$ y comparar automáticamente si el algoritmo de clasificación asigna a cada marcador la misma categoría que tiene en la taxonomía revisada manualmente, lo que arroja la precisión y la cobertura de la clasificación de MDs. Finalmente, evaluamos de manera cualitativa qué precisión tienen las clasificaciones que hace el algoritmo en segunda instancia (Sección 4.3).

| | **Detec. errores** | **Clasif. marcadores** |
|---|---|---|
| **Tp** | 145 | 587 |
| **Fp** | 10 | 22 |
| **Fn** | 41 | 10 |
| **Pre** | 93 % | 96 % |
| **Rec** | 78 % | 98 % |
| **F1** | 85 | 97 |

Tabla 3: Resultados de la evaluación: la columna 1 es la detección de "falsos marcadores", (elementos que no son MDs introducidos de forma deliberada en el listado.), y la columna 2 la clasificación de los MDs asignando una única categoría

### 4.1 Resultado detección de errores

Para la evaluación del desempeño del algoritmo en la operación de detección de errores procedemos con la taxonomía de MDs $T$ descrita en la Sección 3.

Para evaluar el desempeño de la detección de errores, hemos agregado un total de 186 falsos marcadores, para determinar cuántos de esos errores eran detectados frente a cuántos pasaban desapercibidos para el sistema. En la taxonomía hay entonces 619 MDs correctos y 186 errores, 805 en total.

La Tabla 3 muestra que el resultado arrojó una alta precisión en la detección de errores (93 % precisión, con 10 falsos positivos después de 805 ensayos). De esta manera hemos conseguido también depurar los resultados de nuestra taxonomía anterior, lo que nos permite entrar en un círculo de reproducción de los experimentos entrenando al clasificador nuevamente con una taxonomía más depurada.

### 4.2 Resultado clasificación de MDs

Para evaluar el desempeño en la clasificación, realizamos un total de 619 ensayos de clasificación tomando cada vez uno de los MDs y determinando si es posible clasificarlo en función de los 618 MDs ya clasificados, estrategia conocida como *leave-one-out cross validation* (Mitchell, 1997, p. 235). Con este procedimiento resulta sencillo evaluar los resultados de la monoclasificación, aquella en la que cada MD recibe una sola categoría, ya que se procede automáticamente mediante el contraste con el listado inicial.

En la Tabla 3 se muestran los resultados de esta clasificación. Como se puede ver, los resultados presentan una alta precisión en la

clasificación de MDs asignando una sola categoría (96 % precisión en 619 ensayos).

### 4.3 Resultado detección de polifuncionalidad

En esta tercera parte de la evaluación medimos el desempeño del algoritmo en el descubrimiento de segundas categorías propuestas por el clasificador, en tanto podrían ser indicativas de polifuncionalidad. La evaluación de la policlasificación, donde un mismo MD es asignado a distintas categorías, resulta más compleja. Aquí solo podemos proceder mediante el análisis cualitativo, aceptando o descartando las propuestas de clasificación del algoritmo según nuestro conocimiento de la materia.

Debido a que estos experimentos se hicieron en función de dos parámetros $w$ y $z$, en la Tabla 4 se exponen los resultados obtenidos según se modifiquen esos valores. La Ecuación 3 define $w$ como la intensidad de la asociación entre los elementos y $z$ como la proporción que existe entre la primera y la segunda categoría.

| **Precisión en detección de polifuncionalidad** | | | | | | |
|---|---|---|---|---|---|---|
| % | w=1 | w=2 | w=3 | w=4 | w=5 | w=6 |
| z=.1 | 28 | 28 | 31 | 28 | 25 | 50 |
| z=.2 | 35 | 28 | 31 | 29 | 29 | 50 |
| z=.3 | 36 | 28 | 33 | 30 | 25 | 50 |
| z=.4 | 32 | 29 | 30 | 30 | 27 | **75** |
| z=.5 | 30 | 30 | 30 | 30 | 27 | 75 |
| z=.6 | 31 | 31 | 31 | 31 | 27 | 75 |
| z=.7 | 33 | 33 | 33 | 33 | 28 | 66 |
| z=.8 | 66 | 66 | 66 | 66 | 66 | 0 |

Tabla 4: Resultados de la evaluación del proceso de policlasificación de los MDs, es decir, de aquellos casos en los que el clasificador indica que hay otra categoría, además de la principal

En el mejor de los casos, cuando el clasificador asigna una segunda categoría a un MD, muestra una precisión de 75 % en 4 ensayos. El número bajo de ensayos se explica porque a mayor $z$ y $w$, menor cantidad de ensayos posibles. Por ejemplo, en el caso de $z = 0{,}1$ y $w = 1$ tenemos 99 ensayos y en $z = 0{,}4$ y $w = 4$, tenemos 23. Como es natural, a medida que es más restrictivo, menos arriesga.

No evaluamos cobertura por no disponer de un marco de referencia, que podría ser el relevamiento manual previo de los elementos polifuncionales que existen actualmente en la taxonomía $T$, tarea que dejamos para el futuro.

## *5 Conclusiones*

En este artículo hemos presentado una propuesta para clasificar MDs en categorías funcionales generadas inductivamente en un estudio exploratorio previo. Utilizamos los resultados de ese estudio como datos de entrenamiento con el fin de generar un sistema que permita asignar un MD no incluido en la taxonomía inicial a una o más categorías emergentes. Con esto, hemos superado la limitación inicial de asignar un elemento a solo una categoría, con lo cual hacemos una primera aproximación al estudio de la polifuncionalidad de los MDs desde un enfoque inductivo.

A diferencia de trabajos anteriores que han propuesto clasificaciones automáticas de MDs (Alonso, Castellón, y Padró, 2002; Alonso et al., 2002; Hutchinson, 2004; Muller et al., 2016), no hemos partido de ninguna lista de marcadores codificada previamente y abarcamos no solo elementos de conexión, sino una variedad más amplia de MDs. Hemos destacado ya, además, la independencia de la lengua debido a la naturaleza cuantitativa del método. Con esto, creemos que este estudio complementa las clasificaciones previas con una aproximación derivada naturalmente de datos de la lengua en uso.

Además de consecuencias teóricas, este trabajo tiene varias aplicaciones prácticas tales como la segmentación discursiva, la extracción de información o la traducción automática, debido a que los MDs son importantes señales de la estructura del discurso (Popescu-Belis y Zufferey, 2006). Una vía de trabajo futuro será mejorar los resultados de clasificación múltiple de MDs con métodos de aprendizaje automático.

## *Agradecimientos*

Este trabajo ha sido posible gracias a una Beca Doctoral Conicyt otorgada por el Gobierno de Chile al primer autor. Agradecemos también a los revisores por sus comentarios.

## *Bibliografía*

Aijmer, K. 2002. *English discourse particles: Evidence from a corpus*. John Benjamins.

Aijmer, K., A. Foolen, y A.-M. Vandenbergen. 2006. Pragmatic markers in translation: a methodological proposal. En K. Fischer, editor, *Approaches to discourse particles*. Elsevier, páginas 101–114.

Alonso, L., I. Castellón, K. Gibert, y L. Padró. 2002. An empirical approach to discourse

markers by clustering. En *Proceedings of the 5th Catalonian Conference on AI: Topics in Artificial Intelligence*, páginas 173–183.

Alonso, L., I. Castellón, y L. Padró. 2002. Lexicón computacional de marcadores del discurso. *Procesamiento del lenguaje natural*, 29:239–246.

Bazzanella, C. 2006. Discourse markers in italian: towards a 'compositional' meaning. En K. Fischer, editor, *Approaches to discourse particles*. Elsevier, páginas 449–464.

Blakemore, D. 1987. *Semantic constraints on relevance*. Blackwell.

Brinton, L. J. 1996. *Pragmatic markers in English: Grammaticalization and discourse functions*. Mouton de Gruyter.

Casado, M. 1993. *Introducción a la gramática del texto del español*. Arco/Libros.

Cuartero, J. 2002. *Conectores y conexión aditiva. Los signos incluso, también y además en español actual*. Gredos.

Fischer, K. 2006. Towards an understanding of the spectrum of approaches to discourse particles: introduction to the volume. En K. Fischer, editor, *Approaches to discourse particles*. Elsevier, páginas 1–20.

Fraser, B. 1990. An approach to discourse markers. *Journal of pragmatics*, 14(3):383–398.

Fraser, B. 1996. Pragmatic markers. *Pragmatics*, 6(2):167–190.

Fraser, B. 1999. What are discourse markers? *Journal of pragmatics*, 31(7):931–952.

Fraser, B. 2006. Towards a theory of discourse markers. En K. Fischer, editor, *Approaches to discourse particles*. Elsevier, páginas 189–204.

Fuentes, C. 2009. *Diccionario de conectores y operadores del español*. Arco/Libros.

Hutchinson, B. 2004. Acquiring the meaning of discourse markers. En *Proceedings of the 42nd Annual Meeting on ACL*, página 684.

Jucker, A. y Y. Ziv. 1998. *Discourse Marker: Description and Theory*. John Benjamins.

Knott, A. 1996. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. Ph.D. tesis, University of Edinburgh.

Knott, A. y R. Dale. 1995. Using linguistic phenomena to motivate a set of coherence relations. *Discourse processes*, 18(1):35–62.

Lenk, U. 1997. Discourse markers. En J. Verschueren, editor, *Handbook of pragmatics. Installment*. John Benjamins, páginas 1–17.

Loureda, Ó. y E. Acín. 2010. Cuestiones candentes en torno a los marcadores del discurso en español. En Ó. Loureda y E. Acín, editores, *Los estudios sobre marcadores del discurso en español, hoy*. Arco/Libros, páginas 7–59.

Marcu, D. 1997. From discourse structures to text summaries. En *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, páginas 82–88.

Martín Zorraquino, M. A. y J. Portolés. 1999. Los marcadores del discurso. En I. Bosque y V. Demonte, editores, *Gramática descriptiva de la lengua española, Vol. 3*. Espasa-Calpe, páginas 4051–4213.

Maschler, Y. y D. Schiffrin. 2015. Discourse markers: Language, meaning, and context. En D. Tannen H. E. Hamilton, y D. Schiffrin, editores, *The handbook of discourse analysis*. John Wiley & Sons, páginas 189–221.

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

Montolío, E. 2001. *Conectores de la lengua escrita: contraargumentativos, consecutivos, aditivos y organizadores de la información*. Ariel.

Muller, P., J. Conrath, S. Afantenos, y N. Asher. 2016. Data-driven discourse markers representation and classification. En *TextLink–Structuring Discourse in Multilingual Europe. Károli Gáspár University of the Reformed Church in Hungary, Budapest*, página 93.

Pons, S. 1998. *Conexión y conectores: estudio de su relación en el registro informal de la lengua. Anejo XXVII de la revista Cuadernos de filología*. Universitat de València.

Pons, S. 2000. Los conectores. En A. Briz y Val.Es.Co, editores, *¿Cómo se comenta un texto coloquial?* Ariel, páginas 193–220.

Popescu-Belis, A. y S. Zufferey. 2006. Contrasting the automatic identification of two discourse markers in multiparty dialogues. *ISSCO Working Paper 65*.

Robledo, H., R. Nazar, y I. Renau. 2017. Un enfoque inductivo y de corpus para la categorización de los marcadores del discurso en español. En *Proceedings of the 5th International Conference "Discourse Markers in Romance Languages: Boundaries and Interfaces"*, páginas 91–93. Université Catholique de Louvain, Belgium.

Schiffrin, D. 1987. *Discourse markers*. Cambridge University Press.

Sperber, D. y D. Wilson. 1986. *Relevance: communication and cognition*. Harvard University Press.

Waltereit, R. 2006. The rise of discourse markers in italian: a specific type of language change. En K. Fischer, editor, *Approaches to discourse particle*. Elsevier, páginas 61–76.

Wierzbicka, A. 2003. *Cross-Cultural Pragmatics: The Semantics of Human Interaction*. Mouton/de Gruyter.

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 117-124 recibido 04-04-2018 revisado 30-04-2018 aceptado 04-05-2018

# Lexicon Adaptation for Spanish Emotion Mining

## *Adaptación de lexicones para la minería de emociones en Español*

**Flor Miriam Plaza-del-Arco, María Dolores Molina-González, Salud María Jiménez-Zafra, M. Teresa Martín-Valdivia**,
Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{fmplaza, mdmolina, sjzafra, maite}@ujaen.es

**Abstract:** Emotion mining is an emerging task that is still at a first stage of research. Most of the existing works and resources focus on English, but there are other languages, such as Spanish, whose presence on the Internet is greater every day. In WASSA-2017 Shared Task on Emotion Intensity, it was found that the best systems included features from affect lexicons. This fact combined with the scarcity of resources in Spanish, led us to build a new Spanish lexicon that has been tested over the dataset released at SemEval 2018 Task 1. Moreover, it has been compared with the unique emotion intensity lexicon existing in Spanish, SEL lexicon, and it has shown the difficulty of the task and the importance of continuing working on the development of resources.
**Keywords:** emotion mining, emotion intensity, lexicon, iSAL, SEL

**Resumen:** La minería de emociones es una tarea emergente que todavía se encuentra en una primera etapa de investigación. La mayoría de los trabajos y recursos existentes se han realizado para textos en inglés, pero la presencia en Internet de otras lenguas, como el español, es cada vez mayor. En la tarea *Shared Task on Emotion Intensity* de la competición WASSA-2017 se llegó a la conclusión de que los sistemas que mejor realizaban la clasificación de intensidad de las emociones eran aquellos que incluían características de lexicones afectivos. Este hecho, combinado con la escasez de recursos en español, nos llevó a construir un nuevo lexicón para el español que ha sido probado sobre el conjunto de datos liberado en la tarea 1 de la competición SemEval 2018. Además, se ha comparado con el único lexicón de intensidades existente en español, el lexicón SEL, y se ha demostrado la dificultad de la tarea y la importancia de continuar trabajando en el desarrollo de recursos.
**Palabras clave:** minería de emociones, intensidad de la emoción, lexicón, iSAL, SEL

## *1 Introduction*

Sentiment Analysis (SA) is an area of Natural Language Processing (NLP) that is focused on identifying, extracting, quantifying, and studying affective states and subjective information (Liu, 2015). SA includes the study of different tasks, from the simpler opinion detection to the more complex emotion mining (Cambria, 2016).

Although SA is a relative new discipline, an extensive number of research has been focused on some basic tasks in SA, such as polarity classification or subjectivity detection, even treating complex problems related to different languages or domains (Molina-González et al., 2015; Jiménez-Zafra et al., 2016; Montejo-Ráez et al., 2014). However, emotion mining in textual documents that studies emotional states such as "angry", "sad", and "happy", is still in a first stage of research and it has a long way to proceed (Yadollahi, Shahraki, and Zaiane, 2017).

Some works are starting to explore the potential of emotion detection and emotion classification systems (Mohammad, 2017) but, as usual, most of the studies are oriented to treat English documents. However, it is necessary to adapt the systems in order

ISSN 1135-5948. DOI 10.26342/2018-61-13

to develop real emotion mining applications for a specific language. In this paper, we focus on the emotion mining task and specifically for Spanish documents. Unfortunately, one of the main problems to be resolved is the generation and integration of specific resources.

In the case of English there are some interesting resources that can be integrated into real systems such as WordNet-Affect (WNA) (Strapparava and Valitutti, 2004), Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Booth, and Francis, 2007) or NRC word-emotion association lexicon (Mohammad and Turney, 2010). All these lexicons can be applied to develop English emotion recognition systems. However, when we move to other languages different from English, the first problem to be resolved is the lack of resources. Thus, in this paper, we focus on adapting the NRC English resource for emotion recognition in order to be used for Spanish emotion systems. We have applied several strategies combining automatic machine translation and manual revision in order to increase the recall and the precision.

In order to test the effectiveness of the resources generated, we have conducted several experiments over the Semeval 2018 corpora (Mohammad et al., 2018). Specifically we have focused on the EI-oc emotion intensity ordinal classification task.

The rest of the paper is organized as follows: Section 2 describes some related studies; dataset and lexicon adaptation for Spanish are presented in Section 3; Section 4 shows the results and discussion, and finally, our conclusions are presented in Section 5.

## 2 Background

In the last decade, most of the work has focused on SA whose most important task is polarity classification. Pang and Lee (2008) give an excellent summary. However, one of the most complex areas that has not yet been studied in depth is Emotion Mining.

Emotion mining has been studied from many disciplines such as neuroscience, cognitive sciences, and psychology. However, only recently this area has attracted the attention in computer science perhaps due to the multiple and interesting applications (Yadollahi, Shahraki, and Zaiane, 2017). Gupta, Gilbert, and Fabbrizio (2013) describe a method that uses salient features to identify emotional emails in the customer care domain. In this way they improve contact center efficiency and the quality of the overall customer care experience. In Human Computer Interaction, the systems can monitor user's emotions to suggest suitable music or movies (Voeffray, 2011). In the field of psychology there are several works of great application. De Choudhury et al. (2013) detect if a patient is facing depression, stress or even is thinking about committing suicide. This fact is quite useful since the user can be referred to counseling services (Luyckx et al., 2012). Moreover, this area is becoming very popular, and some of the main conferences dealing with data mining and evaluation are currently including workshops and share tasks related to it. These include Semantic Evaluation (SemEval), Computational Approaches to Subjectivity and Sentiment Analysis (WASSA) and workshops on Computational Modeling of People's opinions, personality and emotions in Social Media (PEOPLE).

Emotion recognition is part of the broader area of Emotion Mining with aims to enable computers recognize and express emotions (Picard and others, 1995). Emotion mining techniques can be classified into two categories: lexicon based approaches and machine learning approaches (Cambria, 2016). The first one is based on lexical resources such as lexicons, bags of words or ontologies. Kim, Valitutti, and Calvo (2010) follow lexical-based approaches to evaluate the merit of the discrete emotion theory and the dimensional model using the WNA lexicon. The second one applies algorithms based on linguistic features. Luyckx et al. (2012) focus on a dataset of notes written by people who have committed suicide. The objective is to predict label(s) of a note among 15 possible emotions. First, they split all multi-labeled notes to single-labeled fragments manually. Then a Support Vector Machine (SVM) with Radial Basis Function is trained on these single-labeled data.

Almost all the emotion mining works focus on using and integrating different resources such as lexicons and corpora. Specifically, affect lexicons are very valuable because they provide prior information about the type and strength of emotion carried by each word of the text. Actually, in WASSA-2017 Shared Task on Emotion Intensity it was demon-

strated that using features from affect lexicons is beneficial for emotion mining tasks (Mohammad and Bravo-Marquez, 2017).

However, the availability of resources in textual emotion mining is scarce and most of them are for English language. For example, WordNet-Affect (WNA) (Strapparava and Valitutti, 2004) is an emotional lexical resource based on the synsets of WordNet (Fellbaum, 1998) that has been applied in several studies. WNA contains a set of affective concepts correlated with affective words in English and one of its main problem is the low recall. Linguistic Inquiry and Word Count (LIWC) is another lexical resource that divides the words into different categories including emotional states (Pennebaker, Booth, and Francis, 2007). Finally, NRC word-emotion association lexicon (Mohammad and Turney, 2010) is a resource generated using a crowdsourcing annotation by Amazon's Mechanical Turk and provides real-valued affect intensity scores for four basic emotions (*anger*, *fear*, *sadness*, *joy*). This is one of the most popular emotion lexicons because it has been integrated in different systems combined with other resources or supporting machine learning approaches.

Regarding the availability of emotional resources in other languages different from English, we find that the number is very limited. Specifically, for Spanish we can metion the Spanish Emotional Lexicon (SEL) (Sidorov et al., 2012) although the obtained results in different experiments are not very promising.

In this work, we focus our attention on the relatively small amount of work in the generation of lexicons and on computational analysis of the emotional content of texts. We have adapted the NRC Affect intensity lexicon to Spanish, obtaining different versions that have been evaluated over the dataset released at SemEval 2018 Task 1: Affect in Tweets (Mohammad et al., 2018).

## *3 Resources*

### 3.1 Dataset

To run our experiments, we used the Spanish dataset provided by the organizers in SemEval 2018 Task 1: Affect in Tweets (Mohammad et al., 2018). The dataset EI-oc is composed by a set of tweets that belong to an emotion E (*anger*, *fear*, *joy*, and *sadness*). Thus, separate datasets are provided for *anger*, *fear*, *joy*, and *sadness* emotions. The complete dataset is composed by 1,986 tweets for *anger*, 1,986 tweets for *fear*, 1,990 tweets for *joy* and 1,991 tweets for *sadness*. We trained our models on the train and dev sets, and tested the model on the test set. Table 1 shows the number of tweets for Spanish language used in our experiments.

| Dataset | train+dev | test | Total |
|---|---|---|---|
| anger | 1,359 | 627 | 1,986 |
| fear | 1,368 | 618 | 1,986 |
| joy | 1,260 | 730 | 1,990 |
| sadness | 1,350 | 641 | 1,991 |
| EI-oc | 5,337 | 2,616 | 7,593 |

Table 1: Number of tweets per dataset

### 3.2 Lexicon generation

In a first instance we generated a parallel list of affect terms in Spanish from the affect lexicon in English (NRC Affect Intensity Lexicon) provided by Mohammad (2017). NRC is composed of 5,814 words (1,483 *anger* words, 1,765 *fear* words, 1,268 *joy* words and 1,248 *sadness* words). Each of them has an intensity score associated to one of the following basic emotions: *anger*, *fear*, *joy* and *sadness* and does not belong to any domain. The score ranges from 0 to 1, where 1 indicates that the word has a high association to the emotion and 0 that the word has a low association to the emotion. Our parallel list was generated by applying automatic machine translation techniques. Google translator was used for the automatic translation, taking into account the first translated word that this system returned for each original word from NRC. According to Google Translator, this first translated word is the most frequently used regardless of grammatical category. The new resource is called SAL (Spanish Affect Lexicon). Following, each translated word was written by using non capital letters. Thus, SAL is composed of 5,814 affect translations, with the same proportion of affect terms that the NRC list. During this process, we have found some issues that must be resolved. These issues are the following:

(i) 1,267 repeated Spanish terms

(ii) 110 English words untranslated

(iii) 45 English expression untranslated

(iv) 326 n-grams terms

The solution of each problem led us to the generation of new versions of the SAL list.

Below, we describe the improved SAL versions (iSAL) that we have generated automatically:

- **iSALv1a**: The first automatic version solves the issue related to the repetition of terms in general. We noticed that some English words shared the same first translation. For this reason, we decided to discard all of them from the SAL list in our first approach. The resulting list was called iSALv1a (improved SAL version 1 automatic) and it is composed of 2,338 unique affective Spanish terms with only one emotion associated.
- **iSALv2a**: The second automatic version improves the issue related to the repetition of terms. In this case, we considered that the repeated terms with intensity score associated in different emotions must be included. Thus, the new list is the sum of iSALv1a and 2,053 Spanish terms. In total, iSALv2a is composed of 4,391 Spanish terms. Table 2 shows some English words that shared the same first translation in different emotions.

| English word | Spanish meaning | Emotion |
|---|---|---|
| abandonment | abandonando | anger, fear, sadness |
| hellish | infernal | anger, fear, sadness |
| unbeaten | invicto | sadness, joy |
| youth | juventud | anger, fear,joy |
| treat | tratar | anger, fear, joy, sadness |

Table 2: Some English words that shared the same first translation in different emotions

- **iSALv3a**: The third automatic version improves the issue related to the repetition of terms too. In this case, we considered that the repeated terms with different intensity scores associated to the same emotion must be taken into account. The adopted solution was to compute the average of the intensities of each equal term in the same emotion and to include an unique term associated to that emotion with the average intensity. Thus, the new list is the sum of iSALv2a and 764 Spanish terms. In total, iSALv3a is composed of 5,155 Spanish terms. Table 3 shows some English words that shared the same first translation in the same emotion.

| English words | Spanish meaning | Emotion |
|---|---|---|
| murderer, murderous, assassin, slayer, cutthroat | asesino | anger, fear |
| harass, harry, harassing | acosar | anger |
| cheerfull, jolly, joyful, glad, merry, cheery, rollicking | alegre | joy |
| dilapidated, bankrupt, blighted, ruined | arruinado | sadness |
| executioner, hangman | verdugo | fear |

Table 3: Some English words that shared the same first translation in the same emotion

After resolving the issues related to the repeated terms in SAL, the next step in the process of the generation of an affect resource in Spanish was to solve manually the problems (ii), (iii) and (iv). We found many misspelled words or expressions in the NRC list, which should not be considered mistakes and could be improved manually. Therefore the new generated lexicons improved the previous versions.

These are the versions of improved SAL (iSAL) that were manually generated:

- **iSALv1m**: In the 2,338 terms of iSALv1a we found 110 English words and 45 expressions untranslated. It is important to note that all words did not have a properly assignment and therefore, they have been removed from the original list iSALv1a to generate the iSALv1m list. Then, from the 110 English words, only 33 had a manual as-

signment according to the Spanish language. From the 45 expressions, only 40 had a manual assignment according to the Spanish language. Tables 4 and 5 show some examples of this kind of words.

Finally, the last issue was related to the fact that the translation of an English word returned two or more words (n-grams). For these cases we assigned manually the best synonym (composed of only one term) for the translated word. In the iSALv1a list we found 266 n-grams but only 237 had a synonym composed of one term. Table 6 shows some examples of some manually reviewed translations. Thus, the new list iSALv1m is composed of 2,227 Spanish terms.

| English word without translation | Spanish meaning |
|---|---|
| cantbreathe | asfixia |
| wracking | exprimiendo |
| sux | chupar |
| stoopid | estupido |
| grump | gruñón |

Table 4: Some English words without translation

| English expression | Spanish meaning |
|---|---|
| xoxo | bss |
| hee | eh |
| woohoo | viva |
| meh | bah |
| hohoho | jojojo |

Table 5: Some English expressions without translation

| English word | n-gram Spanish meaning | Spanish term |
|---|---|---|
| fuming | echando humo | encolerizar |
| makememad | me enfurece | enloquecerme |
| uphill | cuesta arriba | agotador |
| astray | por mal camino | descarriado |

Table 6: Some Spanish n-grams found in iSALv1a

- **iSALv2m**: The second manually version improves the issue related to the n-grams. We found 45 n-grams in the 2,053 Spanish terms that were repeated in different emotions, but only 39 of them have been modified by only one synonym term. Table 7 shows some examples of this kind of words. Finally, iSALv2m are composed of iSALv1m and the 2,031 terms improved manually, in total 4,258 words.

| English words | n-gram Spanish meaning | Spanish term |
|---|---|---|
| forcibly | a la fuerza | forzosamente |
| misconception | idea equivocada | malentendido |
| wince | contraerse de dolor | estremecerse |
| disreputable | de mala fama | desacreditado |
| landslide | deslizamiento de tierra | derrumbre |

Table 7: Some Spanish n-grams found in iSALv2a

- **iSALv3m**: The third manually version improves the issue related to the n-grams, as well. We found 15 n-grams in the 764 Spanish terms that were repeated in the same emotion. Some of these n-grams have been modified by only one synonym term. Finally, iSALv3m is composed of iSALv2m and the 754 terms improved manually, in total 5,012 words.

## *4 Experiments and result analysis*

In this section we describe the systems developed to test the different versions of iSAL lexicon in the SemEval's EI-oc task (Mohammad et al., 2018). Moreover, we analyze the results obtained.

### 4.1 Experiments

EI-oc is an emotion intensity ordinal classification task. Given a tweet and an emotion E, it consists of classifying the tweet into one of four ordinal classes of intensity of E that best represents the mental state of the tweeter. Separate datasets are provided for *anger*, *fear*, *joy*, and *sadness* emotions.

First, we preprocessed the corpus of tweets. We applied the following preprocess-

ing steps: the documents were tokenized using NLTK TweetTokenizer[1], stemming was performed using NLTK Snowball stemmer[2] and all letters were converted to lower-case.

To solve this task we have used the following methodologies:

- **Heuristic 1 (H1)**. To perform the classification, we checked the presence of lexicon terms in the tweet and then we added the intensity value of these words grouping them by the emotional category (*anger*, *fear*, *sadness* and *joy*). The result is a vector of four values for each lexicon. Moreover, each tweet is represented as a vector of unigrams using the TF-IDF weighting scheme. The union of the lexicon vectors and the TF-IDF representation of the tweet are used as features for the classification using the SVM algorithm. We selected the SVM formulation, known as C-SVC, the value of the C parameter was 1.0 and the kernel chosen was linear.
- **Heuristic 2 (H2)**. In this case, we checked the presence of lexicon terms in the tweet and then we computed the sum, the average and the maximum of the intensity value of the words of the tweet grouping them by the emotional category (*anger*, *fear*, *sadness* and *joy*). The result is a vector of twelve values for each lexicon. The union of the lexicon vectors and the TF-IDF representation of the tweet are used as features for the classification using the SVM algorithm with the same configuration as that used in the first methodology.

For each version of iSAL we applied the two methodologies described above. The official competition metric to evaluate the systems in EI-oc subtask is the Pearson Correlation Coefficient (PCC) between semantic similarity scores of machine assigned and human judgments. The results of our experiments are shown in Tables 8, 9, 10, 11, 12 and 13.

On the other hand, we have also performed the experiments with the unique emotion intensity lexicon existing in Spanish, the SEL lexicon, to compare the results. The results are shown in Table 14.

| | H1 | H2 |
|---|---|---|
| anger | 0.395 | 0.403 |
| fear | 0.54 | 0.524 |
| sadness | 0.500 | 0.493 |
| joy | 0.513 | 0.514 |
| macro-avg | 0.487 | 0.483 |

Table 8: Results of experiments using iSALv1a

| | H1 | H2 |
|---|---|---|
| anger | 0.403 | 0.399 |
| fear | 0.542 | 0.546 |
| sadness | 0.507 | 0.506 |
| joy | 0.516 | 0.517 |
| macro-avg | 0.492 | 0.492 |

Table 9: Results of experiments using iSALv1m

| | H1 | H2 |
|---|---|---|
| anger | 0.418 | 0.409 |
| fear | 0.528 | 0.534 |
| sadness | 0.504 | 0.492 |
| joy | 0.492 | 0.49 |
| macro-avg | 0.486 | 0.481 |

Table 10: Results of experiments using iSALv2a

| | H1 | H2 |
|---|---|---|
| anger | 0.418 | 0.405 |
| fear | 0.524 | 0.522 |
| sadness | 0.514 | 0.500 |
| joy | 0.514 | 0.510 |
| macro-avg | 0.493 | 0.484 |

Table 11: Results of experiments using iSALv2m

| | H1 | H2 |
|---|---|---|
| anger | 0.404 | 0.398 |
| fear | 0.519 | 0.520 |
| sadness | 0.499 | 0.499 |
| joy | 0.517 | 0.515 |
| macro-avg | 0.485 | 0.483 |

Table 12: Results of experiments using iSALv3a

## 4.2 Result analysis

If we take a look at the results it can be seen that the manual versions of the lexicons work better than the automatic ones. Specifically,

[1] http://www.nltk.org/api/nltk.tokenize.html

[2] http://www.nltk.org/_modules/nltk/stem/snowball.html

| | H1 | H2 |
|---|---|---|
| anger | 0.401 | 0.391 |
| fear | 0.521 | 0.519 |
| sadness | 0.496 | 0.500 |
| joy | 0.529 | 0.537 |
| macro-avg | 0.487 | 0.487 |

Table 13: Results of experiments using iSALv3m

| | H1 | H2 |
|---|---|---|
| anger | 0.402 | 0.395 |
| fear | 0.545 | 0.540 |
| sadness | 0.478 | 0.495 |
| joy | 0.489 | 0.502 |
| macro-avg | 0.479 | 0.483 |

Table 14: Results of experiments using SEL

iSALv2m is the best lexicon. There are expressions that are specific for each language and even they are used with different purposes and intensities, making the manual versions work better.

Focusing on the emotions, the lowest correlation has been obtained on *anger* emotion and the best correlation on *fear* emotion. On the contrary, in WASSA-2017 Shared Task on Emotion Intensity (Mohammad and Bravo-Marquez, 2017), most of the systems performed better on *anger* emotion and worse on *fear* and *sadness* emotions. In this competition, it was found that the best systems included features from affect lexicons. This fact combined with the scarcity of resources in Spanish encourage us to build a new Spanish lexicon.

In order to compare our results, we have accomplished experiments over the emotion intensity lexicon existing for Spanish, SEL (Sidorov et al., 2012). It can be seen that most of the results obtained with the proposed lexicons are better than those of the SEL lexicon, but they are only slightly better. This shows the difficulty of the task and the importance of continuing working on the development of resources for languages other than English. In addition, we have evaluated our results with the obtained in the SemEval's EI-oc task[3]. If we participated, we would be in sixth position in the ranking as can be seen in the summary Table 15.

[3] `https://bit.ly/2Gj8Anm`

| (r) Team name | Pearson | | | | |
|---|---|---|---|---|---|
| | macro-avg | anger | fear | joy | sadness |
| (1) AffectThor | 0.664 | 0.606 | 0.706 | 0.667 | 0.667 |
| (5) UWB | 0.504 | 0.361 | 0.606 | 0.544 | 0.506 |
| (6) AIT2018 Organizers | 0.481 | 0.444 | 0.546 | 0.451 | 0.483 |
| (15) AIT2018 Organizers | -0.022 | 0.011 | -0.069 | -0.005 | -0.027 |

Table 15: Results of SemEval's EI-oc task in Spanish language

## 5 Conclusion

In this work, it has been presented the process performed to generate a new emotion intensity lexicon for Spanish. We have generated a parallel list to the NRC Affect Intensity Lexicon for English (Mohammad, 2017) by automatically translating the terms with Google translator. In the generated list (SAL list) we found some translation problems, such as repeated Spanish terms, English words untranslated and English expression untranslated. This led us to the generation of new versions of the SAL list using an automatic approach and a manual approach.

The different versions of the lexicon have been tested over the dataset released at SemEval 2018 Task 1 and iSALv2m has been the one that has provided the best results. Moreover, the generated lexicon has been compared with the unique emotion intensity lexicon existing for Spanish, SEL lexicon, and it has shown that emotion intensity is a difficult task that is still at a first stage of research and that is very important to continuing working on the development of resources.

## Acknowledgements

This work has been partially supported by a grant from the Ministerio de Educación Cultura y Deporte (MECD - scholarship FPU014/00983), Fondo Europeo de Desarrollo Regional (FEDER) and REDES project (TIN2015-65136-C2-1-R) from the Spanish Government.

## References

Cambria, E. 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107.

De Choudhury, M., M. Gamon, S. Counts, and E. Horvitz. 2013. Predicting depression via social media. *ICWSM*, 13:1–10.

Fellbaum, C. 1998. *WordNet*. Wiley Online Library.

Gupta, N., M. Gilbert, and G. D. Fabbrizio. 2013. Emotion detection in email customer care. *Computational Intelligence*, 29(3):489–505.

Jiménez-Zafra, S. M., M. T. Martín-Valdivia, E. Martínez-Cámara, and L. A. Ureña-López. 2016. Combining resources to improve unsupervised sentiment analysis at aspect-level. *Journal of Information Science*, 42(2):213–229.

Kim, S. M., A. Valitutti, and R. A. Calvo. 2010. Evaluation of unsupervised emotion models to textual affect recognition. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 62–70. Association for Computational Linguistics.

Liu, B. 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Luyckx, K., F. Vaassen, C. Peersman, and W. Daelemans. 2012. Fine-grained emotion detection in suicide notes: A thresholding approach to multi-label classification. *Biomedical informatics insights*, 5:BII–S8966.

Mohammad, S. M. 2017. Word affect intensities. *arXiv preprint arXiv:1704.08798*.

Mohammad, S. M. and F. Bravo-Marquez. 2017. Wassa-2017 shared task on emotion intensity. *arXiv preprint arXiv:1708.03700*.

Mohammad, S. M., F. Bravo-Marquez, M. Salameh, and S. Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA.

Mohammad, S. M. and P. D. Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics.

Molina-González, M. D., E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña-López. 2015. A spanish semantic orientation approach to domain adaptation for polarity classification. *Information Processing & Management*, 51(4):520–531.

Montejo-Ráez, A., E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña-López. 2014. A knowledge-based approach for polarity classification in twitter. *Journal of the Association for Information Science and Technology*, 65(2):414–425.

Pang, B. and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Pennebaker, J. W., R. J. Booth, and M. E. Francis. 2007. Liwc2007: Linguistic inquiry and word count. *Austin, Texas: liwc. net*.

Picard, R. W. et al. 1995. Affective computing.

Sidorov, G., S. Miranda-Jiménez, F. Viveros-Jiménez, A. Gelbukh, N. Castro-Sánchez, F. Velásquez, I. Díaz-Rangel, S. Suárez-Guerra, A. Treviño, and J. Gordon. 2012. Empirical study of machine learning based approach for opinion mining in tweets. In *Mexican international conference on Artificial intelligence*, pages 1–14. Springer.

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

Voeffray, S. 2011. Emotion-sensitive human-computer interaction (hci): State of the art-seminar paper. *Emotion Recognition*, pages 1–4.

Yadollahi, A., A. G. Shahraki, and O. R. Zaiane. 2017. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):25.

# *Proyectos*

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 127-130 recibido 20-03-2018 revisado 28-04-2018 aceptado 03-05-2018

# Plataforma inteligente para la recuperación, análisis y representación de la información generada por usuarios en Internet

## *Intelligent framework for retrieving, analysing and representing user generated content on Internet*

**Yoan Gutiérrez, José M. Gómez, Fernando Llopis, Lea Canales, Antonio Guillén**
GPLSI - Universidad de Alicante San Vicente del Raspeig, s/n, 03690, Alicante
{ygutierrez, jmgomez, llopis, lcanales, aguillen}@dlsi.ua.es

**Resumen:** Este proyecto viene motivado por la necesidad de definir una plataforma basada en Tecnologías del Lenguaje Humano que sea capaz de procesar la información de manera inteligente y de forma automática, combinando múltiples técnicas y herramientas. Dicha plataforma flexibilizará el modo de mostrar visualmente los datos resultantes para ser adaptados a las necesidades de los usuarios desde un punto de vista analítico. El avance científico de cada una de las tecnologías involucradas en la creación de la plataforma propuesta, así como su combinación e integración en una única infraestructura, supondrá un paso importante dentro de las tecnologías del lenguaje humano, siendo a su vez, de valiosa utilidad para la sociedad actual y futura.
**Palabras clave:** Tecnologías del Lenguaje Humano, Plataforma Inteligente, Tecnologías Integradas, Analitícas

**Abstract:** This project is motivated by the need of defining a platform based on Human Language Technologies capable of intelligently processing textual information, by combining multiple techniques and tools. In addition, the way of displaying the obtained results will be adapted to the users needs from an analytical point of view. The scientific progresses of each technology involved, as well as their combination and integration in a single infrastructure, will contribute to the progress of human language technologies, being in turn of valuable use for the current and future society.
**Keywords:** Human Language Technologies, Intelligent Platform, Processing Textual Content, Integrating Technologies, Analytics

### *1 Introducción*

Actualmente, Internet cuenta con más de 4,156 millones de usuarios[1], dato que indica que el 54,4 % de la población mundial está conectada a la red de redes y por consiguiente consumiendo y generando información.

La web 2.0 o web social supone uno de los mayores atractivos para los usuarios de internet. Si se analizan los datos de dos de las redes sociales más conocidas hoy en día, Twitter[2] y Facebook[3], encontramos que Twitter cuenta con más de 310 millones de usuarios activos, que generan 500 millones de tweets al día[4], mientras que Facebook cuenta con más de 1.860 millones de usuarios[5] y más de 78 millones de páginas[6]. Si a ello le sumamos el resto de los sitios Web, incluyendo otros tipos de redes sociales, páginas Web, enciclopedias, blogs, foros, contenido multimedia, etc. encontramos más de 4.26 billones de páginas Web indexadas[7].

Sin embargo, el principal inconveniente de toda esta gran cantidad de información disponible es la complejidad para poder analizarla, sobre todo si la persona interesada desea obtener información precisa sobre da-

[1] http://www.internetworldstats.com/stats.htm (Marzo 2017)
[2] https://twitter.com (Marzo 2017)
[3] https://www.facebook.com (Marzo 2017)
[4] http://www.internetlivestats.com/twitter-statistics (Marzo 2017)
[5] https://www.trecebits.com/2017/02/02/facebook-ya-tiene-1-860-millones-de-usuarios/ (Febrero 2017)
[6] http://www.statisticbrain.com/facebook-statistics/ (Marzo 2018)
[7] http://www.worldwidewebsize.com/ (Febrero, 2018)

ISSN 1135-5948. DOI 10.26342/2018-61-14

tos formulados en lenguaje natural que necesiten ser interpretados.

Un modo de reducir el tiempo invertido por los usuarios en analizar grandes cantidades de información es mediante el uso de las Tecnologías del Lenguaje Humano (TLH).

Las herramientas y recursos de TLH desarrollados en los últimos años han permitido mejorar los procesos de búsqueda, recuperación y extracción de información (El-Helw, Farid, y Ilyas, 2012) (Irfan et al., 2015), clasificación de textos (Iglesias, Seara Vieira, y Borrajo, 2013) (Zhang, Zhao, y LeCun, 2015), detección y minería de opiniones (Fernández et al., 2013) (Ravi y Ravi, 2015) o síntesis de información (Cadilhac et al., 2015) (Moen et al., 2016) así como los procesos intermedios involucrados en cada una de estas tareas tales como el análisis semántico (Li y Joshi, 2012) (Gutiérrez, Vázquez, y Montoyo, 2017) que son clave para su interpretación.

Por tanto, se hace necesario aunar esfuerzos en las distintas tareas hacia la creación de una plataforma capaz de identificar el tipo de información que necesita el usuario, recuperarla, procesarla y presentársela de manera adecuada y flexible.

## 2 Estado del arte

Hoy en día existen algunas herramientas y sistemas informáticos que de una manera u otra son capaces de incorporar tecnologías de TLH para proporcionar infraestructuras analíticas. Por ejemplo Atribus[8] que es capaz de rastrear, buscar, recoger, filtrar y devolver todo lo que se está diciendo de un cliente en la red a partir de las palabras clave para cada uno en tiempo real; Natural Opinions[9] que analiza todo lo que se está diciendo en cada momento en Internet sobre una persona, una marca, una institución o un producto, y detectar automáticamente las entidades, conceptos y opiniones más relevantes; Textalytics[10] el cual se presenta como un motor de análisis de texto que extrae elementos con significado de cualquier contenido y lo estructura para que puedas procesarlo y gestionarlo fácilmente; Sentimentviz[11] propone un medios estimar y visualizar el sentimiento asociado a textos cortos e incompletos, Tweet Reach[12] permite obtener informes estadísticos a partir del análisis de Twitter, SocialBro[13] que propone una solución avanzada para la gestión y el análisis de comunidades de Twitter, permitiendo a los profesionales del Marketing y el Social Media analizar a fondo sus contactos, gestionarlos y definir sus estrategias en función de ello; SumAll[14] entre otras cosas obtiene estadísticas sobre seguidores de varias las redes sociales, como son: números de *me gusta*, cantidad de mensajes, localidades, etc. y los muestra por medios de gráficas de intervalos de tiempo (días, semanas, meses).

Nuestra propuesta estaría básicamente más alineada con las soluciones de Atribus y Natural Opinions. Además seríamos capaces de aportar con la plataforma de TLH valores añadidos como los que siguen a continuación.

## 3 Características distintivas de la propuesta de proyecto

Las características que se añaden en este proyecto a diferencia de las tecnologías ya existentes son: Nube de conceptos vs nubes de palabras/etiquetas; Dominios relevantes; Mapas de emociones (clasificación de tipos de emociones vs. Simple clasificación de polaridad); Extracción de mensajes de usuarios más relevantes en un intervalo de tiempo (resumen de tweets); Mapas de polaridad donde geográficamente se puedan representar las opiniones expresadas o inferidas de los usuarios de las redes sociales; Detección de conjuntos de términos que caracterizan e indican una localidad (e.g. postiguet, hogueras, arroz, Alicante) vs. Simple geolocalización que proporciona Twitter; Tratamiento multilinge de la información; y Flexibilidad para establecer métricas de análisis de reputación de entidades digitales.

## 4 Propuesta

Nuestra propuesta de plataforma de TLH se ilustra en la Figura 1. Dicha plataforma permite a los usuarios extraer información que se encuentre dispersa en la Web Social y representarla visualmente desde un punto de vista analítico, tras un intenso procesamiento de datos estructurados y no estructurados.

[8] https://www.atribus.com (Marzo, 2018)
[9] https://www.bitext.com (Marzo, 2018)
[10] https://www.meaningcloud.com/es/ (Marzo, 2018)
[11] https://www.csc2.ncsu.edu/faculty/healey/tweet_viz (Marzo, 2018)
[12] https://tweetreach.com/ (Marzo, 2018)
[13] http://es.socialbro.com (Dec, 2017)
[14] https://sumall.com (Marzo, 2018)

Las herramientas y procesos de TLH son el elemento central de este proyecto, ya que en él se tienen en cuenta tres roles fundamentales.

El primero es relacionado con la **extracción y recuperación** de Contenidos Generados por Usuarios (UGC). La plataforma debe ser capaz de ofrecer mecanismos para que los usuarios puedan definir sus propias búsquedas de información y de este modo el sistema pueda recuperar y extraer la información precisa.

El segundo rol es concerniente a la posibilidad de considerar **diferentes tipos de tecnologías de TLH** con el fin de poder aplicar **Minería de Textos** y obtener diversos rasgos de caracterización.

Y por último y no menos importante, debe ser capaz de ofrecer mecanismos para **mostrar de modo visual y sencillo analíticas** resultantes tras procesar la información obtenida.

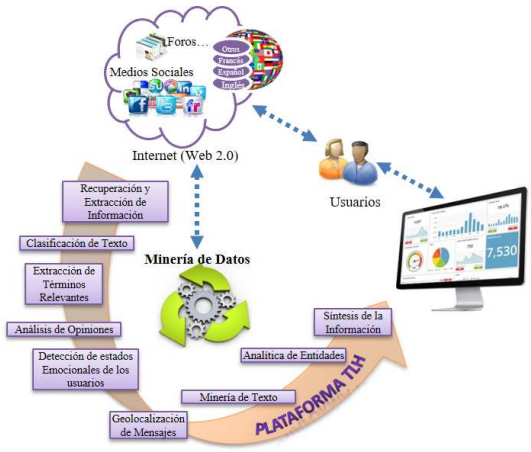


Figura 1: Plataforma de TLH

## 5 *Objetivo general*

El objetivo general de este proyecto es analizar, proponer y evaluar diferentes enfoques novedosos para el procesamiento de UGC desde un punto de vista analítico, creando una plataforma que combine, integre y visualice la información resultante de distintos procesos de TLH.

La información resultante puede materializarse de distintas formas dependiendo de las necesidades y preferencias del usuario. Por ejemplo, pueden ser sintetizadas en forma de resúmenes, de tweets, valoración de opiniones, términos más relevantes, pasajes, recopilación de fuentes relevantes, geolocalización de los mensajes, autores, etc. Siendo conscientes que dicha información procede de la Web 2.0.

El núcleo del proceso tanto de recuperación y extracción como de procesamiento de la información estaría formado por técnicas y herramientas que conforman las TLH. Para ello, se integrarán tecnologías tales como el análisis semántico, recuperación y extracción de información, minería de opiniones, clasificación de textos, computación afectiva (o análisis de emociones), síntesis de textos y otras que puedan ser de utilidad durante el transcurso del proyecto. Aunque la plataforma no está limitada a la integración de otras tecnologías, sí que es cierto que estas tareas serán las que conformen su núcleo central, y por tanto, serán cruciales para el correcto desarrollo del proyecto.

Por tanto, el proceso de clasificación, análisis y presentación del contenido social implicaría en primer lugar decidir qué información se debe recuperar y seleccionarla. Posteriormente habrá que ser capaces de procesar dicha información. Para ello, será necesari: identificar el tipo de información; clasificarla; detectar lo realmente importante y discriminar aquello que no es relevante; determinar información redundante, complementaria y/o contradictoria; e integrar y combinar todo el conocimiento obtenido. Finalmente, todo el conocimiento obtenido quedará almacenado en un repositorio de minería de textos capaz de indexar toda aquella información que el usuario considere relevante para ser mostrado desde una óptica analítica mediante interfaces visuales.

## 6 *Oportunidades de explotación*

En la actualidad muchas empresas se preocupan considerablemente por su reputación en la Web 2.0, ya que las redes se han convertido en las vías más populares, rápidas y efectivas de *marketing*. Es por ello, que nuevos perfiles laborales surgen de la mano de las nuevas tecnologías. Por ejemplo, podemos mencionar la figura del analista social[15], que entre otras funciones: evalúa y propone mejoras para la estrategia en los medios sociales y campañas comerciales; monitoriza y recolecta información sobre marca, productos, competencia y sector; clasifica las consultas de los clientes

[15] http://www.concepto05.com/2011/01/que-es-un-social-media-analyst-i-un-nuevo-puesto-de-trabajo (pub Enero, 2011)

para mejorar los sistemas de atención técnica; analiza la reputación online de marcas; realiza análisis sectoriales y comparativas con la competencia; etc.

## 7 Enlace al proyecto y resultados

En la web del proyecto[16] podéis encontrar la relación de artículos científicos que sustentan las tecnologías desarrolladas, así como registros de software y herramientas de demostración.

## Agradecimientos

Este proyecto con referencia GRE16-01: Plataforma inteligente para recuperación, análisis y representación de la información generada por usuarios en Internet, está financiado por las Ayudas a Proyectos Emergentes de la Universidad de Alicante.

## Bibliografía

Cadilhac, A., A. Chisholm, B. Hachey, y S. Kharazmi. 2015. Hugo: Entity-based News Search and Summarisation. En *Proceedings of the Eighth Workshop on Exploiting Semantic Annotations in Information Retrieval - ESAIR '15*, páginas 51–54.

El-Helw, A., M. H. Farid, y I. F. Ilyas. 2012. Just-in-time information extraction using extraction views. En *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, páginas 613–616, New York, NY, USA. ACM.

Fernández, J., Y. Gutiérrez, J. M. Gómez, P. Martínez-Barco, A. Montoyo, y R. Muñoz. 2013. Sentiment Analysis of Spanish Tweets Using a Ranking Algorithm and Skipgrams. *XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN 2013)*, páginas 133–142.

Gutiérrez, Y., S. Vázquez, y A. Montoyo. 2017. Spreading semantic information by Word Sense Disambiguation. *Knowledge-Based Systems*, 132:47–61.

Iglesias, E. L., A. Seara Vieira, y L. Borrajo. 2013. An HMM-based oversampling technique to improve text classification. *Expert Systems with Applications*, 40(18):7184–7192.

Irfan, R., C. K. King, D. Grages, S. Ewen, S. U. Khan, S. A. Madani, J. Kolodziej, L. Z. Wang, D. Chen, A. Rayes, N. Tziritas, C. Z. Xu, A. Y. Zomaya, A. S. Alzahrani, y H. X. Li. 2015. A survey on text mining in social networks. *Knowledge Engineering Review*, 30(2):157–170.

Li, Y. y K. D. Joshi. 2012. The state of social computing research: A literature review and synthesis using the latent semantic analysis approach. En *18th Americas Conference on Information Systems 2012, AMCIS 2012*, volumen 1, páginas 33–40.

Moen, H., L. M. Peltonen, J. Heimonen, A. Airola, T. Pahikkala, T. Salakoski, y S. Salanterä. 2016. Comparison of automatic summarisation methods for clinical free text notes. *Artificial Intelligence in Medicine*, 67:25–37.

Ravi, K. y V. Ravi. 2015. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46.

Zhang, X., J. Zhao, y Y. LeCun. 2015. Character-level convolutional networks for text classification. En *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, páginas 649–657, Cambridge, MA, USA. MIT Press.

[16] https://gplsi.dlsi.ua.es/gplsi13/es/node/396

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 131-134 recibido 15-03-2018 revisado 28-04-2018 aceptado 03-05-2018

# Extracción automática de equivalentes multilingües de colocaciones

## *Automatic extraction of multilingual collocation equivalents*

**Marcos Garcia**
Universidade da Coruña
Grupo LyS, Departamento de Letras, Facultade de Filoloxía
Campus da Zapateira, 15008, A Coruña, Galiza
marcos.garcia.gonzalez@udc.gal

**Resumen:** Este trabajo presenta el proyecto *Extracción automática de equivalentes multilingües de colocaciones*, financiado por el programa *Becas Leonardo a Investigadores y Creadores Culturales 2017* dentro del área de Humanidades. El objetivo principal del proyecto es extraer automáticamente equivalentes multilingües de colocaciones fraseológicas. Para ello proponemos diversas estrategias que combinan análisis sintáctico y estadístico con técnicas de semántica distribucional, tanto monolingües como multilingües. Las lenguas de trabajo del proyecto son el portugués, el español y el inglés.
**Palabras clave:** colocaciones, fraseología, léxico, sintaxis, semántica distribucional

**Abstract:** This work presents the project *Automatic extraction of multilingual collocation equivalents*, funded by a *2017 Leonardo Grant for Researchers and Cultural Creators*, in the area of Humanities. The main objective of this project is to automatically extract multilingual equivalents of phraseological collocations. We propose several strategies which combine syntactic and statistical analysis with distributional semantic techniques, both monolingual and multilingual. The languages of the project are Portuguese, Spanish, and English.
**Keywords:** collocations, phraseology, lexicon, syntax, distributional semantics

### *1 Introducción y objetivos*

Un alto porcentaje de las expresiones lingüísticas que producimos los hablantes de una lengua está formado por estructuras más o menos preconstruidas de elementos léxicos (Erman y Warren, 2000). Entre estas estructuras encontramos diferentes tipos, tales como las locuciones (expresiones no composicionales desde un punto de vista semántico), o las colocaciones (entendidas aquí como combinaciones composicionales restringidas léxicamente) (Mel'čuk, 1995; Alonso-Ramos, 1995).

A pesar de que estas expresiones no presentan grandes problemas en el proceso de adquisición de las lenguas maternas, sí que generan dificultades en situaciones en las que coexiste más de un idioma, como en el aprendizaje de lenguas extranjeras (Altenberg y Granger, 2001), o en sistemas de traducción automática (Orliac y Dillinger, 2003).

Así, la colocación del inglés *brown sugar*, podría tener como equivalentes *azúcar moreno* –en español–, o *açúcar mascavado* –en portugués–, pero no otras alternativas como **azúcar marrón*, **açúcar castanho*, etc.

Conocer qué combinaciones son posibles y cuáles no son habituales en un dado idioma puede ser útil para diversos propósitos, desde la enseñanza de lenguas o la traducción automática, a la compresión y generación de lenguaje natural.

Teniendo esto en cuenta, el objetivo principal de este proyecto consiste en implementar métodos basados en lingüística computacional que permitan extraer automáticamente, con alta precisión y a gran escala equivalentes de colocaciones en portugués, español e inglés.

Especial interés tendrá la obtención de equivalentes *incongruentes* de colocaciones, i.e., aquellos en los que la traducción de sus constituyentes no es coherente (Nesselhauf, 2003). Por ejemplo, el par inglés-español "pay attention"–"prestar atención", a diferencia de equivalentes *congruentes* como "formulate

ISSN 1135-5948. DOI 10.26342/2018-61-15

[a] hypothesis"–"formular [una] hipótesis".

En el proceso de obtención de equivalentes multilingües de colocaciones podemos identificar dos tareas bien diferenciadas:

1. Extracción de colocaciones monolingües.
2. Identificación de equivalentes multilingües.

Para la primera utilizaremos análisis sintáctico de dependencias y medidas de asociación estadística. Además, compararemos métodos distribucionales composicionales y no composicionales con el objetivo de discriminar aquellas colocaciones fraseológicas de las puramente estadísticas.

Para identificar equivalentes de colocaciones en diferentes idiomas usaremos modelos multilingües de semántica distribucional, que serán obtenidos tanto de corpus paralelos y comparables como de textos monolingües.

Este proyecto continúa y amplía un conjunto de trabajos previos sobre la extracción automática de colocaciones multilingües (Garcia, García-Salido, y Alonso-Ramos, 2018; Garcia, García-Salido, y Alonso-Ramos, 2017) y, del mismo modo, publicará y distribuirá utilizando licencias libre todos los recursos generados durante su realización.

## 2 Metodología y plan de trabajo

Para llevar a cabo nuestro objetivo, el proyecto se divide en cuatro partes, cada una de ellas organizada en diferentes subtareas:

1. Compilación, análisis y clasificación de corpus paralelos, comparables y monolingües.
2. Identificación de candidatos a colocaciones mediante técnicas lingüístico-estadísticas.
3. Creación y evaluación de modelos multilingües de semántica distribucional.
4. Clasificación automática de equivalentes multilingües de colocaciones.

La primera etapa (compilación y análisis de corpus) consistirá en obtener a través de la web grandes cantidades de corpus, tanto multilingües (paralelos y comparables) como monolingües. Los diferentes recursos compilados pertenecerán a diferentes registros y dominios (enciclopédico, periodístico, oral, etc.), con el fin de conseguir un conjunto amplio y diverso de información fraseológica y distribucional. Los corpus serán clasificados, utilizando técnicas semiautomáticas, en función del registro y de la variedad lingüística.

En la segunda fase aplicaremos métodos híbridos, que combinan información sintáctica y fraseológica con modelos estadísticos, para la identificación de candidatos a colocaciones monolingües (Seretan, 2011; Evert et al., 2017). Aquí será necesario realizar un análisis automático de los corpus utilizando sistemas de procesamiento del lenguaje natural (Garcia y Gamallo, 2015; Straka y Straková, 2017). El análisis sintáctico será realizado con *Universal Dependencies*, lo que nos permitirá, entre otras cosas, obtener corpus etiquetados con una anotación homogénea en las diversas lenguas de trabajo (Nivre, 2015). Además, evaluaremos diferentes estrategias composicionales y no composicionales para diferenciar colocaciones fraseológicas de combinaciones estadísticas (Pearce, 2001; Kiela y Clark, 2013; Rodríguez-Fernández et al., 2016; Farahmand y Henderson, 2016).

El resultado de esta segunda fase serán grandes listas monolingües de colocaciones candidatas, ordenadas tanto mediante información estadística (*log-likelihood*, *DeltaP*, etc.), como por su carácter fraseológico.

Denominamos colocaciones fraseológicas a aquellas combinaciones de dos unidades léxicas (*base* y *colocativo*) en las que el significado de la base se mantiene intacto en la colocación, mientras que el del colocativo depende del significado de la propia colocación (e.g., $odio_{Base}$ $mortal_{Colocativo}$) (Mel'čuk, 1995). A diferencia de las visiones puramente estadísticas de las colocaciones, la tradición fraseológica exige una relación sintáctica directa entre los dos elementos de una colocación. Además, la elección del colocativo no es libre, sino que está restringida por la base (Alonso-Ramos, 1995).

En este proyecto nos centraremos en tres tipos diferentes de colocaciones:

- Verbo-Objeto: $ceño_b$, $fruncir_c$ ("fruncía el ceño"); $moción_b$, $secundar_c$ ("secundaron la moción").
- Nombre-Adjetivo: $dinero_b$, $negro_c$; $saludo_b$, $cordial_c$.
- Nombre-Nombre: $coral_b$, $arrecife_c$ ("arrecifes de coral"); $lana_b$, $ovillo_c$ ("ovillo de lana").

Téngase en cuenta que el uso de dependencias sintácticas permite identificar colocaciones a larga distancia ("fruncía con dolor pero rápidamente el ceño") y en diferente orden ("saludos$_{\text{nombre}}$ cordiales$_{\text{adjetivo}}$", "mayor$_{\text{adjetivo}}$ cantidad$_{\text{nombre}}$"). Además, el análisis en dependencias universales establece relaciones sintácticas entre palabras léxicas (e.g., *cartón*→`nmod`→*tabaco* en "cartón de tabaco"), lo que agiliza el proceso de extracción de candidatos a colocaciones.

Tanto la clasificación de los corpus multilingües realizada en la primera etapa, como la información obtenida mediante el análisis computacional, serán utilizadas en la tercera parte del proyecto, cuyo objetivo es la creación de modelos de semántica distribucional multilingües.

Así, en esta tercera fase crearemos diferentes modelos bilingües (portugués-inglés, portugués-español y español-inglés) utilizando corpus paralelos, comparables y monolingües. Además, cada par de idiomas contará en esta etapa con modelos construidos utilizando tres estrategias: (i) monoléxica, en los que cada vector representa una única palabra (ortográfica), (ii) contextual, en los que los diferentes sentidos de cada palabra serán agrupados en función de su distribución semántica, y (iii) no composicional, que representa en un único vector las dos unidades léxicas de cada colocación candidata.

Durante esta fase evaluaremos la precisión de los diferentes modelos distribucionales en la identificación de equivalentes multilingües de dos tipos de colocaciones: (i) *congruentes*, en las que la traducción de las unidades léxicas es coherente interlingüísticamente (por ejemplo, entre inglés y español: *vermouth$_b$, red$_c$ – vermú$_b$, rojo$_c$*), e (ii) *incongruentes*, en las que el carácter impredecible de las colocaciones implica que los equivalentes multilingües no sean traducciones directas de la base y el colocativo (*wine$_b$, red$_c$ – vino$_b$, tinto$_c$*).

En la última parte del proyecto utilizaremos los candidatos a colocaciones monolingües (obtenidos en la fase 2) y los modelos distribucionales (fase 3) para generar los equivalentes multilingües de colocaciones. Primero, unificaremos y seleccionaremos las colocaciones candidatas obtenidas por las diferentes medidas de asociación estadística y de análisis fraseológico, generando así listas de colocaciones de alta confianza para cada uno de los tipos (verbo-objeto, adjetivo-nombre y nombre-nombre). Después, utilizando estas listas monolingües de colocaciones, aplicaremos los modelos distribucionales para identificar (i) equivalentes de la base y del colocativo mediante similaridad semántica (modelos monoléxicos, aplicados principalmente a equivalentes congruentes), y (ii) equivalentes de la colocación como combinación contextual o como unidad (modelos contextuales y no composicionales, para la identificación de equivalentes incongruentes).

Por último, los métodos de mayor precisión y cobertura se aplicarán en los tres pares de lenguas analizadas (portugués-inglés, portugués-español y español-inglés), extrayendo así grandes cantidades de equivalentes bilingües de colocaciones. Una vez obtenidas estas listas bilingües, se aplicará un método de fusión por transitividad para generar un recurso multilingüe portugués-inglés-español. Este proceso permitirá también aumentar el número de equivalentes bilingües, ya que podrán descubrirse, por ejemplo, pares portugués-inglés no reconocidos previamente, a través de equivalentes español-portugués y español-inglés correctamente identificados.

## *3 Equipo de trabajo*

El proyecto se lleva a cabo en el Grupo LyS (Lengua y Sociedad de la Información), de la Universidade da Coruña. El Grupo LyS es un equipo interdisciplinar de investigación en Lingüística Computacional del que forman parte diferentes profesores, investigadores y estudiantes tanto del área de Lingüística General como de Ciencias de la Computación.

Además del investigador principal del proyecto, en diferentes etapas del mismo serán contratadas dos personas que colaboren tanto en las tareas de compilación, anotación y análisis de los corpus, como en el diseño, implementación y evaluación de las varias estrategias que analizaremos.

Por otro lado, el hecho de formar parte de un grupo interdisciplinar nos permite trabajar en estrecha colaboración con otros miembros del equipo que, sin ser parte directa del proyecto, aportan conocimientos fundamentales para su desarrollo.

## *Agradecimientos*

Proyecto realizado con una Beca Leonardo a Investigadores y Creadores Culturales 2017

(Fundación BBVA), y parcialmente financiado por un contrato *Juan de la Cierva - incorporación* (IJCI-2016-29598, Ministerio de Economía y Competitividad) y por *RELEX: Rede de Lexicografía* (ED341D R2016/046).

## *Bibliografía*

Alonso-Ramos, M. 1995. Hacia una definición del concepto de colocación: de J.R. Firth a I.A. Mel'čuk. *Revista de Lexicografía*, 1:9–28.

Altenberg, B. y S. Granger. 2001. The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics*, 22(2):173–195.

Erman, B. y B. Warren. 2000. The idiom principle and the open choice principle. *Text-Interdisciplinary Journal for the Study of Discourse*, 20(1):29–62.

Evert, S., P. Uhrig, S. Bartsch, y T. Proisl. 2017. E-VIEW-alation–a Large-scale Evaluation Study of Association Measures for Collocation Identification. En *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*, páginas 531–549, Leiden.

Farahmand, M. y J. Henderson. 2016. Modeling the non-substitutability of multiword expressions with distributional semantics and a log-linear model. En *Proceedings of the 12th Workshop on Multiword Expressions (MWE 2016) at ACL 2016*, páginas 61–66, Berlin. ACL.

Garcia, M. y P. Gamallo. 2015. Yet another suite of multilingual NLP tools. En José-Luis Sierra-Rodríguez and José Paulo Leal and Alberto Simões, editor, *Languages, Applications and Technologies. Communications in Computer and Information Science*, International Symposium on Languages, Applications and Technologies (SLATE 2015), páginas 65–75. Springer.

Garcia, M., M. García-Salido, y M. Alonso-Ramos. 2017. Using bilingual word-embeddings for multilingual collocation extraction. En *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017) at EACL 2017*, páginas 21–30, Valencia. ACL.

Garcia, M., M. García-Salido, y M. Alonso-Ramos. 2018. Discovering bilingual collocations in parallel corpora: A first attempt at using distributional semantics. En I. Doval y M. T. Sánchez Nieto, editores, *Parallel corpora for contrastive and translation studies: New resources and applications*. John Benjamins Publishing.

Kiela, D. y S. Clark. 2013. Detecting compositionality of multi-word expressions using nearest neighbours in vector space models. En *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, páginas 1427–1432, Copenhagen.

Mel'čuk, I. 1995. Phrasemes in language and phraseology in linguistics. En *Idioms: Structural and psychological perspectives*. Lawrence Erlbaum Associates, páginas 167–232.

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

Nivre, J. 2015. Towards a Universal Grammar for Natural Language Processing. En *International Conference on Intelligent Text Processing and Computational Linguistics*, páginas 3–16. Springer.

Orliac, B. y M. Dillinger. 2003. Collocation extraction for machine translation. En *Proceedings of Ninth Machine Translation Summit (MT Summit IX)*, páginas 292–298, New Orleans, Lousiana.

Pearce, D. 2001. Synonymy in collocation extraction. En *Proceedings of the Workshop on WordNet and other lexical resources at NAACL 2001*, páginas 41–46, Pittsburgh. ACL.

Rodríguez-Fernández, S., L. Espinosa Anke, R. Carlini, y L. Wanner. 2016. Semantics-driven recognition of collocations using word embeddings. En *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, páginas 499–505, Berlin.

Seretan, V. 2011. *Syntax-based collocation extraction*, volumen 44 de *Text, Speech and Language Technology Series*. Springer Science & Business Media.

Straka, M. y J. Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. En *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, páginas 88–99, Vancouver.

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 135-138 recibido 16-03-2018 revisado 29-04-2018 aceptado 03-05-2018

# ARAP: Arabic Author Profiling Project for Cyber-Security

## *ARAP: Proyecto sobre Perfiles de Autoría en Árabe para la Ciber-Seguridad*

**Paolo Rosso**[1], **Francisco Rangel**[1], **Bilal Ghanem**[1], **Anis Charfi**[2]
[1]PRHLT Rsearch Center, Universitat Politècnica de València
Camino de Vera s/n, 46022 Valencia, Spain
prosso@dsic.upv.es, francisco.rangel@autoritas.es, bigha@doctor.upv.es
[2]Cargenie Mellon University Qatar, Education City PO Box 24866 Doha, Qatar
acharfi@andrew.cmu.edu

**Abstract:** In this paper we describe the current state of the ARAP project on Arabic Author Profiling for Cyber-Security funded by the Qatar National Research Fund via the Carnegie Mellon University in Qatar. The project focuses on determining whether a suspicious message is actually a potential threat and, in that case, aims at profiling its author. The contribution of the project lies in the lack of research of this type for Arabic.
**Keywords:** Arabic, cyber-security, author profiling, gender, age, native language, language variety, deception, irony

**Resumen:** En este artículo describimos el estado actual del proyecto ARAP de perfilado de autores en árabe para la ciberseguridad financiado por la Qatar National Research Fund va la Universidad de Carnegie Mellon en Qatar. El proyecto se centra en determinar cuando un mensaje sospechoso realmente es una amenaza potencial, y en tal caso, perfilar a su autor. La contribución del proyecto reside en la falta de investigaciones de este tipo para el árabe.
**Palabras clave:** Árabe, ciberseguridad, perfiles de autor, sexo, edad, idioma nativo, variedad del lenguaje, engaño, ironía

### *1 Cyber-Security in Social Media*

The anonymity of social media provides with new ways of communication without censorship. However, the lack of knowledge about the authors may contribute to new cyber-security issues, such as threatening messages or terrorism propaganda. Security in the fifth domain, the cyber space, is nowadays one of the defense priorities in many nations[1]. Generating intelligence from social networks content is important to prevent cyber threats. To profile potential terrorists from the messages that they share in their social circles allows detecting communities whose aim is to undermine the security in our daily life. From 2014 to 2017, the PRHLT research center of the Universitat Politècnica de València has been involved in a research project funded by the Army Research Office of the United States whose objective was the detection of communities in Twitter that shared content about ISIS (large-scale copy detection in social circles)[2]. Since 2017 the PRHLT research center takes part in a project of the Qatar National Research Fund whose aim is determining the linguistic profile of the author of a suspicious or threatening text, with the attempt of profiling potential terrorists (Russell and Miller, 1977). When a suspicious message is analysed, we follow the workflow in Figure 1. Concretely: *(i)* we check the veracity of the threat, discarding those messages that are deceptive (Cagnina and Rosso, 2017) or ironic (Hernández, Patti, and Rosso, 2016), since they do not represent a real threat; finally, *(ii)* we profile the demographics of its author (Rangel and Rosso, 2016) as well as

[1]https://www.economist.com/node/16478792

[2]https://www.prhlt.upv.es/wp/es/project/2016/sococode

ISSN 1135-5948. DOI 10.26342/2018-61-16

her cultural and social context.

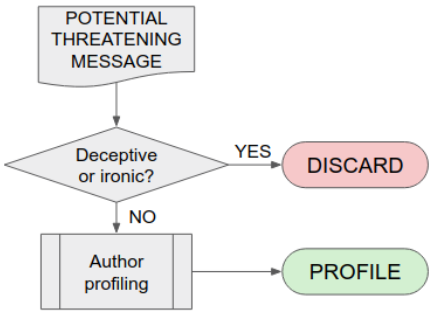


Figure 1: Workflow to profile the author of a potential threatening message.

The contribution of the presented project is relevant due to the lack of this kind of investigations in the Arabic language (Rosso et al., 2018). In the next sections, we describe the current state of the project.

## 2 Threat Messages: Beyond Deceptive Messages and Irony

A suspicious message may not be a threat when it is deceptive or ironic. A message can be considered deceptive when it is written with the intention to sound authentic. Deceptive detection has mainly focused on the detection of spam in opinion reviews (Cagnina and Rosso, 2017) and nowadays there is an increasingly interest in fake news detection (e.g. deception in political debates[3]), also in Arabic. Despite the interest in the area, few are the preliminary works in Arabic in deception detection in reviews and news.

The most common definition of irony is the use of words to express the opposite meaning from what is said. With respect to irony detection in Arabic, even less are the preliminary works. Recently, a preliminary system for irony detection in Arabic in social media was presented in (Karoui, Zitoune, and Moriceau, 2017). At the moment we are helping in the annotation of an enriched version of the corpus the authors employed in the previous work.

## 3 Author Profiling: Gender, Age, Native Language, and Arabic Language Variety

Since 2013 author profiling in social media has been addressed in the framework of the PAN Lab[4] at CLEF[5]. We started addressing the Arabic language since 2017 (Rangel et al., 2017): gender and language variety identification.

In the framework of the ARAP project, we approached both with the Low Dimensionality Statistical Embedding (LDSE) (Rangel, Rosso, and Franco-Salvador, 2018) representation. This method represents documents on the basis of the probability distribution of occurrence of their words in the different classes. The key concept of LDSE is a weight, representing the probability of a term to belong to one of the different categories: for gender (female vs. male) and for language variety (e.g. Egypt vs. Levantine vs. ...). The distribution of weights for a given document should be closer to the weights of its corresponding category. In order to compare the obtained results, we have proposed the following two state-of-the-art baselines: *i)* BASELINE-stat that emulates random choice, depending on the number of classes. For example, in gender identification with two balanced classes, this baseline obtain 50% of accuracy; and *ii)* BASELINE-bow that represents documents as a bag-of-words with the 1,000 most common words in the training set, weighted by absolute frequency of occurrence. The texts are preprocessed as follows: lowercase words, removal of punctuation signs and numbers, and removal of stop words. A Support Vector Machine classifier with a linear kernel and default parameters is used.

In the next subsections, we describe the corpora used and the obtained results with LDSE compared to the described baselines. In case of using the PAN corpus, we have also compared ourselves with the best result obtained in the official task.

### 3.1 Gender Identification

We have collected two corpora and annotated with gender information: *i)* PAN-AP-2017[6]; and *ii)* CMUQ-ARAP. To build the PAN-AP-2017 corpus, we have retrieved tweets geolocated in the following cities: Cairo, Abu Dhabi, Doha, Kuwait, Manama, Mascate, Riyadh, Sana'a, Amman, Beirut, Damascus,

[3] http://alt.qcri.org/clef2018-factcheck

[4] http://pan.webis.de

[5] http://clef2018.clef-initiative.eu

[6] The PAN-AP-2017 corpus has been also annotated with language variety information. Hence, the methodology described here applies to the next section.

Jerusalem, Algiers, Rabat, Tripoli, and Tunis. We have selected the unique authors who wrote these tweets and downloaded their timelines. Authors with their location outside the given regions have been discarded, as well as retweets or tweets written in other languages than Arabic. We have annotated gender automatically with a dictionary of proper nouns and performed a manual review of their profiles to fix errors and discard ambiguous profiles. The corpus consists of a total of 4,000 authors completely balanced by gender, with 100 tweets per author, and split into training and test following 60/40 proportion. The Carnegie Mellon University in Qatar has developed the CMUQ-ARAP corpus. This corpus consists of a total of 10,140 authors with a variable number of tweets (from a few ones to thousands), imbalanced with respect to gender, and split into training and test following 60/40 proportion. Corpora statistics are shown in Table 1.

| Set | Males | Females | Total |
|---|---|---|---|
| **PAN-AP-2017** | | | |
| Training | 1,200 | 1,200 | 2,400 |
| Test | 800 | 800 | 1,600 |
| Total | 2,000 | 2,000 | 4,000 |
| **CMUQ-ARAP** | | | |
| Training | 5,481 | 3,645 | 9,126 |
| Test | 609 | 405 | 1,014 |
| Total | 6,090 | 4,050 | 10,140 |

Table 1: Arabic corpora annotated with gender information.

The following tweets are examples written by a female and a male respectively:

هههه كنت مشغولة يومها عندي بزنس مش فاضية

انا مش متخيل حيجي يوم نسوق في كوبري

In the female sentence (*hahaha I was busy on that day, I have a business so I'm not free*), the words مشغولة (*busy*) and فاضية (*free*) indicate that the writer is female. In the Arabic language when the adjective ends with ـة (*Taa': one of the Arabic letter*) (this is the letter shape in an independent situation: ة ) letter, it gives an indicator that the speaker is a female. While in the male sentence (*I don't imagine that in the future we will drive in a subway*), the word متخيل (*conceived*) without the same previous letter (ة) implies that the speaker is a man.

We have used the LDSE representation with SVM with Gaussian kernel and gamma equal to 0.02 to approach the task on the PAN-AP-2017 corpus, whereas SVM with linear kernel for the CMUQ-ARAP corpus[7]. The obtained results in terms of accuracy can be seen in Table 2.

| Corpus | Method | Accuracy |
|---|---|---|
| PAN-AP-2017 | LDSE | 73.19 |
| | BASELINE-bow | 53.00 |
| | BASELINE-stat | 50.00 |
| | Best at PAN'17 | 80.31 |
| CMUQ-ARAP | LDSE | 66.96 |
| | BASELINE-bow | 61.56 |
| | BASLINE-stat | 60.06 |

Table 2: Gender results in terms of accuracy.

In case of PAN-AP-2017 corpus, the proposed method obtains about 20% higher accuracy than the best baseline (38% of improvement), although it obtains lower results than the best participant at PAN (7.12%). In case of CMUQ-ARAP, the proposed method obtains about 5.4% higher accuracy than the best baseline (8.77% of improvement). Despite the imbalance in the number of authors per gender besides the uneven number of tweets per author, the results are not much lower than the obtained with the PAN-AP-2017 corpus[8].

### 3.2 Language Variety Identification

Following previous works (Sadat, Kazemi, and Farzindar, 2014), the aforementioned PAN-AP-2017 corpus has been annotated with four varieties of Arabic: Egypt, Gulf, Levantine and Maghrebi. There are 1,000 authors per variety, divided into 600/400 for training and test respectively. Each author contains 100 tweets.

The following tweet is an example of the Gulf language variety:

طاريك بمليني فرح ، شعاد لقياك

(*Your remembrance makes me rejoice, but what about a meeting with you!*). The words طاريك (*your remembrance*), شعاد (*but what*

[7]Several algorithms and parameters have been tested and we have selected the configuration that obtained the best results.

[8]The CMUQ team is working on balancing the corpus in terms of number of authors per gender and number of tweets per author.

*about*) and لقياك (*meeting you*) are only used in the Gulf variety.

Table 3 shows the results for the language variety task on the PAN-AP-2017 corpus. We have used LDSE and SVM with Gaussian kernel and default parameters. As can be seen, LDSE outperforms both baselines in 48.56% and 57.50% of accuracy (143% and 230% of improvement respectively). The difference with respect to the best result at PAN (0.63%) is not statistically significant.

| Method | Accuracy |
|---|---|
| LDSE | 82.50 |
| BASELINE-bow | 33.94 |
| BASELINE-stat | 25.00 |
| Best at PAN'17 | 83.13 |

Table 3: Gender results in terms of accuracy.

## 4 Conclusions and Future Work

In this paper we have presented the current state of the ARAP project on Arabic Author Profiling for Cyber-Security. So far, we have focused mainly on gender and language variety identification since 2017. We have included the Arabic language in the organisation of the PAN shared task and built corpora labeled with gender and language variety information. We have proposed a method to approach both problems and obtained competitive results. At the moment of writing of this paper, the colleagues at the Carnegie Mellon University in Qatar are tagging the CMUQ-ARAP corpus with the information about age. Moreover, the colleague in Qatar are organising the Fact Checking Lab[9] at CLEF on deceptive detection in political debates in English and Arabic.

As future work, at PAN@CLEF in 2018 we will address gender identification in Twitter from a multimodal perspective taking into account not only the textual information but also the images of the URLs links in tweets. In the future, together with the Carnegie Mellon University in Qatar, we plan to organise a track at the Forum of the Information Retrieval Evaluation addressing the several aspects of the ARAP research project.

## Agradecimientos

This article was made possible by NPRP grant 9-175-1-033 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

[9] http://alt.qcri.org/clef2018-factcheck

## References

Cagnina, L. C. and P. Rosso. 2017. Detecting deceptive opinions: Intra and cross-domain classification using an efficient representation. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 25(Suppl. 2):151–174.

Hernández, I., V. Patti, and P. Rosso. 2016. Irony detection in twitter: The role of affective content. *ACM Transactions on Internet Technology (TOIT)*, 16(3):19.

Karoui, J., F. B. Zitoune, and V. Moriceau. 2017. Soukhria: Towards an irony detection system for arabic in social media. *Procedia Computer Science*, 117:161–168.

Rangel, F. and P. Rosso. 2016. On the impact of emotions on author profiling. *Information processing & management*, 52(1):73–92.

Rangel, F., P. Rosso, and M. Franco-Salvador. 2018. A low dimensionality representation for language variety identification. In *CICLing-2016, Springer-Verlag, Revised Selected Papers, Part II, LNCS(9624)*. Springer-Verlag, arXiv:1705.10754.

Rangel, F., P. Rosso, M. Potthast, and B. Stein. 2017. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. In *CLEF 2017 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1866.*

Rosso, P., F. Rangel, I. Hernández, L. Cagnina, W. Zaghouani, and A. Charfi. 2018. A survey on author profiling, deception, and irony detection for the arabic language. *Language and Linguistics Compass (in press).*

Russell, C. A. and B. H. Miller. 1977. Profile of a terrorist. *Studies in Conflict & Terrorism*, 1(1):17–34.

Sadat, F., F. Kazemi, and A. Farzindar. 2014. Automatic identification of arabic language varieties and dialects in social media. *Proceedings of SocialNLP*, page 22.

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 139-142 recibido 18-03-2018 revisado 28-04-2018 aceptado 03-05-2018

# MOMENT: Metáforas del trastorno mental grave. Análisis del discurso de personas afectadas y profesionales de la salud mental

## *MOMENT: Metaphors of severe mental disorder. Discourse analysis of affected people and mental health professionals*

**Marta Coll-Florit**[1], **Salvador Climent**[1], **Martín Correa-Urquiza**[2], **Eulàlia Hernández**[1], **Antoni Oliver**[1], **Asun Pié**[1]
[1]Universitat Oberta de Catalunya, Av.Tibidabo, 47, 08035 Barcelona
[2]Universitat Rovira i Virgili, Av.Catalunya, 35, 43002 Tarragona
{mcollfl,scliment,ehernandez,aoliverg,apie}@uoc.edu, martin.correaurquizav@urv.cat

**Resumen:** El proyecto MOMENT pretende contribuir a mejorar la comprensión del trastorno mental grave a partir del análisis del discurso de los dos grandes colectivos implicados, personas diagnosticadas y profesionales de la salud mental, aplicando como método de análisis la Teoría de la Metáfora Conceptual y la lingüística de corpus. En este marco, se constituirá un corpus anotado manualmente de testimonios en primera persona de ambos colectivos, que a su vez pueda servir como banco de experimentación en detección automática de metáforas. De esta manera, se pretende avanzar tanto en la detección y anotación manual de metáforas, como en la detección computacional. Desde un punto de vista social, el objetivo del proyecto es identificar y sistematizar las concepciones y asunciones dominantes sobre el trastorno mental grave, así como promover el cambio de posibles discursos que nieguen la capacidad agentiva de las personas diagnosticadas.
**Palabras clave:** Salud mental, Lingüística de corpus, Teoría de la Metáfora Conceptual, Análisis del discurso

**Abstract:** The MOMENT project aims to contribute to a better understanding of severe mental disorders by analyzing the discourse of the two main groups involved, affected people and mental health professionals, in the light of the Conceptual Metaphor Theory and Corpus Linguistics methodology. In this framework, a corpus of first-person accounts from both groups will be designed, built and manually annotated. In turn, the corpus will serve as an experimental bank for automatic detection of metaphors. Therefore, MOMENT aims to improve both manual and automatic metaphor detection and annotation. From a social point of view, the goal of the project is twofold: on the one hand, to detect and systematize dominant conceptualizations and ideas about the disorder; and, on the other hand, to identify and promote the change of potential discourses which deny the agentive capacity of affected people.
**Keywords:** Mental Health, Corpus Linguistics, Conceptual Metaphor Theory, Discourse Analysis

### *1 Participantes en el proyecto*

MOMENT es un proyecto de carácter claramente interdisciplinar, con un equipo formado por lingüistas, lingüistas computacionales y expertos en salud mental de diversos ámbitos, puesto que pretende movilizar el conocimiento complementario de diversos campos científicos para afrontar uno de los grandes retos de nuestra sociedad: mejorar la comprensión de los trastornos mentales graves.

El equipo está formado por 6 doctores, tres de los cuales son lingüistas y tres especialistas en salud mental. Los lingüistas son miembros del grupo GRIAL-UOC (Universitat Oberta de Catalunya) adscrito a GRIAL (Grupo interuniversitario de investigación en aplicaciones lingüísticas, UOC, UAB, UB, UdL) y están especializados en lingüística cognitiva, procesamiento del lenguaje natural (PLN) y lingüística de corpus. Los especialistas en salud mental pertenecen a los grupos PSiNET (Psychology, Health and ICT, UOC), CareNet (Care and Preparedness in the Network Society, UOC) y GAS (Grupo en

ISSN 1135-5948. DOI 10.26342/2018-61-17

Antropología Social, URV) y están especializados en psicología, educación social y antropología médica.

Como entidades promotoras observadoras (EPO) figuran el eHealth Center (UOC), centro de investigación en el uso de tecnologías para un cambio de paradigma en los estudios de salud; y la Asociación Radio Nikosia, entidad para la participación social de personas diagnosticadas de problemas de salud mental.

## 2 Institución financiadora y duración

El proyecto está financiado por la Agencia Estatal de Investigación (AEI) y el Fondo Europeo de Desarrollo Regional (FEDER), en el marco de la Convocatoria 2017 Retos Investigación, del Programa de I+D+i Orientada a los Retos de la Sociedad, Plan Estatal de I+D+i 2013-2016 (Referencia: FFI2017-86969-R). MOMENT se inició el 01/01/2018 y tiene una duración de tres años.

## 3 Motivación y antecedentes

El proyecto MOMENT pretende contribuir a mejorar la comprensión del trastorno mental grave (TMG) a partir del análisis del discurso de los dos grandes colectivos implicados, personas diagnosticadas y profesionales de la salud mental, aplicando como método de análisis la Teoría de la Metáfora Conceptual de la lingüística cognitiva (Lakoff y Johnson, 1980; Lakoff, 1993). La detección y sistematización de las concepciones dominantes sobre el trastorno contribuirá a identificar y promover el cambio de posibles discursos que nieguen la capacidad agentiva de las personas afectadas.

El análisis de discursos mediante la detección de la metáfora conceptual (MC) se ha aplicado a múltiples ámbitos, como el económico o el religioso (Soriano, 2012: 117). Probablemente el más conocido sea el análisis del discurso político, especialmente a partir del trabajo de Lakoff (2004) sobre el uso de MCs en el discurso de demócratas y republicanos estadounidenses. Ya de modo directamente relacionada con este proyecto, se ha analizado el uso de MCs en relatos de diversas enfermedades, tales como cáncer (Gibbs Jr y Franks, 2002; Semino et al., 2017) o ictus (Boylstein, Rittman, y Hinojosa, 2007). Sin embargo, no se ha aplicado de manera exhaustiva a los trastornos mentales graves.

La detección automática de metáforas es una área de investigación relativamente actual y muy activa, sobre la que se celebran numerosos workshops y competiciones. Durante este proyecto pretendemos adentrarnos en esta disciplina y utilizar los corpus anotados manualmente para evaluar técnicas de detección existentes y desarrollar nuevas técnicas.

## 4 Objetivos

El proyecto MOMENT se articula en los siguientes objetivos específicos:

1. Analizar testimonios de personas con un diagnóstico de TMG y profesionales de la salud mental, publicados en Internet (blogs, foros, etc.), con el fin de detectar: (a) qué tipo de metáforas usan estos colectivos de forma espontánea cuando hablan sobre el trastorno y las vivencias relacionadas; y (b) qué tipo de marcos o discursos de interpretación del trastorno se construyen a partir de las metáforas usadas.
2. Constituir un corpus anotado manualmente de producciones en primera persona de afectados y profesionales que, a su vez, sirva como banco de experimentación en detección automática de metáforas; de esta manera, se pretende potenciar la investigación sobre metáfora conceptual en el PLN.
3. Afinar el método de detección y anotación en corpus de metáforas conceptuales y otras figuras semánticas relacionadas, con el objetivo de contribuir a solucionar problemas metodológicos de identificación y formulación en el ámbito de aplicación de la Teoría de la Metáfora Conceptual.
4. Contribuir a la reflexión sobre cómo se debería hablar sobre el trastorno mental. En este sentido, en función de los resultados obtenidos, se elaborará una guía de recomendaciones de uso de metáforas que eviten la estigmatización de las personas diagnosticadas.

## 5 Metodología y planificación

El proyecto se divide en cinco grandes estadios metodológicos que se concretan en las siguientes tareas:

- Constitución del corpus.
- Anotación.
- Experimentación en detección automática de expresiones metafóricas.
- Análisis.
- Integración de resultados y acciones de transferencia.

### 5.1 Constitución del corpus

El corpus MOMENT se constituirá, con previa gestión de los permisos correspondientes, a partir de testimonios de personas diagnosticadas por un TMG y profesionales de la salud mental publicados en Internet. Se establecen cuatro tipologías de participantes dentro de cada colectivo analizado:

- Tipo de trastorno de la persona diagnosticada: esquizofrenia, trastorno bipolar, depresión grave, TOC.
- Tipo de profesional: psiquiatra, psicólogo, enfermero, educador social.

A su vez, para cada grupo participante nos basaremos en tres grandes tipos de fuentes documentales, las cuales representan diferentes grados de espontaneidad del discurso: 1) Foros: generalmente coordinados por profesionales de la salud mental, en que las personas diagnosticadas pueden compartir sus experiencias; 2) Blogs personales o integrados en webs de asociaciones de salud mental; y 3) Entrevistas.

Además del corpus objeto principal del estudio se ha constituido también un corpus de contraste sobre trastornos mentales compuesto por textos científicos o divulgativos. El objetivo principal de este corpus es realizar estadísticas sobre el uso de determinadas palabras en un contexto de utilización diferente. Para constituir el corpus en español se han utilizado las siguientes fuentes:

- El manual MSD en español, tanto en su versión profesional (sección trastornos psiquiátricos), como en la versión para el hogar (trastornos de la salud mental).
- Wikipedia en español: se han tenido en cuenta todas las categorías relacionadas con la categoría "Salud mental"hasta un nivel 5, lo que supone un total de 58 categorías y 901 entradas de la Wikipedia.

### 5.2 Anotación del corpus

Para analizar lingüísticamente el corpus con rigor, se aplicará el método Metaphor Identification Procedure (MIP) definido en Praglejaz Group (2007) y Steen et al. (2010), ampliado en Climent y Coll-Florit (2017) y empleado en la anotación de corpus en español por Martínez Santiago et al. (2014). El MIP se basa en la detección de unidades léxicas que, en el contexto del discurso analizado, se utilizan en un sentido distinto del determinado como básico en recursos léxicos de referencia. Este procedimiento se ha convertido en estándar en los trabajos de investigación basados en la identificación de metáforas conceptuales. Este método se evolucionará en función de las especificidades del proyecto.

Con carácter previo a la anotación del corpus MOMENT, se estructurará una interfaz y se realizará una primera anotación prototipo con un subcorpus equilibrado para cada una de las 24 dimensiones analizadas (8 subtipos de participantes y 3 géneros documentales), con el fin de detectar problemas y fijar los parámetros del corpus final. Asimismo se planteará una fase de predetección automática que contribuirá a constituir la base de corpus a analizar por los lingüistas.

Simultáneamente, se establecerá una guía de criterios de anotación, basada en trabajos previos de anotación de expresiones metafóricas en corpus (Deignan, 2006; Shutova, 2010). La guía incluirá el alcance de la anotación: campos semánticos tratados y nivel de granularidad de la anotación.

### 5.3 Detección automática de expresiones metafóricas

La detección automática de expresiones metafóricas (Shutova, 2010) es un campo de investigación que despierta mucho interés en el área de PLN. Se han desarrollado diversas técnicas, entre las que cabe destacar: el uso de dominios semánticos (Schulder y Hovy, 2014), patrones léxicos (Goatly, 2011), patrones oracionales (Birke y Sarkar, 2006) o semántica distribucional (Tsvetkov et al., 2014). El corpus que se anotará de manera manual en este proyecto nos servirá para desarrollar nuevas técnicas para la detección automática de metáforas.

### 5.4 Clasificación de metáforas y análisis comparativo

El método analítico que se usará para la anotación reside en la detección y clasificación de los dominios origen y destino de las metáforas. En trabajos previos realizados por los miembros del equipo (Coll-Florit, Miranda, y Climent, 2018; Climent y Coll-Florit, 2017), se han detectado dos grandes tipos de metáforas recurrentes sobre el dominio destino del trastorno mental y la vida de la persona diagnosticada: las metáforas de guerra y las metáforas de viaje:

EL TRASTORNO ES UNA GUERRA
*"Son grandes luchadores, se enfrentan con su enfermedad con las armas que la propia enfermedad les deja".*
EL TRASTORNO ES UN VIAJE
*"La mayoría avanza, si están bien tratados, la mayoría avanza".*

La detección de metáforas conceptuales aporta interpretación al discurso, ya que las metáforas revelan unas determinadas asunciones y conceptualizaciones, de entre otras posibles. En el ejemplo, el dominio origen GUERRA destaca la idea de confrontación y oculta otros aspectos como el progreso, una idea que se resalta con el uso del dominio origen VIAJE.

Una vez realizada la anotación, se clasificarán las metáforas detectadas y se realizarán análisis comparativos de uso intergrupales (personas diagnosticadas vs. profesionales) e intragrupales. Finalmente, a partir de las regularidades detectadas, el último estadio consistirá en la identificación de discursos dominantes, tomando como referencia los tres grandes paradigmas teóricos para interpretar la relación entre el ser humano y la enfermedad: los modelos biomédico, biopsicosocial y hermenéutico-crítico (Martínez, 2008).

## 5.5 Acciones de transferencia

En la última fase del proyecto, en función de los resultados obtenidos, se elaborará una guía de recomendaciones de uso de metáforas, con el fin de que los profesionales de la salud mental, así como la sociedad en general, puedan usar ciertas metáforas sobre el trastorno mental de manera más consciente y respetuosa con las personas diagnosticadas.

## *Bibliografía*


Birke, J. y A. Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. En *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Boylstein, C., M. Rittman, y R. Hinojosa. 2007. Metaphor shifts in stroke recovery. *Health communication*, 21(3):279–287.

Climent, S. y M. Coll-Florit. 2017. La metáfora conceptual en el discurso psiquiátrico sobre la esquizofrenia. *Ibérica*, (34):187–208.

Coll-Florit, M., X. Miranda, y S. Climent. 2018. Metáforas de la esquizofrenia. un estudio sobre el discurso de afectados y profesionales. *Revista Española de Lingüística Aplicada (RESLA) / Spanish Journal of Applied Linguistics (SJAL)*, (En prensa).

Deignan, A. 2006. The grammar of linguistic metaphors. En A. Stefanowitsch y S. T. Gries, editores, *Corpus-Based Approaches to Metaphor and Metonymy*. Mouton de Gruyter, Berlin.

Gibbs Jr, R. W. y H. Franks. 2002. Embodied metaphor in women's narratives about their experiences with cancer. *Health Communication*, 14(2):139–165.

Goatly, A. 2011. *The language of metaphors*. Routledge.

Lakoff, G. 1993. *The contemporary theory of metaphor*. Cambridge University Press, New York.

Lakoff, G. 2004. *Don't think of an elephant!: Know your values and frame the debate*. Chelsea Green Publishing.

Lakoff, G. y M. Johnson. 1980. *Metaphors we live by*. University of Chicago press.

Martínez Santiago, F., M. A. García Cumbreras, M. C. Díaz Galiano, y A. Montejo Ráez. 2014. Etiquetado de metáforas lingüísticas en un conjunto de documentos en español. *Procesamiento del Lenguaje Natural*, (53).

Martínez, A. 2008. *Antropología médica: Teorías sobre la cultura, el poder y la enfermedad*. Anthropos, Barcelona.

Pragglejaz Group. 2007. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and symbol*, 22(1):1–39.

Schulder, M. y E. Hovy. 2014. Metaphor detection through term relevance. En *Proceedings of the Second Workshop on Metaphor in NLP*, páginas 18–26.

Semino, E., Z. Demjén, A. Hardie, S. Payne, y P. Rayson. 2017. *Metaphor, Cancer and the End of Life: A Corpus-based Study*. Routledge.

Shutova, E. 2010. Models of metaphor in nlp. En *Proceedings of the 48th annual meeting of the association for computational linguistics*, páginas 688–697. Association for Computational Linguistics.

Steen, G. J., E. Biernacka, A. G. Dorst, A. Kaal, C. I. López, y T. Pasma. 2010. Finding metaphorically used words in natural discourse. En G. Low Z. Todd A. Deignan, y L. Cameron, editores, *Researching and Applying Metaphor in the Real World*. John Benjamins, Amsterdam, páginas 165–184.

Tsvetkov, Y., L. Boytsov, A. Gershman, E. Nyberg, y C. Dyer. 2014. Metaphor detection with cross-lingual model transfer. En *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volumen 1, páginas 248–258.

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 143-146 recibido 20-03-2018 revisado 28-04-2018 aceptado 03-05-2018

# QUALES: Estimación Automática de Calidad de Traducción Mediante Aprendizaje Automático Supervisado y No-Supervisado

## *QUALES: Machine Translation Quality Estimation via Supervised and Unsupervised Machine Learning*

**Thierry Etchegoyhen**[1], **Eva Martínez Garcia**[1], **Andoni Azpeitia**[1], **Iñaki Alegria**[2], **Gorka Labaka**[2], **Arantza Otegi**[2], **Kepa Sarasola**[2], **Itziar Cortes**[3], **Amaia Jauregi**[3], **Igor Ellakuria**[4], **Eusebi Calonge**[5], **Maite Martin**[5]
[1]Vicomtech - {tetchegoyhen, emartinez, aazpeitia}@vicomtech.org
[2]IXA taldea, University of the Basque Country (UPV/EHU)
{i.alegria, gorka.labaka, arantza.otegi, kepa.sarasola}@ehu.eus
[3]Elhuyar - {i.cortes, a.jauregi}@elhuyar.eus
[4]ISEA - isantos@iseamcc.net
[5]Ametzagaiña - ecalonge@ametza.com, maite@adur.com

**Resumen:** La estimación automática de calidad (EAC) de la traducción automática consiste en medir la calidad de traducciones sin acceso a referencias humanas, habitualmente mediante métodos de aprendizaje automático. Un buen sistema EAC puede ayudar en tres aspectos del proceso de traducción asistida por medio de traducción automática y posedición: aumento de la productividad (descartando traducciones automáticas de mala calidad), estimación de costes (ayudando a prever el coste de posedición) y selección de proveedor (si se dispone de varios sistemas de traducción automática). El interés en este campo de investigación ha crecido significativamente en los últimos años, dando lugar a tareas compartidas a nivel mundial (WMT) y a una fuerte actividad científica. En este artículo, se hace un repaso del estado del arte en este área y se presenta el proyecto QUALES que se está realizando.
**Palabras clave:** Estimación de calidad, Traducción automática, Aprendizaje automático

**Abstract:** The automatic quality estimation (QE) of machine translation consists in measuring the quality of translations without access to human references, usually via machine learning approaches. A good QE system can help in three aspects of translation processes involving machine translation and post-editing: increasing productivity (by ruling out poor quality machine translation), estimating costs (by helping to forecast the cost of post-editing) and selecting a provider (if several machine translation systems are available). Interest in this research area has grown significantly in recent years, leading to regular shared tasks in the main machine translation conferences and intense scientific activity. In this article we review the state of the art in this research area and present project QUALES, which is under development.
**Keywords:** Quality Estimation, Machine Translation, Machine Learning

## *1 Participantes y entidades financiadoras*

QUALES es un proyecto de investigación subvencionado por el Gobierno Vasco a través de la convocatoria de ayudas ELKARTEK 2017 de la Agencia Vasca de desarrollo empresarial Spri.[1]

El proyecto tiene una duración total de 21 meses, con comienzo el 1 de abril de 2017 y finalización el 31 de diciembre de 2018.

QUALES ha sido diseñado y se está lle-

[1]http://www.spri.eus

ISSN 1135-5948. DOI 10.26342/2018-61-18

vando a cabo por el siguiente consorcio: Vicomtech[2], grupo IXA de la UPV/EHU[3], Elhuyar[4], ISEA[5] y Ametzagaina[6]. Son empresas adheridas Argia, Mondragon Lingua y Eleka. El proyecto tiene asignado el código KK-2017/00094 y el sitio web asociado es http://quales.eus/

## 2 Contexto y motivación

Con los avances obtenidos por los métodos basados en datos, estadísticos o enmarcados en redes neuronales profundas, la traducción automática (TA) ha logrado niveles de calidad suficientes para su uso en la industria. En particular, los proveedores de servicios lingüísticos de traducción requieren poder integrar componentes de traducción automática para dar una respuesta eficiente y de alta calidad a las demandas de traducción en entornos y dominios variados. Pese a estos progresos notables, la calidad de las traducciones automáticas puede variar significativamente según el dominio, los idiomas considerados o la complejidad de los segmentos individuales por traducir. Esta variabilidad genera problemas bien identificados, entre los cuales los principales son:

- *Productividad*: las traducciones automáticas de pésima calidad requieren un esfuerzo cognitivo importante por parte de los traductores profesionales, en particular para determinar si ciertas partes de la traducción automática son recuperables y qué correcciones aplicar. El enfrentarse a traducciones de baja calidad genera así pérdidas de productividad y frustración para los profesionales del sector. De forma similar, traducciones automáticas correctas a nivel gramatical pero con errores a nivel semántico implican un esfuerzo importante de identificación y corrección de los errores de traducción.
- *Estimación de costes*: los proveedores de servicios lingüísticos ofertan varios servicios según las demandas y la complejidad de las traducciones. Estos servicios pueden ser traducción automática completa, con posedición humana, o realizada por completo por traductores humanos, con o sin el apoyo de memorias de traducción existentes. A cada uno de estos servicios le corresponde una tarificación precisa según el esfuerzo necesario, desde el uso de TA únicamente, con coste mínimo, hasta la traducción humana por completo, con coste máximo. Para poder establecer los costes correctos, es necesario poder determinar la calidad de cada traducción automática y establecer así automáticamente el tipo de servicio óptimo correspondiente.
- *Selección de traducciones*: cada industria con necesidades multilingües puede acceder a una gama variada de servicios de traducción automática, desde sistemas propietarios entrenados para diferentes dominios hasta las ofertas de traducción genéricas online. Cada sistema de TA suele producir traducciones diferentes debidas a los diferentes datos y métodos empleados para el modelado del sistema, con niveles de calidad variables. Resulta imprescindible, entonces, la estimación automática de la calidad de cada una de las traducciones generada por los diferentes sistemas, para poder seleccionar así el mejor conjunto de traducciones.

Tradicionalmente, la calidad de las traducciones automáticas se ha medido de forma estática, comparando un subconjunto de las traducciones con referencias creadas por profesionales humanos. Las comparaciones se establecen usando métricas automáticas que calculan, con cierta aproximación, la distancia entre las traducciones automáticas y las referencias. Este enfoque sigue siendo fundamental para obtener una medida de la calidad general de los sistemas de TA en los dominios considerados y permitir avances incrementales en el desarrollo de los sistemas en función de los resultados objetivos obtenidos con estas métricas.

Pese a estos aspectos positivos, este tipo de medida de calidad es problemático en dos aspectos principales. En primer lugar, la correlación entre los resultados de las métricas automáticas y las valoraciones humanas es baja a nivel de frases o segmentos, lo que no permite una estimación adecuada de la calidad de los segmentos individuales. En segundo lugar, este enfoque implica disponer

[2]http://www.vicomtech.org
[3]http://ixa.eus
[4]http://www.elhuyar.eus/
[5]http://www.iseamcc.net
[6]http://www.ametza.com

de traducciones de referencia, lo cual no es realista a la hora de evaluar la calidad de los amplios volúmenes de traducciones automáticas generadas.

Esta limitación de las métricas estáticas a la hora de medir la calidad de las traducciones automáticas impone el desarrollo de métodos alternativos. La estimación automática de calidad (EAC) (Blatz y otros, 2004; Specia, Raj, y Turchi, 2010) se centra en responder a este reto, a través de sistemas que permitan medir la calidad de traducciones individuales, sin acceso a referencias humanas, empleando habitualmente métodos de aprendizaje automático. Este campo de investigación y desarrollo ha crecido significativamente en los últimos años, dando lugar a tareas compartidas a nivel mundial y una fuerte actividad científica.

Una EAC exitosa permitiría aportar una solución a los tres problemas principales indicados anteriormente, al ofrecer mecanismos adecuados de medida de calidad para cada segmento de traducción automática generado por cualquier sistema de TA. A estos retos responde el proyecto Quales, mediante el desarrollo de métodos supervisados y no-supervisados para la estimación automática de calidad.

## 3 Estado del arte

Los enfoques supervisados han sido el paradigma dominante en EAC, donde traducciones automáticas anotadas o poseditadas se usan para entrenar clasificadores (Blatz y otros, 2004; Quirk, 2004) o regresores (Specia y otros, 2009). Los sistemas participantes en las tareas compartidas de las conferencias WMT han sido así tipicamente basados en enfoques supervisados, con diferencias centradas en diferentes conjuntos de características (*features*) o en los métodos de aprendizaje automático utilizados, p. ej. Support Vector Machines o Gaussian Processes (Callison-Burch y otros, 2012).

Los sistemas baseline estándares para la tarea se generan habitualmente con las herramientas QUEST (Specia et al., 2013), o QUEST++ (Specia, Paetzold, y Scarton, 2015), en base a 17 características que incluyen puntuación de perplejidad en base a modelos de lenguaje, probabilidades de traducción léxica, o ratios de ocurrencias de palabras, entre otros.

En trabajos recientes, los enfoques basados en redes neuronales han sido empleados también de forma exitosa para la tarea de estimación automática de calidad, bien sea mediante características adicionales (Shah y otros, 2016) o como sistemas EAC completos (Kim, Lee, y Na, 2017; Martins, Kepler, y Monteiro, 2017). En la última edición de la tarea compartida en WMT 2017, los mejores sistemas basados en redes neuronales incrementaron notablemente las prestaciones de la baseline (Bojar y otros, 2017). Así, en la traducción de alemán a inglés los valores del índice de Pearson de los dos mejores sistemas fueron de 0,728 y 0,715, muy por encima del valor baseline (0,441). Para el sentido de traducción inverso, los resultados fueron similares con 0,695 y 0,673 para los mejores sistemas, y 0,307 para la baseline.

Aunque permitan obtener las estimaciones las más precisas a día de hoy, los enfoques supervisados dependen de anotaciones o posediciones humanas. Este aspecto es problemático dada la gran cantidad de diferentes dominios y pares de idiomas en los que se requiere aplicar la tecnología. Considerando los recursos y esfuerzos necesarios para anotar o poseditar conjuntos de muestras adecuados para entrenar sistemas de calidad, el coste de los métodos supervisados puede resultar prohibitivo.

Enfoques no-supervisados que no necesiten datos anotados, basados estrictamente en las características de las frases de origen y/o de las frases traducidas, tienen la notable ventaja de poder adaptarse más fácilmente a distintos dominios y pares de idiomas. Pese a ofrecer este tipo de ventajas, pocos estudios se han centrado en enfoques no-supervisados para EAC. Uno de ellos es (Moreau y Vogel, 2012), donde la estimación de calidad se ejecuta en base a amplios conjuntos de n-gramas y medidas de similitud. (Popovic, 2012) es otra alternativa, basada en la combinación de puntuaciones obtenidas por modelos de lenguaje y probabilidades de traducción léxica sobre morfemas y categorías gramaticales. Ninguno de estos enfoques ha logrado superar hasta hoy las baselines supervisadas.

## 4 QUALES

En el marco de QUALES, se investigan tanto métodos supervisados avanzados basados en redes neuronales como métodos no-supervisados que permitan desarrollar estimadores de calidad de forma eficiente en dis-

tintos casos de uso.

Tras su puesta en marcha en 2017, el proyecto ha logrado los primeros resultados siguientes:

- Creación de datos de entrenamiento y validación manuales y sintéticos para los pares de idiomas euskera-castellano e inglés-castellano en el dominio de las noticias.
- Despliegue de baselines supervisadas basadas en QUEST++.
- Desarrollo de sistemas de EAC basados en redes neuronales y aprendizaje profundo, explotando espacios vectoriales bilingües.
- Desarrollo de un sistema de EAC basado en características mínimas, en versión supervisada y no-supervisada. Ambas versiones superan significativamente las baselines sobre los datos de las tareas compartidas WMT 2015, 2016 y 2017.

Los primeros resultados del proyecto son satisfactorios, en particular los obtenidos mediante métodos minimalistas no-supervisados que superan significativamente a sistemas supervisados robustos y permiten un despliegue eficiente de estimadores fiables para nuevos dominios.

QUALES aportará además los primeros resultados para el par de idiomas euskera-castellano en el campo de la estimación automática de calidad, lo cual constituye un objetivo importante del proyecto.

Durante 2018, el esfuerzo se centrará en extender y mejorar los primeros sistemas desarrollados, y en validar los resultados obtenidos. La convocatoria en la que se enmarca el proyecto apoya a proyectos de investigación con alto potencial industrial y se validará el potencial de los métodos explorados para un uso en entornos profesionales.

## *Bibliografía*


Blatz, J. et al. 2004. Confidence estimation for machine translation. En *Proceedings of COLING*, páginas 315–321.

Bojar, O. et al. 2017. Findings of the 2017 conference on machine translation. En *Proceedings of the Second Conference on Machine Translation*, páginas 169–214, Copenhagen, Denmark.

Callison-Burch, C. et al. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. En *Proceedings of the Seventh Workshop on Statistical Machine Translation*.

Kim, H., J.-H. Lee, y S.-H. Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. En *Proceedings of the Second Conference on Machine Translation*, páginas 562–568, Copenhagen, Denmark.

Martins, A. F. T., F. Kepler, y J. Monteiro. 2017. Unbabel's participation in the wmt17 translation quality estimation shared task. En *Proceedings of the Second Conference on Machine Translation*, páginas 569–574, Copenhagen, Denmark.

Moreau, E. y C. Vogel. 2012. Quality estimation: an experimental study using unsupervised similarity measures. En *Proceedings of the Seventh Workshop on Statistical Machine Translation*, páginas 120–126.

Popovic, M. 2012. Morpheme- and pos-based IBM1 and language model scores for translation quality estimation. En *Proceedings of the Seventh Workshop on Statistical Machine Translation*, páginas 133–137.

Quirk, C. 2004. Training a sentence-level machine translation confidence measure. En *Proceedings of LREC*, páginas 825–828.

Shah, K. et al. 2016. SHEF-LIUM-NN: Sentence level Quality Estimation with Neural Network Features. En *Proceedings of the First Conference on Machine Translation*, volumen 2, páginas 838–842.

Specia, L. et al. 2009. Estimating the sentence-level quality of machine translation systems. En *13th Conference of the European Association for Machine Translation*, páginas 28–37.

Specia, L., G. Paetzold, y C. Scarton. 2015. Multi-level translation quality prediction with QUEST++. *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, páginas 115–120.

Specia, L., D. Raj, y M. Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50.

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 147-150 recibido 20-03-2018 revisado 29-04-2018 aceptado 03-05-2018

# AMIC: Affective multimedia analytics with inclusive and natural *communication*

## *AMIC: Análisis afectivo de información multimedia con comunicación inclusiva y natural*

**Alfonso Ortega[1], Eduardo Lleida[1], Rubén San-Segundo[2], Javier Ferreiros[2], Lluis Hurtado[3], Emilio Sanchís[3], María Ines Torres[4], Raquel Justo[4]**
[1] ViVoLab-Universidad de Zaragoza C/ María de Luna 1. 50018 Zaragoza
[2] GTH-Universidad Politécnica Madrid Ciudad Universitaria s/n. Madrid
[3] ELiRF-Universitat Politècnica València Camino de Vera s/n 46022 Valencia
[4] SPIN-Universidad del País Vasco, Campus de Leioa. 48940 Leioa
ortega@unizar.es, ruben.sansegundo@upm.es, lhurtado@dsic.upv.es, manes.torres@ehu.eus

**Abstract:** Traditionally, textual content has been the main source of information extraction and indexing, and other technologies that are capable of extracting information from the audio and video of multimedia documents have joined later. Other major axis of analysis is the emotional and affective aspect intrinsic in human communication. This information of emotions, stances, preferences, figurative language, irony, sarcasm, etc. is fundamental and irreplaceable for a complete understanding of the content in conversations, speeches, debates, discussions, etc. The objective of this project is focused on advancing, developing and improving speech and language technologies as well as image and video technologies in the analysis of multimedia content adding to this analysis the extraction of affective-emotional information. As additional steps forward, we will advance in the methodologies and ways for presenting the information to the user, working on technologies for language simplification, automatic reports and summary generation, emotional speech synthesis and natural and inclusive interaction.
**Keywords:** Audio, Speech, Language, Multimedia Analytics, Affective, natural inclusive communication

**Resumen:** Tradicionalmente, el análisis de los contenidos textuales ha sido la principal fuente de extracción y catalogación de contenidos multimedia y a él se han ido sumando tecnologías que son capaces de extraer información del audio y del video. Un nuevo eje de análisis es la vertiente emocional-afectiva intrínseca en la comunicación humana. Esta información de emociones, posiciones, preferencias, lenguaje figurativo, ironía, sarcasmo, etc. Es fundamental para una comprensión total del contenido de conversaciones, discursos, debates, etc. El objetivo de este proyecto se centra en avanzar en el desarrollo y mejora de prestaciones de las tecnologías del habla, el lenguaje, la imagen y el vídeo para el análisis de contenidos multimedia y añadir a este análisis la extracción de información afectiva-emocional. Como pasos adicionales, se avanzará en los métodos de presentación de resultados al usuario, trabajando en tecnologías de simplificación del lenguaje, generación automática de resúmenes e informes, síntesis de voz emocional e interacción natural e inclusiva.

**Palabras clave:** Audio, voz, lenguaje, análisis multimedia, afectivo, comunicación natural inclusiva.

## *1 Project consortium*

In this project are involved four partners:
- Universidad de Zaragoza (ViVoLab).
- Universidad Politécnica de Madrid (GTH).
- Universidad Politécnica de Valencia (ELiRF).
- Universidad del País Vasco (SPIN).

The project is founded by the Ministerio de Economía y Competitividad under the grant TIN2017-85854-C4-1-R and lasts for three

ISSN 1135-5948. DOI 10.26342/2018-61-19 

years. It can be considered a natural extension of the previous one ASLP-MULAN project (Ferreiros et al., 2016), where the same partners were involved.

## *2 Introduction*

The scientific community is highly interested in developing new effective and efficient content-based indexing and retrieval methods and techniques to extract all the information from multimedia data. The AMIC project aims to advance on audio, speech and language technologies as well as image and video technologies in the analysis of multimedia content adding to this analysis the extraction of affective-emotional information providing to the user a natural and inclusive interaction.

The challenges and goals of affective computing and multimedia analytics must merge to accomplish the task of providing a comprehensive analysis of each topic of interest. Automatic extraction of the information contained in multimedia data including those pieces of information coming from social media is essential nowadays for multiple purposes. Affective information plays a key role in audiovisual and textual information that must be taken into account in order to assist multimedia computing (processing, indexing, retrieval …). Despite of the progress in multimedia content analysis there is much work to be done in the field to provide effective ways of integrating all possible sources of information in this analysis: emotional cues, figurative language, stance, preference, reputation…

## *3 Technologies involved in the project*

The main technologies involved in the AMIC project are:

- *Machine learning*: Recent advances in deep learning (Hinton, Simon & Teh, 2006) have provided a great progress in the field of machine learning, especially in tasks where we have to deal with raw signals like audio processing (Hinton et al, 2012), image classification (Krizhevsky, Sutskever, & Hinton, 2012), natural language processing (Mikolov et al., 2013), but there are new challenges when dealing with multimedia content, and these techniques must be adapted to different domains and contexts.
- *Speech technologies*: In the last years, deep learning has had a great impact on ASR core technology and its applications (Deng, 2016), (Amodei et al., 2016), but there are new challenges when dealing with multimedia content: open and dynamic vocabularies, adaptation to new speakers, and adaptation to new acoustic environments.

In the field of language and speaker identification/diarization also Deep Neural Networks (DNNs) have been recently used, mainly as a bottleneck feature extraction, in order to improve the behavior of the classical i-vector approaches (Miguel et al., 2017), (Viñals et al., 2017), (Martínez-González et al., 2017).
On the other hand, emotion analysis can be considered from the point of view of speech recognition, by means of classifiers based on features associated to emotions, or speech synthesis using generative models that can be adapted to different emotions, styles and expressivity intensities (Lorenzo-Trueba et al, 2015).

- *Natural language technologies*: Two main areas can be considered in the scope of this project. On the one hand, information about the emotional states of the speaker can be extracted from the speech transcriptions by using statistical multilayer classifiers or deep learning methods. The intentionality of the text can be also identified by detecting figurative language forms, like sarcasm, irony, nastiness, or humor (Justo et al, 2014) (Hurtado et al, 2017). This information is very useful for the Sentiment Analysis problem, where phenomena such as irony, sarcasm and some types of humor like pun, limits the accuracy of the systems and new methods and resources must be used.

On the other hand, a deeper analysis of the text, based on its semantics, is necessary to tackle with important task as Social Media Analytics. Sentiment Analysis at different levels, user profiling, and stance detection are some interesting problems that can be addressed for monitoring social networks as Twitter, Facebook, or Instagram. One of the successes of the neural networks applied to the Natural Language Processing (NLP) area is to model complex relationships between the words of a document generating continuous representations rich in syntactic/semantic information. Unsupervised generation of these continuous of words and phrases (word

embeddings) take into account the complex relationships between the words of a document. Knowledge graphs could be used to enrich the representation of the semantics of words.

- *Audiovisual technologies*: The analysis of visual content can help to catalogue the type of scene, to segment the video in clips or to search for specific concepts. This information can be used for video retrieval, or to improve the video content analytics.

  Multimedia content summarization refers to text, audio and video-based summarization. The aim is to produce a condensed representation that captures the core meaning. Recent advances of DNNs have given promising results by using convolutional neuronal networks for feature extraction and Long-Short Term Memory (LSTM) to model temporal dependency among video frames (Zhang et al, 2016).

- *Inclusive communication technologies*: People can have difficulty in communication for many different reasons. Physical disabilities, motor co-ordination problems or learning difficulties can make hard to produce speech or handle spoken language. Augmentative and alternative communication (AAC) includes all forms of communication (other than oral speech) that are used to express thoughts, needs, wants, and ideas. Recently there are some approaches for text-to-pictogram translation using simple NLP tools at the sentence level (García et al, 2015). However, there is still a lot of research to be done to make more useful document-to-pictograms translation, where keyword extraction, text summarization and language simplification are key NLP tools.

## *4 Project objetives*

### 4.1 Strategic objectives

The main strategic goal is to progress a set of diverse technologies and use them to deal with affective analysis on multimedia documents and affect-aware person-computer interactive systems. We are committed to contribute to an improved study on all kind of sources of information including traditional broadcast media and new massive and heterogeneous social media. We aim at proposing novel technological solutions to support a comprehensive information extraction of multimedia sources that includes developing audio, image, speech and language technologies devoted to: multimedia information extraction and processing, affect aware multimedia analytics, and natural, affective and inclusive communication. Additionally, we are committed to find efficient methods to manage and integrate these new sources of information into multimedia analytics systems and to provide effective ways of including affective aspects into natural and inclusive human machine interaction systems.

### 4.2 Scientific-Technological objectives

Following the project structure, our scientific-technological goals are:

- To develop technologies for audio, video, speech and text processing intended to
  1. Transcribe the speech content of multimedia documents into text.
  2. Use Web of Data as a source of knowledge to improve language, understanding, and aspect-based polarity models.
  3. Identify the language and the speaker automatically from the audio.
  4. Analyze the video of each multimedia document to extract useful information such as scenarios or characters.
- To develop technologies for affective analysis intended to
  1. Extract emotional cues from video, text and audio documents
  2. Study the impact of multimedia content on users while they are watching or listening to this content.
  3. Process figurative language, detecting and interpreting pun, irony and sarcasm.
  4. Automatically detect the stance of people involved in conversations, identify the reputation of an institution or company and track trends in social media.
- To develop technologies for natural, affective and inclusive communication devoted to
  1. Automatically generate reports and summaries out of the information extracted from multimedia documents using simple language.
  2. Synthesize speech with affective aspects such as expressivity control through emotion and style transplantation.

3. Develop person-computer interactive systems taking into account emotional and inclusive aspects including alternative and augmentative communication.

### 4.3 Transferring knowledge objectives

We propose three main objectives related to the knowledge transfer to the society:

- To develop and evaluate an application demonstrator: Affective Multimedia analysis platform with inclusive and natural interface
- To develop multimedia annotated resources and software tools freely available.
- To train experts in the developed technologies that may be employed by companies interested in our results.

## *Acknowledgments*

This work is supported by Ministerio de Economía y Competitividad under the grants TIN2017-85854-C4-(1, 2, 3, 4)-R.

## *References*

Amodei, D., S. Ananthanarayanan, R. Anubhai, , J. Bai, E. Battenberg, C. Case and J. Chen. 2016. Deep speech 2: End-to-end speech recognition in English and mandarin. In Int. Conf. on Machine Learning, pp. 173-182.

Deng, L. 2016. Deep learning: from speech recognition to language and multimodal processing. APSIPA Transactions on Signal and Information Processing.

Ferreiros, J., J.M. Pardo, L.F. Hurtado, E. Segarra, A. Ortega, E. Lleida, M.I. Torres, and R. Justo, 2016. ASLP-MULAN: Audio speech and language processing for multimedia analytics. Procesamiento del Lenguaje Natural, Vol 57, pp.147-150.

García P., E. Lleida, D. Castán, J.M. Marcos, and D. Romero, 2015. Context-Aware Communicator for All. In Universal Access in Human-Computer Interaction. Lecture Notes in Computer Science, vol 9175. Springer.

Hinton, G. E., O. Simon and Y.W. The, 2006. A fast learning algorithm for deep belief nets. Neural computation 18.7, pp. 1527-1554.

Hinton, G., L. Deng, D. Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen and T.N. Sainath 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups, Signal Processing Magazine, IEEE, vol. 29, no. 6, p. 829.

Hurtado, L., E. Segarra, F. Pla, P. Carrasco and J.A. González 2017. ELiRF-UPV at SemEval-2017 Task 7: Pun Detection and Interpretation. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017).

Justo, R., T. Corcoran, S. Lukin, M. Walker and M.I. Torres 2014. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web, Knowledge-Based Systems 69.

Krizhevsky, A., I. Sutskever and G. Hinton 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems.

Lorenzo-Trueba J., R. Barra-Chicote, R. San-Segundo, J. Ferreiros, J. Yamagishi and J.M. Montero 2015. Emotion Transplantation through Adaptation in HMM-based Speech Synthesis. Computer Speech and Language. Volume 34, Issue 1, pp. 292–307.

Martinez-González, B., J.M. Pardo, R. San-Segundo, and J.M. Montero 2016. Influence of Transition Cost in the Segmentation Stage of Speaker Diarization. In Proc of Odyssey, Bilbao-Spain.

Miguel, A., J. Llombart, A. Ortega, and E. Lleida 2017 Tied Hidden Factors in Neural Networks for End-to-End Speaker Recognition. In Proc. of Interspeech.

Mikolov, T., K. Chen, G. Corrado, and J. Dean 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Viñals, I., A. Ortega, J. Villalba, A. Miguel and E. Lleida 2017. Domain Adaptation of PLDA models in Broadcast Diarization by means of Unsupervised Speaker Clustering. Proc. Interspeech 2017.

Zhang, K., W.L. Chao, F. Sha and K. Grauman 2016. Video Summarization with Long Short-term Memory, arXiv:1605.08110v2.

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 151-154 recibido 21-03-2018 revisado 19-04-2018 aceptado 03-05-2018

# Proyecto TAGFACT: Del texto al conocimiento. Factualidad y grados de certeza en español

## *TAGFACT Project: From text to knowledge. Factuality and degrees of certainty in Spanish*

**Laura Alonso[1], Irene Castellón[2], Hortensia Curell[3], Ana Fernández-Montraveta[3], Sonia Oliver[3], Gloria Vázquez[4]**

[1]Universidad de la República, [2]Universitat de Barcelona,
[3]Universitat Autònoma de Barcelona, [4]Universitat de Lleida
[1]lauraalonsoalemany@gmail.com, [2]icastellon@ub.edu,
[3]{hortensia.curell, ana.fernandez, sonia.oliver}@uab.cat, [4]gvazquez@dal.udl.cat

**Resumen**: El objetivo general de este proyecto es crear una herramienta para la anotación de la factualidad expresada en textos en español a través del procesamiento automático. Pretendemos que dicha representación sea muy rica, por lo que se llevará a cabo desde tres ejes distintos: multinivel, multidimensional y multitextual. El análisis multinivel da cuenta de las distintas marcas lingüísticas que expresan el grado de certeza de un evento a nivel morfológico y sintáctico, pero también discursivo; el análisis multidimensional, de un número variado de las voces que evalúan dicho evento; y el análisis multitextual, de distintos textos sobre un mismo evento, siendo este último uno de los aspectos más innovadores de la propuesta.
**Palabras clave**: factualidad, procesamiento semántico, implicación textual, presuposición, modalidad, polaridad, certeza, asertividad.

**Abstract**: The main aim of this project is to create a tool for the automatic annotation of factuality-expressed in Spanish texts. We intend the representation to be very rich, so it will be carried out from three different perspectives: multilevel, multidimensional and multitextual. The multilevel analysis gives account of the various linguistic markers that express the degree of certainty of an event at the morphological and syntactic, as well as discourse, levels. The multidimensional analysis accounts for the varied number of voices that assess an event. Finally, the multitextual analysis will take into account the various texts about the same event, which is one of the most innovative aspects of our proposal.
**Keywords**: factuality, semantic processing, textual implication, presupposition, modality, polarity, certainty, assertivity.

### *1 Introducción*

En los textos se predica sobre eventos y situaciones (a partir de ahora, eventos). La asignación de valores de factualidad (certeza) a los eventos es un campo que tiene un interés muy destacado en el ámbito de la lingüística de corpus y en el PLN. La factualidad de un evento no es una propiedad intrínseca de este sino que siempre está en relación al modo en que es presentado dicho evento, teniendo en cuenta que puede ser evaluado por una o más personas y en el mismo momento o en momentos distintos.

El objetivo general del proyecto que presentamos (TAGFACT), que está en la fase inicial de desarrollo, es crear una herramienta para anotar automáticamente la factualidad de los eventos expresados en textos en español. Pretendemos desarrollar un sistema de anotación en el que se tengan en cuenta tres ejes distintos: multinivel, multidimensional y multitextual.

En este sentido, nuestra propuesta tiene de innovadora que es globalizadora, ya que pretendemos construir un sistema automático capaz de aportar las diversas interpretaciones factuales de los eventos a partir de las distintas marcas de factualidad que se pueden identificar en los textos, abarcando desde la morfología, la sintaxis y el discurso (análisis multinivel), pero también un abanico de fuentes muy

ISSN 1135-5948. DOI 10.26342/2018-61-20 

diversificado (análisis multidimensional), que van más allá de las que se encuentran en los límites de la frase y del propio texto, ya que se tienen en cuenta también distintos textos que tratan sobre el mismo evento global (análisis multitextual).

Además, queremos investigar las ventajas que puede aportar crear una herramienta que use exclusivamente conocimiento lingüístico en el caso del español, ya que para esta lengua se ha trabajado muy poco en este ámbito y básicamente usando conocimiento estadístico.

## 2 *La factualidad en la lingüística de corpus y el PLN*

Uno de los sistemas automáticos pioneros en la anotación de la factualidad son el de Light et al. (2004) y el de Medlock y Briscoe (2007). FactBank (Sauri, 2008) supuso una propuesta innovadora para la representación de información factual para el inglés. Es un referente en este ámbito, ya que el tratamiento que hace de los diferentes fenómenos lingüísticos relacionados con la factualidad es muy amplio. Contiene 8.837 eventos (208 documentos) etiquetados manualmente en función de distintas fuentes. Cabe decir que un aspecto que no se considera en el modelo de anotación propuesto es la interpretación factual del evento según el momento temporal. Tampoco distingue los eventos futuros y condicionales del resto y no queda claro cómo se evalúan en términos de factualidad los estados atemporales ni los eventos habituales.

En el marco de su tesis la autora también implementa un anotador de factualidad (De Facto). Aunque esta herramienta contiene un algoritmo para etiquetar automáticamente la factualidad, algunos de los módulos de conocimiento utilizados han sido creados manualmente, por lo que no se puede considerar que la etiquetación sea un proceso automático. Los resultados de este sistema respecto a cobertura y precisión, así como de la F-measure son buenos aunque no tiene en cuenta algunas construcciones que aportan información de factualidad, como las condicionales o las temporales, entre otras.

Aproximadamente a partir de la constitución de FactBank y De Facto, la comunidad de PLN consolida su interés sobre la factualidad. Son diversas las contribuciones que se han ido realizando en la última década en este ámbito. Asimismo, también aparecen trabajos en los que se avanza en el campo de la anotación de la modalidad (Hendrickx et al., 2012; Rupenhofer y Rehbein, 2012; Morante y Daleemans, 2012).

Velupillai (2011) trabaja con corpus del dominio de la medicina clínica y aporta una herramienta de etiquetación automática de eventos respecto a la polaridad y la certeza. El campo de la biomedicina es uno en los que más se ha trabajado en relación a la anotación de la factualidad, donde destaca Vincze et al. (2008) y Nawaz et al. (2010). Otros sistemas basados en el aprendizaje supervisado en este ámbito son los de Farkas et al. (2010), Tang et al. (2010) y Morante et al. (2010). Como herramientas construidas con conocimiento lingüístico mencionamos Harkema et al. (2009) y Wu et al. (2009).

Otros trabajos bastante recientes basados en Sauri (2008), Van Son et al. (2014) y Prabhakaran et al. (2015). Respecto a este último, el corpus anotado manualmente es de aproximadamente un millón de palabras (es el más grande que existe para el inglés), del cual un 10% aproximadamente se usa para evaluar distintas herramientas automáticas. En otros trabajos, contrariamente al marco usado en FactBank y De Facto, se opta por considerar la factualidad ligada al conocimiento del mundo (Marneffe et al., 2012).

Para el español existen algunos corpus etiquetados con información temporal. Pero en el campo de la factualidad solo conocemos el trabajo de Wonsever et al. (2016). Estos autores adoptan el modelo de Sauri (2008), con algunos cambios, y crean un corpus anotado con información sobre factualidad y una herramienta de anotación automática basada en aprendizaje automático supervisado.

Si se pretende llevar a cabo un tratamiento global de la factualidad, es necesario también identificar las equivalencias entre eventos dentro del mismo texto y entre textos distintos. Este campo está muy poco explorado, pero existe una línea en la que se trabaja la similitud textual (Agirre et al., 2014) que puede aportar luz en el tema. También los sistemas de identificación de correferencias anafóricas son importantes para tratar las correferencias de eventos. Para el español destacamos Björkelund y Kuhn (2014) y Durret y Klein (2013).

## 3 *Metodología*

Para llevar a cabo el proyecto TAGFACT el primer paso es estudiar la expresión de la factualidad en español a diferentes niveles lingüísticos (morfológico, sintáctico y

discursivo). Para ello, en este momento, estamos confeccionando un corpus sobre noticias relacionadas con la política extraído de diarios españoles de distintas tendencias. Dicho corpus se anotará manualmente por diversos miembros del proyecto y se evaluará el *agreement* con las medidas oportunas. Así obtendremos un Gold Standard que servirá como referente para evaluar la anotación automática.

En lenguas como el español, hay diversos marcadores lingüísticos de la factualidad como algunos elementos y el tiempo verbal, ya que, inicialmente el futuro y el condicional no expresan hechos. Además hay determinadas construcciones que afectan a la interpretación de la factualidad y un subconjunto de predicados que subcategorizan eventos y que proyectan información factual sobre estos (Palmer, 1986; Quirk et al., 1985). Por ejemplo, en "María *ha conseguido* pasar de curso" el verbo *conseguir* claramente proyecta la certeza del aprobado de María.

Otro elemento clave en la expresión de la factualidad es la modalidad, que ha sido categorizada en diversos tipos, entre los cuales la que tiene un efecto claro respecto a la factualidad es la epistémica. Las tres categorías que se suelen utilizar para describir la modalidad epistémica asociada a las proposiciones son: certeza, probabilidad y posibilidad. La polaridad negativa completaría la escala del continuum con los no hechos (Sauri, 2008; Repiso, 2015).

Los distintos marcadores de factualidad de las categorías mencionadas pueden interactuar entre sí en la descripción de un mismo evento. Uno de los retos del proyecto en esta fase de estudio es analizar estas combinaciones y dar cuenta del resultado que se obtiene en relación con los valores factuales.

Los desafíos más destacables en la elaboración de la herramienta giran en torno a tres ejes: en primer lugar, la desambiguación de los posibles valores de algunos elementos léxicos y construcciones; en segundo lugar, el establecimiento de correferencias entre eventos, más allá de la frase, dentro de una misma noticia y entre noticias distintas; y, en tercer lugar, la resolución automática de las distintas interpretaciones que se van construyendo en las oraciones que incorporan distintas fuentes (discurso directo e indirecto) y también en los casos de sintaxis altamente compleja a nivel de subordinación.

## 4 Conclusiones

En TAGFACT pretendemos aportar una herramienta que represente la información diferenciando los hechos reales de aquellas expresiones que presentan creencias, opiniones o posibilidades y constituir un corpus de referencia en la anotación de la factualidad.

En este sentido, la detección del grado de certeza de los eventos sobre los que se predica en los textos y su representación en un lenguaje formal se considera vital para poder realizar inferencias con el objetivo de adquirir nuevo conocimiento y actualizar y crear ontologías y bases de conocimiento, basadas en hechos, de diferentes ámbitos. Hay que tener en cuenta que muchas aplicaciones de diversos ámbitos utilizan ontologías, y el mantenimiento de estas ontologías es un proceso muy costoso. La automatización de esta tarea permite ahorrar esfuerzos en el desarrollo del conocimiento necesario para cualquier aplicación automática que trabaje con lenguaje. Así, el recurso creado tendrá gran utilidad en otras aplicaciones como el análisis de opiniones (*sentiment analysis*), la extracción de información textual, los sistemas de pregunta-respuesta, en las que la interpretación del grado de factualidad es una tarea primordial.

Finalmente, otro valor añadido del proyecto es que el anotador automático que se creará, al estar basado en conocimiento lingüístico, será más estable entre diferentes dominios que los sistemas basados en aprendizaje automático, que son más dependientes del dominio de desarrollo. Esta característica dotará a la herramienta de más potencialidad para procesar de forma automática grandes volúmenes de textos y de gran diversidad.

## Agradecimientos

Este proyecto está siendo desarrollado por miembros del grupo GRIAL (http://grial.uab.es) y está financiado por el Ministerio de Economía, Industria y Competitividad - FFI2017-84008-P.

## Bibliografía

Agirre, E., C. Baneab, C. Cardiec, D. Cerd, M. Diabe, A. González-Agirre, W. Guof, R. Mihalceab, G. Rigau y J. Wiebe. 2014. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. *Proceedings of the 8th International Workshop on Semantic Evaluation*, 81-91.

Björkelund, A. y J. Kuhn. 2014. Learning structured perceptrons for coreference resolution with latent antecedents and non-

local features. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 47-57.

Durret, G. y D. Klein. 2013. Easy victories and uphill battles in coreference resolution. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1971-1982.

Farka, R., V. Vincze, G. Móra, J. Csirik y G. Szarvas. 2010. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. *Proceedings of the 14th CoNLL*, 1-12.

Harkema, H., J. N. Dowling, T. Thornblade y W. W. Chapman. 2009. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics*, 42: 839-851.

Hendrickx, I., A. Mendes y S. Mencarelli. 2012. Modality in text: a proposal for corpus annotation. *LREC 2012, Eighth International Conference on Language Resources and Evaluation*, 1805-1812.

Light, M., X. Ying Qiu y P. Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. En L. Hirschman y J. Pustejovsky (Eds.), *HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases, 17-24.*

Marneffe, M. C., C. D. Manning y C. Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics* 38-2:301-333.

Medlock, B. y T. Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. *Proceedings of the 45th ACL*, 992-999.

Morante, R., V. Van Asch and W. Daelemans. 2010. Extraction of biomedical events. *LOT Occasional Series*, 16. 91-105.

Morante, R. y W. Daelemans. 2012. Annotating Modality and Negation for a Machine Reading Evaluation. *CLEF 2012 Evaluation Labs and Workshop Online Working Notes.* Disponible en: http://www.clef-initiative.eu/documents/71612/463956e9-2b22-4e68-aa39-b711302c97b1

Nawaz, R., P. Thompson y S. Ananiadou. 2010. Evaluating a meta-knowledge annotation scheme for bio-events. *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, 69-77.

Palmer, F. R. 2004. *Mood and Modality*. Cambridge: Cambridge University Press.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Repiso, I. 2015. Talking about counterfactual worlds. A comparative study of French and Spanish. *Journal of Romance Studies* 15-1, 52-72.

Rupenhofer, J. e I. Rehbein. 2012. Annotating the senses of the English modal verbs. *LREC 2012, Eighth International Conference on Language Resources and Evaluation*, 1538-1545.

Sauri, R. 2008. *A Factuality Profiler for Eventualities in Text*. Ph.D. Thesis. Brandeis University.

Tang, B., X. Wang, X. Wang, B. Yuan y S. Fan. 2010. A cascade method for detecting hedges and their scope in natural language text. In *Proceedings 14th CoNLL*, 13-17.

Van Son, C., M. van Erp, A. Fokkens y P. Vossen. 2014. Hope and Fear: Interpreting Perspectives by Integrating Sentiment and Event Factuality. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, 26-31.

Velupillai, S. 2011. Automatic classification of factuality levels. A case study on Swedish diagnoses and the impact of local context. En A. Moen, S. K. Andersen, J. Aarts y P. Hurlen (Eds.) *Proceedings of the XXIII International Conference of the European Federation for Medical Informatics. User Centred Networked Health Care*, Oslo: IOS Press, 559-563.

Vincze, V., G. Szarvas, R. Farkas, G. Móra y J. Csirik. 2008. The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9, 38-45.

Wonsever, D., A. Rosá y M. Malcuori. 2016. Factuality annotation and learning in Spanish texts. *Proceedings of Language Resources and Evaluation*, 2076-2080.

Wu, A. S., B. H. Do, J. Kim y D. Rubin. 2009. Evaluation of negation and uncertainty detection and its impact on precision and recall in search. *Journal of Digital Imaging*, 24(2): 234-242.

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 155-158 recibido 21-03-22018 revisado 13-04-2018 aceptado 03-05-2018

# Open Data for Public Administration: Exploitation and semantic organization of institutional web content

## *Datos Abiertos para la Administración Pública: Explotación y organización semántica del contenido web institucional*

**Paula Peña, Rocío Aznar, Rosa Montañés, Rafael del Hoyo**
Grupo de Big Data y Sistemas Cognitivos
ITAINNOVA (Instituto Tecnológico de Aragón)
C/ María de Luna, nº 7. 50018 Zaragoza
{ppena,raznar,rmontanes,rdelhoyo}@itainnova.es

**Abstract:** The project presented has been financed by Government of Aragon and is part of the 'Open Data' initiative promoted by that organization. Given the amount of unstructured information related to the Government of Aragon currently published on the Internet, with slightly or no standardization and decentralized, it emerges the need to gather it systematically to be offered to all interested collectives from a single access point in a public and structured way. Within this context, 'Aragon Open Data' project aims to collect, organize, store and maintain updated, Administration's web information by means of human language and semantic technologies. Firstly, crawling is performed over websites in order to retrieve textual data over which Natural Language Processing (NLP) and ontology-based techniques are applied. Thereafter, results are stored into NoSQL databases, allowing future open access and simple data exploitation. NLP techniques used in the project involve named-entities recognition and classification (NERC) and texts semantic classification and summarization.
**Keywords:** Open Data, ontologies, natural language processing, crawlers

**Resumen:** El proyecto presentado, financiado por el Gobierno de Aragón, se enmarca dentro de la iniciativa de 'Open Data' promovida por dicho organismo. Dada la cantidad de información no estructurada relacionada con el Gobierno de Aragón, publicada en Internet de forma no estandarizada y descentralizada, surge la necesidad de recopilarla sistemáticamente para ser ofrecida a los colectivos de interés desde un único punto de acceso, pública y estructuradamente. En este contexto el objetivo del proyecto 'Aragón Open Data'es extraer, organizar, almacenar y mantener actualizada la información web de la administración, mediante el uso de tecnologías semánticas y del lenguaje. Concretamente, se realiza un crawling exhaustivo de páginas web para extraer los datos textuales sobre los cuáles se aplican técnicas basadas en ontologías y de procesamiento de lenguaje natural (PLN). Finalmente se almacenan los resultados en bases de datos NoSQL, permitiendo su futura explotación de manera sencilla, abierta y transparente al ciudadano. Las técnicas de PLN utilizadas en el proyecto incluyen el reconocimiento y clasificación de entidades nombradas (NERC) y la clasificación semántica y resumen de textos.
**Palabras clave:** Open Data, ontologías, procesamiento del lenguaje natural, crawlers

## *1 Introduction*

Nowadays, Open Data is a worldwide movement whose philosophy aims to provide data openness and availability to citizens. Many countries have introduce an Open Data Initiative using government data[1]. Particularly, in Europe, Open Data has positioned itself as a focus of interest among policymakers for

[1] "Open Data Barometer Global Report". Available on https://opendatabarometer.org

ISSN 1135-5948. DOI 10.26342/2018-61-21

over a decade[2].

Aragón Open Data is a project framed by the agreement of July 17, 2012 of the Government of Aragon guided by the core idea of opening public data to the general public in order to be accessed, used and shared in a standard and valuable way. Its Internet portal `http://opendata.aragon.es` was publicly presented with the final purpose of creating economic value within the ICT sector through the reuse of public information, increasing Administration's transparency, promoting innovation, improving information systems of the Administration, adopting technical standards in the information society field and generating data interoperability between public sector websites.

Aragonese public administration presents a complex casuistry in their data generation processes that is reflected in the proliferation of a large number of websites, subdomains, portals and downloadable documents in heterogeneous formats under its root domain: `aragon.es`. This circumstances make it difficult to easily access and make use of the information by users and Government services as well, generating the popular sense of certain lack of transparency from the public administration.

In order to deal with these issues, a solution is required to allow data and institutional information that is currently dispersed, non-homogeneous, uncontrolled and non-exploitable, is converted into structured data that can be analyzed together, be accessible, be browsed by related concepts and be exploited and served to third parties (other institutional websites, media, developers or citizens). For this reason, an open data solution has been implemented. It retrieves all the institutional information published over the Internet by means of web crawling or *spider* techniques within the existing domains of the Government of Aragon. Data analysis and structuring is performed by NLP techniques, such as semantic classification, NERC or summarization, combined with the use of an ontology designed and implemented within the context of the project, this is, the Interoperable Information Scheme of Aragon (EI2A). Likewise, the data obtained throughout this process will be used to verify and, if necessary, enrich the operation of the ontology, leading to a continuous update of the system knowledge. Therefore, EI2A emerges with the main idea of generating a framework in which the open data and regional government data in general, can begin to be automated in a much deeper way.

## 2 *Proposed Approach*

In this section, the proposed approach followed to develop the system of capture and exploitation of the institutional web content is described. The functional design of the system is visually displayed in Figure 1. It is focused on three main modules:

1. Textual information retrieval from websites under `aragon.es` domain.
2. Natural language processing (NLP) techniques application on extracted data.
3. Results storage into NoSQL databases, conforming EI2A structure.

This process is executed periodically, allowing exploitation and query of updated results in real time.

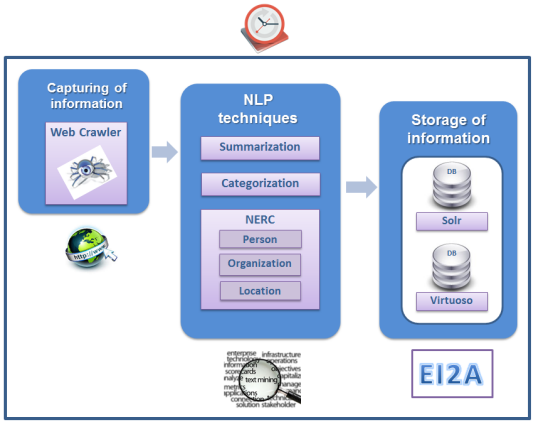


Figure 1: Functional Design

These functionalities have been implemented through *Moriarty*[3] (Peña et al., 2016), an ITAINNOVA's framework that allows the development of advanced software solutions for Big Data and Artificial Intelligence. In the following subsections a detailed explanation of each one of these three modules is provided.

[2] "The Re-use of Public Sector Information Regulations". Available on http://www.legislation.gov.uk/uksi/2015/1415/made

[3] "Moriarty". Information available on: http://www.ita.es/moriarty/

### 2.1 Information Retrieval

A set of information sources as websites, subdomains or portals, are considered the seed of this approach, from which a list of URLs is created with the depth of the analysis, maximum number of pages to analyze and number of crawling-threads desired. Afterwards, *web crawling* techniques are applied in order to extract all the institutional textual information possible from these sources.

Once information has been extracted, a cleaning task is performed applying customized meta-data removing rules, which results in a text prepared for the later application of NLP techniques. Elimination of headers, footnotes or indexes are some of the functionalities applied in this phase.

Since web information could change frequently and new pages may appear, web crawling is executed periodically, analyzing the new webs that appear or reprocessing the webs that have changed.

### 2.2 NLP techniques

Main NLP techniques used over the textual data, as shown in Figure 1, involve text summarization, thesaurus-based semantic classification and the named entities recognition and classification (NERC). Before applying these high level NLP techniques, it is necessary to preprocess text, performing some common task such as lowercase transformations, lemmatization or stopwords filtering.

These NLP techniques provide valuable results of general interest to the user, contributing to the performance of the project objective. Summary task offers a synthesis of the textual information with the most relevant sentences by means of graph-based ranking algorithms (Erkan and Radev, 2004), the NERC task identifies implicit information of the texts about the people, organizations and places that are named in them, thanks to the use of neural networks algorithms (Chiu and Nichols, 2015).

The information extracted from the application of NLP techniques and the storage of their results has followed a legal methodology. In particular, due to legal aspects, people extracted from NERC technique, who do not belong to the organizational chart of Government of Aragon, is anonymized in the summary task by asterisks.

### 2.3 Storage and structuring of information: *EI2A ontology*

Extracted knowledge of web textual content is stored in a structured and controlled manner, both into Solr database and Virtuoso through triplets according to the EI2A scheme for further exploitation.

Based on the philosophy of the Semantic Web, well-known ontologies, schemes and vocabularies endorsed by European directives (INSPIRE)[4] and International Consortium (W3C)[5] such as *The Organization Ontology*, *Simple Knowledge Organization System*, *RDF Schema*, *XML Schema*, *Dublin Core Metadata Terms*, *Schema.org*, *Friend Of A Friend*, *ISA Program Person Core Vocabulary*, *ISA Program Location Core Vocabulary*, *WGS84 Geo Positioning*, *The Event Ontology* and *OWL-Time* have been reused to model EI2A ontology. In the process of construction of the EI2A ontology methodological guides (Noy and McGuinness, 2005; Fernández-López, Gómez-Pérez, and Juristo, 1997) have been followed.

*The Organization Ontology* provides concepts and relations to support the representation of organizational structure (notion of an organization, decomposition into sub-organizations and units, purpose and classification of organizations); reporting structure (membership and reporting structure within an organization; roles, posts, and relationships between people and organizations); location information (sites or buldings, locations within sites) and organizational history (merger, neaming). A Government of Aragon domain-especific extension has been added to model the nature of an organic unit or office in Aragon (level of administration, public or private character, classification of a public entity, etc.). EI2A ontology model has been enriched with aspects and metadata of the DIR[6] and ENI[7].

*ISA Programme Person Core Vocabulary* and *ISA Programme Location Core Vocabulary* are reused for describing a natural person and any place in terms of its name, address or geometry. Other institutional data of common interest identified and modeled are focused on concepts related to documents, geolocation, territory, event, temporality and

[4] http://www.idee.es/europeo-inspire
[5] https://www.w3.org/
[6] http://administracionelectronica.gob.es/ctt/dir3
[7] http://administracionelectronica.gob.es/ctt/eni

URIs.

In order to apply EI2A scheme on real data, the ontology has been populated. On one hand, EI2A has been populated with information related to the organizational chart of Government of Aragón extracted from spreadsheets. In this way, semantic information is added to indicate the nature of a person's membership of an organization, that is to say, that a person belongs to a unit or department with a specific role in a valid time interval. On the other hand, motivated by the need of automatic way to extract, structure and standardize information from the huge amount of textual content available on the institutional websites, EI2A ontology has also been populated with new instances of concepts and relations provided by the NERC process. For example, information related to a recognized entity (person, organization and/or location) has been cited on a web that is classified under a categorization of themes of the Government of Aragon is specified semantically, in addition to add data related to the summary, the url and the date of capture of textual web content. A browser through the ontological model EI2A is illustrated in Figure 2.


Figure 2: Browser for EI2A elements

## 3 Conclusions and Future Work

Despite dealing with texts with a great diversity of domains and formats, the work carried out manages to integrate a generic system capable of fulfilling the expectations presented at the beginning. In addition, the results obtained are significantly satisfactory.

In this context, the viability of the proposed project has been verified and new aspects have been detected in which it is necessary to continue exploring. To this end, the need to deploy the solution over the public Aragonese infrastructures is raised, in order to develop on top of this system natural language recognition services with the challenge to understand questions asked by a user and know what needs to be answered (for example, semantic search engine and assistant BOT), to investigate new services in the line of extracting knowledge from the unstructured information that the Government of Aragon has, and to continue expanding and evolving the EI2A schema with the definition of new concepts and relationships based on the information processed as a consequence of the indicated actions.

## Acknowledgements

This work has been partly funded by the Department of Innovation, Research and University of the Government of Aragón in the context of the Aragón Open Data project. Special thanks to General Direction of Electronic Administration and the Information Society, Iciar Alonso and Julián Moyano for their collaboration. Also, this work has been partly financed by the FSE Operative Programme for Aragon (2014–2020).

## References

Chiu, J. P. and E. Nichols. 2015. Named entity recognition with bidirectional lstm-cnns. *arXiv preprint arXiv:1511.08308*.

Erkan, G. and D. R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Fernández-López, M., A. Gómez-Pérez, and N. Juristo. 1997. Methontology: from ontological art towards ontological engineering.

Noy, N. F. and D. L. McGuinness. 2005. Desarrollo de ontologías-101: guía para crear tu primera ontología. Available on http://ocw.uc3m.es/ingenieria-informatica/sistemas-avanzados-de-recuperacion-de-informacion/ejercicios.

Peña, P., R. del Hoyo, J. Vea-Murguía, V. Rodrigálvarez, J. I. Calvo, and J. M. Martín. 2016. Moriarty: Improving 'time to market' in big data and artificial intelligence applications. *International Journal of Design & Nature and Ecodynamics*, 11:230–238.

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 159-162 recibido 02-04-2018 revisado 13-04-2018 aceptado 04-05-2018

# Tecnologías inteligentes para la autogestión de la salud

## *Intelligent technologies for health self-management*

**Óscar Apolinario[1], José Medina-Moreira[1, 2], Katty Lagos-Ortiz[1, 2], Harry Luna-Aveiga[1], José Antonio García-Díaz[3], Rafael Valencia-García[3]**

[1] Facultad de Ciencias Matemáticas y Físicas, Universidad de Guayaquil, Cdla. Universitaria Salvador Allende, Guayaquil, Ecuador

[2] Facultad de Ciencias Agrarias, Universidad Agraria del Ecuador, Av. 25 de Julio, Guayaquil, Ecuador

[3] Facultad de Informática. Universidad de Murcia, Campus de Espinardo 30100 Murcia, España

{oscar.apolinarioa, jose.medinamo, katty.lagoso, harry.lunaa}@ug.edu.ec

{joseantonio.garcia8, valencia}@um.es

**Resumen:** El objetivo del proyecto "Tecnologías inteligentes para la autogestión de la salud" es desarrollar una plataforma inteligente para la autogestión de la salud de enfermedades crónicas como el asma y la obesidad. El sistema pone a disposición de los pacientes un servicio y control médico de primera mano, utilizando información que proporcione el mismo paciente y otra información textual proveniente de medios sociales como Twitter y otras fuentes oficiales que puedan ayudar a informar y apoyar a los usuarios según la enfermedad que padezcan. Esta información textual se procesará mediante tecnologías de análisis de sentimientos y clasificación para determinar si es una información relevante para cada paciente. Este proyecto está siendo desarrollado el grupo de investigación de informática médica de la Universidad de Guayaquil en colaboración con el grupo de investigación TECNOMOD de la Universidad de Murcia dentro del programa de ayudas propio de la Universidad de Guayaquil denominado FCI.
**Palabras clave:** Autogestión de la salud, ontologías, análisis de sentimientos, personalización de contenidos

**Abstract:** The objective of "Intelligent technologies for health self-management" project is the development of an intelligent platform for the self-management of chronic diseases such as asthma and obesity. With this platform, patients can enhance their health management though information provided by themselves and by extracting textual information from Twitter and official sources to inform and help them in the treatment of the disease they suffer. This textual information will be processed through sentiment analysis and classification technologies to determine which information is relevant for each patient. This project is being developed by the medical informatics research group of the University of Guayaquil in collaboration with the TECNOMOD research group of the University of Murcia within the program of grants owned by the University of Guayaquil called FCI.
**Keywords:** Health self-management, ontologies, sentiment analysis, content customization

## *1 Introducción*

Distintos estudios en la bibliografía hacen hincapié en que la auto-gestión de la salud reporta beneficios en cuanto a 1) una reducción de tiempos de hospitalización y 2) mejoras en la calidad de vida de los pacientes (Long et al. 1999). En la sociedad actual, la incidencia de enfermedades crónicas degenerativas como la diabetes o el asma han aumentado vertiginosamente debido a factores ambientales y al sedentarismo (Lazar, 2005).

Para mitigar los efectos de estas enfermedades, las actividades que se pueden hacer se resumen en: 1) diagnóstico, 2)

ISSN 1135-5948. DOI 10.26342/2018-61-22 

tratamiento, 3) prevención y 4) control. Estas actividades están íntimamente relacionadas. Por ejemplo, la falta de diagnósticos efectivos puede resultar en que las actividades de tratamiento se incrementen resultando en un impacto negativo en la salud de los pacientes y un incremento en el gasto sanitario.

Las enfermedades crónicas, como el asma o la diabetes, son aquellas enfermedades que tienen un tiempo de afección muy elevado. En el caso de enfermedades como la diabetes, los síntomas que permiten el diagnóstico de pacientes pueden ocurrir cuando esta se encuentra en un estado ya avanzado, por lo que es importante dedicar esfuerzos en las tareas de prevención y control, tanto a corto como a largo plazo.

Este proyecto nace con el objetivo de aprovechar tecnologías de monitorización, análisis de datos y de procesamiento del lenguaje natural para ayudar a los pacientes a prevenir y controlar algunas enfermedades crónicas como la diabetes, la obesidad o el asma, enfermedades que afectan de forma sistemática a la población mundial y en especial los países desarrollados y en vías de desarrollo. Con la consecución de los objetivos del proyecto también se pretende reducir los costes derivados al tratamiento de enfermedades en centros de salud, tanto públicos como privados, a través de la concienciación de que la salud debe ser controlada por el mismo paciente.

A nivel técnico, esta plataforma permitirá analizar parámetros de salud del paciente y sus tratamientos. Además, se emplearán técnicas de personalización de contenidos basadas en técnicas de reconocimiento de entidades y análisis de sentimientos. El objetivo final es proveer al paciente con la información necesaria sobre su enfermedad para que puedan tomar las mejores decisiones en cuanto a su salud.

## 2 *Estado actual del proyecto*

La plataforma global está formada por cinco módulos principales: 1) Módulo de extracción de información en redes sociales, 2) Repositorio de bases de conocimiento y ontologías, 3) Módulo de monitorización de parámetros y gestión de alertas, 4) Módulo de recomendaciones de salud y personalización de contenidos e 5) Interfaz web adaptativa (ver Figura 1). Por un lado, los pacientes pueden insertar sus parámetros de salud como la presión arterial, peso, ingesta de alimentos, medidas de glucosa, estado de ánimo y medicación, entre otros. El sistema entonces almacena todas las medidas y en base a un sistema de reglas determina si las medidas están dentro de la normalidad y expone recomendaciones de salud personalizadas por paciente. Por otro lado, expertos en salud podrán revisar los parámetros introducidos por los pacientes para verificar si las recomendaciones automáticas y alertas de los pacientes son correctas y para poder dar recomendaciones de salud personalizadas para cada paciente.

Además, los pacientes pueden recibir información textual relacionada con sus enfermedades, síntomas y medicaciones a partir de un proceso de extracción de información procedente de distintas fuentes de la red, tanto a partir de redes sociales, como de información de fuentes oficiales. Esta información se analiza para determinar qué información es la más relevante para el usuario. Además, se proporcionarán estadísticas sobre la información obtenida como número de comentarios relacionados con su perfil durante distintos periodos de tiempo como días, semanas o meses.

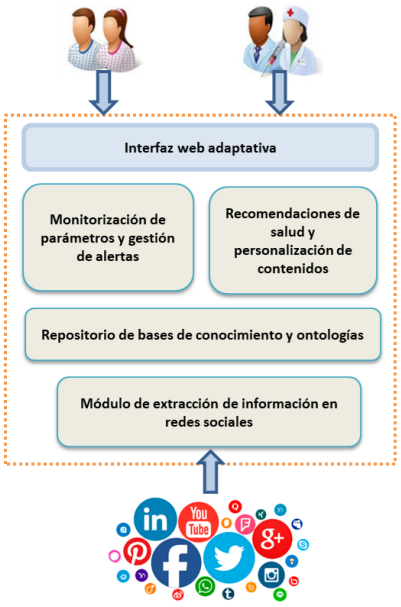


Figura 1: Arquitectura del sistema

A continuación, se describen brevemente cada uno de los módulos de la plataforma.

## 2.1 Módulo de extracción de información en redes sociales

Este módulo se encarga de extraer información textual de Twitter y otras cuentas oficiales relacionadas con las enfermedades que maneja la plataforma. Así, gracias a las ontologías y vocabularios contenidos dentro del repositorio de bases de conocimiento y ontologías, este módulo detecta distintos aspectos incluidos en esas ontologías como síntomas, causas, medicaciones de esas enfermedades. Además, un módulo de análisis de sentimientos basado en trabajos previos del grupo de investigación se utiliza para extraer la polaridad de los tuits y presentárselos de esta manera a los pacientes según la enfermedad que presenten (Salas-Zárate et al., 2017).

## 2.2 Repositorio de bases de conocimiento y ontologías

La plataforma incluye distintas ontologías relacionadas con las enfermedades crónicas que definen entre otras cosas los síntomas, las causas, los parámetros a medir y los medicamentos relacionados con las enfermedades. Para eso, al inicio del proyecto se realizó un estudio de las distintas ontologías sobre todo en el dominio de la diabetes y asma para crear una ontología propia que sirviese de base para el proyecto. Entre las que se seleccionaron para la diabetes se pueden destacar la DIAB (Vasant et al., 2015) y DMO (Rahimi et al., 2014) que se integraron en una nueva ontología y se localizó el contenido para el idioma español. La ontología resultante contiene 286 conceptos y 18 propiedades.

## 2.3 Módulo de monitorización de parámetros y gestión de alertas

Este módulo permite la monitorización de los parámetros relacionados con cada enfermedad. Por ejemplo, dentro de la diabetes se tienen en cuenta distintos parámetros como el nivel de glucosa, la toma de insulina, el pulso, peso y presión arterial del paciente entre otros.

Las distintas métricas son usadas como entrada para el módulo de gestión de alertas. Este módulo permite a los pacientes poder configurar un sistema de recordatorios y avisos que se disparan en cuanto se detecta una situación anómala o bien cuando es necesario recordar a los pacientes que deben de suministrar la medicación.

## 2.4 Módulo de recomendaciones de salud y personalización de contenidos

Este módulo se encarga de sugerir a los usuarios actividades que puedan ayudar a mejorar su calidad de vida. Para ello, se aplican técnicas de minería de datos sobre un conjunto de datos formado por los valores de los parámetros de salud, patrones de administración de medicamentos y hábitos deportivos en conjunto con actividades que han sido efectivas para otros pacientes en condiciones similares.

Por otra parte, la personalización de contenidos permite a los usuarios conocer información actualizada presente en la red. Para ello, se aplican técnicas de reconocimiento de entidades y análisis de sentimientos sobre textos extraídos de redes sociales y fuentes oficiales. La premisa para este módulo es que la información presente en la red, convenientemente explotada, puede ser útil a los pacientes para estar mejor informados sobre su salud. Además, puede servir de apoyo para conocer opiniones, comentarios y artículos sobre las enfermedades, síntomas y medicación (entre otros) propia de cada usuario.

En este módulo, cada usuario podrá acceder a un conjunto de información filtrada relevante relacionada y con las enfermedades, síntomas, medicación o estado de la enfermedad del paciente. La información subjetiva como opiniones y comentarios de Twitter, por ejemplo, se analiza mediante tecnologías de análisis de sentimientos para detectar la polaridad de si son comentarios positivos o negativos. Los pacientes podrán personalizar los contenidos a través de un sistema de filtros según la fuente, la enfermedad a la que alude el texto, la polaridad de la noticia, etc.

Este enfoque ya ha sido usado con éxito en otros proyectos como HealthMap (Freifeld, 2008) en donde se realiza un proceso de selección manual, junto con un sistema automático de minería de datos de noticias presentes en la red con objeto de monitorizar nuevos casos relacionados con las enfermedades infecciosas.

## 2.5 Interfaz web adaptativa

Este módulo se encarga de presentar a los usuarios la información recolectada desde los módulos anteriores de una manera comprensible y accesible. Los casos de uso identificados para la interfaz comprenden el registro de parámetros de salud, personalización

de alertas, registro de administración de medicamentos, recomendaciones de salud y personalización de contenido. Esta información se mostrará usando gráficas que de manera sencilla puedan resumir el estado del paciente.

Esta interfaz ha sido construida utilizando tecnologías de web adaptativa a partir del estándar HTML5. La elección de este tipo de interfaces es porque permiten representar la información ajustada en función del dispositivo que use el usuario, permitiendo adaptar el contenido y los formularios de registro que permitan a los pacientes acceder y registrar sus mediciones y actividades de forma rápida y eficaz. Además, al basarse en estándares web se está garantizando que la interfaz funcionará de manera correcta con dispositivos futuros.

Los profesionales de la salud podrán a su vez consultar la información y parámetros de salud de cada paciente de manera gráfica, así como realizar directamente recomendaciones de salud para cada paciente de manera manual.

## *3 Trabajo futuro*

Actualmente existe una primera versión funcional del prototipo de todos los módulos. La primera versión del módulo de recomendaciones de salud y personalización de contenidos es todavía provisional y actualmente filtra la información en base a los elementos encontrados en la ontología de enfermedades que están relacionados con el paciente. También se aplican técnicas de análisis de sentimientos para detectar la polaridad de esta información.

El próximo año está previsto centrarse en el desarrollo de las tecnologías de este último módulo de recomendación de contenidos en el cual se desarrollarán nuevos métodos para determinar la importancia del contenido textual recomendado y también análisis de sentimientos basados en aspectos.

Otra mejora de la plataforma sería la inclusión de tecnologías de resúmenes de texto como las presentadas en Esteban y Lloret (2017) que permita obtener la información más relevante dentro de todo un documento.

Por último, se realizará un análisis de viabilidad enfocado a otro tipo de enfermedades infecciosas como el zika, dengué y malaria.

## *Agradecimientos*

Este trabajo ha sido financiado por la Universidad de Guayaquil dentro del proyecto "Tecnologías inteligentes para la autogestión de la salud" dentro de las ayudas FCI.

## *Bibliografía*

Esteban, A. y E. Lloret. 2017. Propuesta y desarrollo de una aproximación de generación de resúmenes abstractivos multigénero. *Procesamiento del Lenguaje Natural*, 58:53-60.

Freifeld, C. C., K. D. Mandl, B. Y. Reis y J. S. Brownstein. 2008. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *Journal of the American Medical Informatics Association*, 15(2):150-157.

Lazar, M. A. 2005. How obesity causes diabetes: not a tall tale. *Science*, 307(5708):373-375.

Lorig, K. R., D. S. Sobel, A. L. Stewart, B. W. Brown Jr, A. Bandura, P. Ritter, V. M. Gonzalez, D. D. Laurent y H. R. Holman. 1999. Evidence suggesting that a chronic disease self-management program can improve health status while reducing hospitalization: a randomized trial. *Medical care*, 37(1):5-14.

Rahimi, A., S. T. Liaw, J. Taggart, P. Ray y H. Yu. 2014. Validating an ontology-based algorithm to identify patients with type 2 diabetes mellitus in electronic health records. *International journal of medical informatics*, 83(10):768-778.

Salas-Zárate, M. D. P., J. Medina-Moreira, K. Lagos-Ortiz, H. Luna-Aveiga, M. A. Rodríguez-García, y R. Valencia-García. 2017. Sentiment analysis on tweets about diabetes: an aspect-level approach. *Computational and mathematical methods in medicine*, 2017:Article ID 5140631, 1-9.

Vasant, D., F. Neff, P. Gormanns, N. Conte, A. Fritsche, H. Staiger y P. Robinson. 2015. DIAB: an ontology of type 2 diabetes stages and associated phenotypes. En *Proceedings of Phenotype Day at ISMB 2015*, páginas 24-27.

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 163-166 recibido 04-04-2018 revisado 26-04-2018 aceptado 04-05-2018

# TUNER: Multifaceted Domain Adaptation for Advanced Textual Semantic Processing. First Results Available

## *TUNER: Adaptación a dominio para el procesamiento semántico avanzado. Primeros resultados disponibles*

**Rodrigo Agerri[1], Núria Bel[2], German Rigau[1], Horacio Saggion[2]**
[1] University of the Basque Country (UPV/EHU)
[2] Pompeu Fabra University (UPF)
rodrigo.agerri@ehu.es; nuria.bel@upf.edu;
german.rigau@ehu.es; horacio.saggion@upf.edu

**Abstract:** The TUNER coordinated project (2016-2018) has focused on the development of domain adaptation technologies that reduce the cost of creating linguistic resources to develop systems in different languages and for different domains and genres. In this article we present the demonstrators, prototypes and resources that are already available project results.
**Keywords:** Language Resources, domain adaptation, semantic processing

**Resumen:** El proyecto coordinado TUNER (2016-2018) se ha centrado en el desarrollo de tecnologías de adaptación a dominio que permitan reducir el coste de la creación de recursos lingüísticos para desarrollar sistemas en diferentes lenguas y para diferentes dominios y géneros. En este artículo presentamos los demostradores, prototipos y recursos que son resultados del proyecto ya disponibles.
**Palabras clave:** Recursos lingüísticos, adaptación a dominio, procesamiento semántico

## *1 Introduction*

TUNER project (http://ixa2.si.ehu.es/tuner/) started in 2016 as a coordinated project funded by the Spanish Ministry of Economy, Industry and Competitiveness. The project is coordinated by IXA Group[1] from University of the Basque Country (UPV/EHU) and the participant groups are: TALN Natural Language Research Group[2] and IULATERM-Technologies of Language Resources Group[3] from Pompeu Fabra University; TALG[4] Technologies and Applications of Galician Language, from University of Vigo; GRIAL[5], Linguistic Applications Inter-University Research Group, from University of Barcelona, and Open University of Catalonia; and Natural Language Processing & Information Retrieval Research Group[6] at National Distance Learning University (UNED). Researchers from Elhuyar Fundazioa[7] and Vicometch[8] are also participating in TUNER.

The project has focused on the research and development of domain adaptation technologies. These technologies were used for improving the Natural Language Processing (NLP) tools developed for the different languages in Spain and for different domains, in particular, health and tourism. These technologies have been applied to different tasks for which demonstrations have been built. In Section 3, some of these demonstrators are described, and in Section 4, the datasets

[1] http://ixa.eus/
[2] https://www.upf.edu/web/taln
[3] https://www.upf.edu/web/iulaterm/trl
[4] http://sli.uvigo.gal/
[5] http://grial.uab.es
[6] http://nlp.uned.es/web-nlp/
[7] https://www.elhuyar.eus
[8] http://www.vicomtech.org

ISSN 1135-5948. DOI 10.26342/2018-61-23

delivered by the project are described too to promote their reutilization.

## *2 TUNER objectives*

In the general framework of text analytics and big data industry, the necessity of handling texts which are from different genres, domains and in different languages represents a very expensive investment in creating particular manually enriched datasets to train Natural Language Processing (NLP) methods. This dependency on data is challenging the expansion of companies to emerging business opportunities. TUNER addressed this problem through research and development of domain adaptation techniques to apply them to NLP technologies, to induce the information required to build different NLP tools for different tasks for which there are scarce or no data available.

TUNER research objectives were:

(i) Multilingual Text Processing. Providing language resources and tools for basic processing of textual data in the languages covered by the project: English, Spanish, Catalan, Basque and Galician.

(ii) Knowledge Acquisition and Integration. The improvement of a broad-coverage knowledge base integrated into the Multilingual Central Repository (MCR), linked to the latest versions of the English WordNet, and adapted to specific domains.

(iii) Reasoning and Inferencing. Development of inference and semantic reasoning engines using the knowledge integrated into the MCR.

(iv) Advanced Cross-Lingual Semantic Processing. Providing effective methodologies for developing robust and advanced semantic processing systems suitable to be adapted to other domains and languages.

## *3 Demonstrations and Prototypes*

### 3.1 UMLS Mapper

This demonstrator, available at http://demos-v2.vicomtech.org/umlsmapper, shows the capabilities of a prototype (Perez, Cuadros and Rigau, 2018) to identify medical terms in free text in Spanish and map them to concepts of the UMLS medical terminology compendium[9]. The mapping is based on information retrieval techniques, by indexing these terminologies with Apache Lucene. Acronyms and abbreviations (Montoya, 2017) and IXA-pipes (Agerri, Bermudez and Rigau, 2014) were used to process the input texts. The disambiguation of terms with more than one possible mapping is based on UKB tool (Agirre and Soroa, 2009).

As for the demonstrator, at the main page you can choose between three given clinical texts or enter any free text to analyze. Once the analysis process is finished, two columns are displayed: on the right, the text sent is marked in several colors that highlight the recognized medical terms; on the left, the recognized terms are grouped by general medical concepts. By clicking on any of these terms, a third column shows additional information: hyponyms and hypernyms, synonymous expressions, a definition, etc.

### 3.2 AsisTerm

AsisTerm (http://scientmin.taln.upf.edu/scielo) is a prototype aimed to facilitate the understanding of complex biomedical terms in short texts --currently, abstracts of articles in English and Spanish from the ScieLO parallel corpus[10]-- by means of the identification and annotation of UMLS concepts and the automatic expansion of their definitions. Several resources, tools and methods have been developed to support and/or automate the semantic indexing of biomedical texts, but the vast majority of them are only targeted at English. When parallel corpora are available, as is the case of the ScieLO abstracts, these tools can be exploited in order to produce annotations in English that can then be transferred to texts in other languages. AsisTerm, in particular, makes use of the Becas online service[11] in order to automatically annotate the ScieLO English abstracts with UMLS concepts. Once the relevant concepts are identified, the spans of texts in the Spanish abstracts that best match each of them have to be determined. This is done by computing the similarity between candidate terms obtained from the abstract in Spanish and the Spanish lexicalizations of each retrieved UMLS concept. Once the correct location for each annotation is found, the definition of the corresponding concept is expanded by means of the MedlinePlus Connect service[12]. The system includes a web

[9] https://www.nlm.nih.gov/research/umls

[10] http://www.scielo.org

[11] http://bioinformatics.ua.pt/becas/

[12] https://medlineplus.gov/connect

interface to search and retrieve the ScieLO abstracts in English and Spanish where the annotations and definitions of biomedical terms can be visualized.

### 3.3 Lingaliza

Lingaliza (http://sli.uvigo.gal/lingaliza/) is a web interface designed to test the full new set of NLP tools provided by the IXA pipes for the Galician language. This set of tools includes a rule-based tokenizer and sentence segmenter, a statistical lemmatizer and POS tagger, a state-of-the-art NER tagger a wikification tool based on DBpedia Spotlight, a NED tool based on DBpedia Spotlight and a UKB graph-based tool for word sense disambiguation.

### 3.4 Analhitza

Analhitza is a web interface for easy access to the NLP tools provided by the ixaKat for Basque and IXA pipes for Spanish languages, among others (Otegi et al., 2017). Analthitza (http://ixa2.si.ehu.es/clarink/analhitza.php) uses the basic IXA NLP tools, including segmentation, lemmatization and POS and NER tagging.

### 3.5 Central Unit Detector for Scientific Texts

With the aim of building an automatic discourse analyzer, we developed a tool consisting of a discourse segmenter and an automatic central discourse unit detector. In the future, we will link the segments with discourse relations, on the top of the central unit (CU). The CU detector, for Basque (Bengoetxea, Atutxa and Iruskieta, 2017) (http://ixa2.si.ehu.es/CU-detector/) and for Spanish (Bengoetxea and Iruskieta, 2018) detects the most salient node of a text (http://ixa2.si.ehu.es/clarink/tools/ES-CU-detector/). First, the detector segments the text in discourse units and then it labels the most important text segment.

The CU detector for Basque was developed to handle scientific abstract texts from seven different specialized domains: medicine, health. life, terminology, science, economy and computer science. Scientific texts from psychology and linguistics domains were employed to develop the Spanish CU detector.

### 3.6 Navigating the SEPLN Anthology

As part of our work on scientific text mining, we have developed a tool to process the articles from the SEPLN journal effectively creating the *SEPLN Anthology*, a fully analyzed bi-lingual resource created from SEPLN publications (Saggion et al., 2017) available at http://scipub-taln.upf.edu/sepln/. Furthermore, we have also developed a Web-based information access platform which exploits the SEPLN Anthology documents to provide single-document and collection-based visualizations as a means to explore the rich generated contents.

### 3.7 PDFdigest: Extracting content from PDF

We have developed PDFdigest (Ferrés et al., 2018), a tool for PDF-to-XML content extraction specially designed to extract scientific articles' headings and logical structure (title, authors, abstract, etc.) and its textual content (http://taln.upf.edu/pdfdigest). PDFdigest has been used to extract metadata from the SEPLN journal to create the SEPLN Anthology (Saggion et al., 2017).

## *4 Resources*

### 4.1 Galnet

Galnet is an open wordnet for Galician, aligned with the interlingual index (ILI) generated from the English WordNet3.0, following the expand model for the creation of new wordnets, where the variants associated with the Princeton WordNet synsets are translated using different strategies. Galnet can be searched via a web interface (http://sli.uvigo.gal/galnet/) and can be downloaded in RDF and LMF formats from http://sli.uvigo.gal/download/SLI_Galnet/.

Galnet is part of the Multilingual Central Repository (http://adimen.si.ehu.es/web/MCR/) a database that currently integrates wordnets from six different languages (English, Spanish, Catalan, Galician, Basque and Portuguese) using WordNet 3.0 as ILI and where each *synset* is classified under the WordNet Domains hierarchy, the SUMO ontology and the Top Concept Ontology. The specific interface designed to query Galnet extends the MCR functionalities by providing different navigation types through domain hierarchies and ontologies, allowing an interactive tree-based visualization of *synsets* by their semantic relations, including a variety of terminology-oriented functionalities.

### 4.2 SensoGal Corpus

The SensoGal Corpus is an English-Galician parallel corpus semantically annotated with respect to WordNet 3.0 and aligned at the sentence and word level with the English SemCor corpus. The SensoGal Corpus is based on the Galician translation of the 186 fully tagged texts included in the original English Princeton SemCor, prioritizing the translation of the texts available in the MultiSemCor Corpus. The resulting parallel corpus can already be consulted through a web query interface http://sli.uvigo.gal/SensoGal/. The Galician corpus sentences in SensoGal are used as lexicographic examples of usage of the variants in the Galnet interface.

### 4.3 IULA Spanish Clinical Record Corpus

The IULA Spanish Clinical Record Corpus (IULA-SCRC) is a corpus of 3,194 sentences extracted from anonymized clinical records and manually annotated with negation markers and their scope. Annotation guidelines are documented at Marimon, Vivaldi and Bel (2017) and the corpus access is http://eines.iula.upf.edu/brat/#/NegationOnCR_IULA.

## *Acknowledgments*

TUNER coordinated project was funded by Spanish Ministry of Economy, Industry and Competitiveness (TIN2015-65308-C5, MINECO/FEDER, UE).

## *References*

Agerri, R., J. Bermudez, and G. Rigau (2014). IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), pp. 3823-3828.

Agerri, R.; X. Gómez Guinovart; G. Rigau; M. A. Solla Portela (2018). Developing New Linguistic Resources and Tools for the Galician Language. Proceedings of the 11th Language Resources and Evaluation Conference (LREC'18).

Agirre, E. and A. Soroa. (2009). Personalizing PageRank for Word Sense Disambiguation. In Proceedings of the 12th Conference of the European Chapter of the ACL, pp. 33-41.

Bengoetxea, K.; A. Atutxa; 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.

Bengoetxea, K.; M. Iruskieta (2018). A Supervised Central Unit Detector for Spanish. Procesamiento del Lenguaje Natural, 60, 29-36.

Ferrès, D.; Saggion, H.; Rozano, F.; Bravo, À. (2018). PDFdigest: an Adaptable Layout-Aware PDF-to-XML Textual Content Extractor for Scientific Articles. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Gómez Guinovart, X.; M. A. Solla Portela (2018). Building the Galician wordnet: methods and applications. Language Resources and Evaluation, 52:1, 317-339.

Marimon, M., J. Vivaldi, and N. Bel. (2017). Annotation of negation in the IULA Spanish clinical record corpus. In Proceedings of the Workshop SemBEaR 2017. ACL. p. 43-52.

Montoya, I. (2017). Análisis, normalización, enriquecimiento y codificación de historia clínica electrónica (HCE). Tesis del Máster Universitario Konputazio Ingeniaritza eta Sistema Adimentsuak, (UPV/EHU).

Otegi, A.; O. Imaz; A. Díaz de Ilarraza; M. Iruskieta; L. Uria (2017). ANALHITZA: a tool to extract linguistic information from large corpora in Humanities research. Procesamiento del Lenguaje Natural 58: 77-84.

Perez, N., M. Cuadros y G. Rigau (2018). Biomedical term normalization of EHRs with UMLS. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Saggion, H.; F. Ronzano; P. Accuosto; D. Ferrés (2017). MultiScien: a Bi-Lingual Natural Language Processing System for Mining and Enrichment of Scientific Collections. BIRNDL@SIGIR (1) 2017: 26-40

Solla Portela, M. A. and X. Gómez Guinovart (2017). Diseño y elaboración del corpus SemCor del gallego anotado semánticamente con WordNet 3.0. Procesamiento del Lenguaje Natural, 59, 137-140 (ISSN 1135-5948

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 167-170 recibido 03-04-2018 revisado 26-04-2018 aceptado 04-05-2018

# EMPATHIC: Empathic, Expressive, Advanced Virtual Coach to Improve Independent Healthy-Life-Years of the Elderly

## *EMPATHIC: Coach virtual empático, expresivo y avanzado para mejorar el bienestar de las personas de edad avanzada sanas e independientes*

**Asier López Zorrilla, Mikel de Velasco Vázquez, Jon Irastorza Manso**
**Javier Mikel Olaso Fernández**, **Raquel Justo Blanco**, **María Inés Torres Barañano**
Universidad del País Vasco (UPV/EHU)
{asier.lopezz, mikel.develasco, jon.irastorza,
javiermikel.olaso, raquel.justo, manes.torres}@ehu.eus

**Abstract:** The EMPATHIC project will research, innovate, explore and validate new paradigms and platforms, laying the foundation for future generations of Personalised Virtual Coaches to assist elderly people living independently at and around their home. Innovative multimodal face analytics, adaptive spoken dialogue systems and natural language interfaces are part of what the project will research and innovate, in order to help dependent aging persons and their carers. The project will use remote non-intrusive technologies to extract physiological markers of emotional states in real-time for online adaptive responses of the coach, and advance holistic modelling of behavioural, computational, physical and social aspects of a personalised expressive virtual coach. It will develop causal models of coach-user interactional exchanges that engage elders in emotionally believable interactions keeping off loneliness, sustaining health status, enhancing quality of life and simplifying access to future telecare services
**Keywords:** Multimodal dialogue systems, virtual coach

**Resumen:** El proyecto EMPATHIC tiene como misión investigar, explorar, innovar y validar nuevos paradigmas y plataformas, sentando las bases para las futuras generaciones de *coach* virtuales personalizados para ayudar a personas de avanzada edad que viven de forma independiente en su hogar. Además, EMPATHIC investigará e innovará con el objetivo de ayudar a las personas de avanzada edad dependientes y a sus cuidadores mediante análisis facial multimodal, sistemas de diálogo adaptables e interfaces de lenguaje natural. El proyecto utilizará tecnologías remotas no intrusivas para extraer marcadores fisiológicos de estados emocionales en tiempo real, las cuales influenciarán el comportamiento del coach. También se modelarán los aspectos computacionales, físicos y sociales de los coach virtuales expresivos desde un punto de vista holístico. Finalmente, se desarrollarán modelos que permitan la interacción entre el coach y el usuario, buscando así involucrar a éstos en interacciones que contribuyan a evitar la soledad, mantener el estado de salud, mejorar la calidad de vida y simplificar el acceso a futuros servicios de teleasistencia.
**Palabras clave:** Sistemas de diálogo multimodal, asistente virtual

ISSN 1135-5948. DOI 10.26342/2018-61-24 

## 1 Project Founding and Consortium

The EMPATHIC project has been founded by the European Commission H2020-SC1-2017-RIA grant number 769872: *"Empathic, Expressive, Advanced Virtual Coach to Improve Independent Healthy-Life-Years of the Elderly"*.

The consortium brings together 10 partners from 7 EU and associated countries (Norway and Israel). Among these partners, we find 4 universities or research centres, 3 healthcare centres related to institutions, and 3 companies. The organisations involved are:

- Universidad del País Vasco UPV/EHU (coordinator).
- OSATEK S.A.
- Oslo University Hospital.
- e-Seniors Association.
- Tunstall Healthcare (UK) Ltd.
- Technion - Israel Institute of Technology.
- Intelligent Voice Ltd.
- Acapela Group S.A.
- Institut Mines-Télécom.
- Seconda Università degli Studi di Napoli.

## 2 Introduction

Without undermining the important degree of development that public healthcare services in Europe have achieved, in terms of coverage and intensity, much care for people with limitations in autonomy is still provided in the private sector of the family, i.e. informal care. According to the Survey of Health, Age and Retirement in Europe (SHARE, 2002 - 2004), more than seven out of ten dependent elderly people in Italy, Germany, and Sweden receive informal care exclusively. In Spain, Austria and Holland it is more than half. In Belgium, Denmark and France, despite more coverage of formal services, more than one third receive such informal care. It must be recognised that informal care for the family is unpaid work that provides economic savings to welfare systems. However, the predictions point to less availability of informal caregivers in the future. This situation has led European countries to become interested in policies to support the informal care network, and caregiver support programs have become a priority area of action in the Union.

Studies suggest that attention to the lifestyle of the elderly can help them to maintain independent life (Willcox, Scapagnini, and Willcox, 2014a) (Davies, 2011). However, elder psychological obstacle, lack of knowledge and interpersonal and structural obstacles make it difficult for them. Therefore, besides promoting a healthy lifestyle, attention should be paid to the internal and external difficulties of the elderly through facilities and arrangement of activities. Here, virtual coaching is a very interesting solution (Ding et al., 2010) (Prescott et al., 2012) (Cavanagh and Millings, 2013).

*Virtual Coach* refers to a coaching program or device aiming to guide users through tasks for the purpose of prompting positive behaviour or assisting with learning new skills. Virtual coaches monitor how the user performs activities, provides situational awareness and gives feedback and encouragement matched to their cognitive state and circumstances at the same time. Further, a virtual coach matches its level of support to the user as his/her abilities change, and so the user can upload new options to the virtual coach as desired, and can define new well-being goals without even an office visit.

EMPATHIC will offer a personalised virtual coaching program in three languages (Spanish, French and Norwegian), ranging from low to high intensity levels. Such increased physical activity will help to decrease stress and depressive symptoms, and increase satisfaction with body function. An intervention to improve fruit and vegetable consumption based on the Dietary Guidelines for elders, and a review of practices, will also be applied, since proper nutrition plays a vital role in the health and wellbeing of older adults (Knoops et al., 2004) (Willcox, Scapagnini, and Willcox, 2014)b). EMPATHIC will also offer coaching to focus on brief behavioural techniques like social activation and pleasant event planning.

EMPATHIC will cater to healthy people of 65 years or more, and virtually coach them in order to help increase the years of independent active living and to improve health and slower deterioration. Contact with nature will be encouraged. Recommendations will include also healthy food (and shopping),

having enough sun exposure, vegetables and protein intakes. All counselling will be given in a positive, optimistic manner. The device will include also social activities information, like films, operas, theatres, and conferences in their city.

To this end, this project will develop innovative multimodal face and speech analytics, adaptive spoken dialogue systems, intelligent computational models and natural human-computer interfaces, resulting in an emotionally-expressive virtual coach, designed to help aging users and their carers. Building upon neuroscience research, the project will use unobtrusive remote technologies to extract physiological markers of emotional states in real-time. The virtual coach will monitor facial cues and speech style that underpin the user's basic neural function, and will formulate online adaptive responses, facilitating interaction through mimicking, in turn promoting empathy and support with the user.

## *3 EMPATHIC objectives*

### 3.1 Multidisciplinary research

EMPATHIC proposes multidisciplinary research and development, involving:

- Geriatrician, Neuroscientist, Psychiatric, health and social work specialists to implement the individual coaching goals.
- Psychologist, Neuroscientist and Computer Science experts for detection and identification of the emotional status of the user.
- Engineers and Computer Scientists in speech and language technologies, biometrics, image analysis, and machine learning.
- Telecare services, a senior association and a hospital interested in testing and validating EMPATHIC.
- Companies interested in providing and developing technology for the project and commercialising the products and derived services.

### 3.2 Main objectives

The six main objectives of this multidisciplinary research are listed next:

- Design a virtual coach, to engage the healthy-senior user and reach preset benefits, measured through project-defined metrics, to enhance well-being through awareness of personal physical status, by improving diet and nutritional habits, by developing more physical exercise and by social activity.
- Involve end-users and to reach a degree of fit to their personalised needs and requirements, derived by the coach, which will enhance their well-being.
- Supply the coach with non-intrusive, privacy-preserving, empathic, expressive interaction technologies.
- Validate the coach efficiency and effectiveness across 3 distinct European societies (Norway, Spain, and France), with 200 to 250 subjects who will be involved from the start.
- Evaluate/validate the effectiveness of EMPATHIC designs against relevant user's personalised acceptance and affordance criteria (such as the ability to adapt to users' underlying mood) assessed through well defined Key Performance Indicators.
- Drive the developed methodology and tools to industry acceptance and open-source access identifying appropriate evaluation criteria to improve the "specification-capture-design implementation" software engineering process of implementing socially-centred ICT products.

#### 3.2.1 Scientific goals and research actions

These six objectives will be accomplished through the following sets of goals and research/development actions:

- Implement health-coach goals and actions through an intelligent computational system, intelligent coach and spoken dialogue system adapted to users' intentions, emotions and context.
- Provide automatic personalised advice guidance (through the coach) having a direct impact in empowering elder users into a wide of advanced ICT keeping improving their quality of life and level

their independent independency living status of the people as the age.

- Identifying non-intrusive technologies to detect the individual's emotional and health status.
- Provide the virtual coach with a natural, empathic, personalised and expressive communication model.

#### 3.2.2 Technological goals and actions

- Develop a simulated virtual coach and acquire an initial corpus of dialogues through a Wizard-of-Oz to fulfil the initial end-users and data requirements of the Scientific Goals.
- Integrate and provide a proof-of-concept of the technology running on different devices.
- Validation through Field trials in the aforementioned three languages.

#### 3.2.3 Exploitation goals

- Define a plan for the exploitation of the results by the consortium as a whole and by particular partners.

## *Acknowledgements*

This work has been founded by the European Commission H2020-SC1-2017-RIA grant number 769872: "Empathic, Expressive, Advanced Virtual Coach to Improve Independent Healthy-Life-Years of the Elderly".

## *References*

Cavanagh, K. and A. Millings. 2013. Interpersonal computing: the role of the therapeutic relationship in e-mental health. *Contemporary Psychotherapy*, (43):197–206.

Davies, N. 2011. Promoting healthy ageing: the importance of lifestyle. *Nursing Standard (through 2013)*, 25(19):43.

Ding, D., H. Y. Liu, R. Cooper, R. A. Cooper, A. Smailagic, and D. Siewiorek. 2010. Virtual coach technology for supporting self-care. *Physical medicine and rehabilitation clinics of North America*, (21(1)):179–194.

Knoops, K. T., L. C. de Groot, D. Kromhout, A. E. Perrin, O. Moreiras-Varela, A. Menotti, , and W. A. Van Staveren. 2004. Mediterranean diet, lifestyle factors, and 10-year mortality in elderly european men and women: the hale project. *Jama*, (292(12)):1433–1439.

Prescott, T., T. Epton, V. Evers, K. McKee, M. Hawley, T. Webb, D. Benyon, S. Conran, R. Strand, M. Buning, P. Verschure, P. Dario, and T. Group. 2012. Robot companions for citizens: Roadmapping the potential for future robots in empowering older people. *In BRAID (Bridging Research in Ageing and ICT Development) Final Conference*.

SHARE. 2002 - 2004. Survey of Health, Ageing and Retirement in Europe. http://www.share-project.org/.

Willcox, D. C., G. Scapagnini, and B. J. Willcox. 2014a. Healthy aging diets other than the mediterranean: a focus on the okinawan diet. *Mechanisms of ageing and development*, 136:148–162.

Willcox, D. C., G. Scapagnini, and B. J. Willcox. 2014)b. Healthy aging diets other than the mediterranean: a focus on the okinawan diet. *Mechanisms of ageing and development*, (136):148–162.

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 171-174 recibido 04-04-2018 revisado 26-04-2018 aceptado 04-05-2018

# enetCollect: A New European Network for combining Language Learning with Crowdsourcing Techniques

## *enetCollect: Una nueva red europea para el aprendizaje de idiomas y el crowdsourcing*

**Rodrigo Agerri[*], Montse Maritxalar[*], Verena Lyding[**], Lionel Nicolas[**]**
[*]IXA NLP Group, University of the Basque Country UPV/EHU
[**]Eurac Research Bolzano, Italy
{rodrigo.agerri,montse.maritxalar}@ehu.eus,{verena.lyding,lionel.nicolas}@eurac.edu

**Abstract:** We present enetCollect, a large European COST action network set up with the aim of promoting a research trend combining the well-established domain of Language Learning with recent and successful crowdsourcing approaches. More specifically, the challenge of enetCollect is to foster the language skills of all citizens regardless of their backgrounds by enhancing the production of language learning material using Crowdsourcing techniques. In order to do so, the action will create a balanced interdisciplinary community of active stakeholders related to content-creation, content-usage, and Learning/Content Management Systems to create a theoretical framework for achieving a shared understanding of Language Learning and Crowdsourcing. This will allow to unlock the crowdsourcing potential available for language learning and to facilitate the development of prototypical experiments for the production of language learning material, such as lesson or exercise content. These activities would potentially benefit a wide range of users and languages.
**Keywords:** Crowdsourcing, Language Learning, Language Resources, COST Action

**Resumen:** En este artículo presentamos enetCollect, una extensa acción europea COST diseñada con el objetivo de promover una nueva línea de investigación que combine el dominio del aprendizaje de idiomas con recientes y exitosos enfoques basados en crowdsourcing. Más específicamente, el reto de enetCollect es fomentar el aprendizaje de idiomas para toda la ciudadanía europea mediante la mejora en la producción de materiales para el aprendizaje de idiomas usando técnicas de crowdsourcing. Para ello, la acción creará una comunidad interdisciplinar de agentes activos relacionados con la creación, uso y gestión de contenidos para el aprendizaje de idiomas que permita generar un marco teórico común en el cual investigar sobre el uso de crowdsourcing para la generación de contenidos y tecnología relacionada con el aprendizaje de idiomas. La idea es liberar el potencial de usar crowdsourcing para el aprendizaje de idiomas y facilitar el desarrollo de experimentos y prototipos para la generación de materiales de aprendizaje, tales como ejercicios, lecciones, etc. Estas actividades beneficiarán a la gran mayoría de las personas en proceso de aprender un nuevo idioma.
**Palabras clave:** Crowdsourcing, Aprendizaje de idiomas, Recursos Lingüísticos, acción COST

### *1 Introduction*

enetCollect[1] is a COST Action supported by the EU Framework Programme Horizon 2020. COST Actions are European networking initiatives, which aim at creating new research communities around emerging research subjects with ground-breaking potential. The essence of a COST Action is the generation of a durable network of participants that meet regularly and exchange ideas, approaches and methods, and aim at building new research cooperations. COST funding covers expenses related to travel, meeting organization, and exchange of re-

[1] `https://enetcollect.net`

ISSN 1135-5948. DOI 10.26342/2018-61-25

searchers, while it purposefully does not fund research work in itself. Nonetheless, COST Actions have an excellent record of building consortia and follow up projects that do fund personnel. At this moment, enetCollect includes 36 European countries in the Management Committee. enetCollect started in March 2017 and it will run until March 2021. The authors of the present paper are the Management Committee members representing Spain.

The main challenge of enetCollect, the European Network for Combining Language Learning with Crowdsourcing Techniques is to foster the language skills of all citizens regardless of their backgrounds by enhancing the production of language learning material using Crowdsourcing techniques. Specifically, the Action aims at enhancing the production of learning material combining the well-established domain of language learning with recent and successful crowdsourcing approaches. While primarily focusing on Language Learning, EnetCollect also involves a variety of research players working on language-related topics and having tedious and/or complex tasks to perform that may be approached by crowdsourcing (e.g. Language Resources creation).

enetCollect will research both implicit and crowdsourcing approaches. Briefly, explicit crowdsourcing usually refers to the fact that the crowd intentionally participates in the crowdsourcing task whereas in an implicit crowdsourcing task the crowd is not necessarily aware of the fact that the results of its activity will be used for other, not explicitly explained, objective. The action is also interested in researching issues such as user-orientation and usability of technological applications for language learning driven by ethical, legal and commercial dimensions of developing such technology.

The underlying capacity of crowdsourcing to achieve ground-breaking results has been proven in several impressive ways. For example, Wikipedia[2] completely redefined the well-established domain of encyclopedias while reCAPTCHA[3] tackled the highly laborious task of manually transcribing vast amounts of text by obtaining an unparalleled and continuous workforce from the crowd. In enetCollect, similar approaches will be developed or adapted to facilitate the creation of language learning materials and language-related datasets.

## *2 Objectives*

COST Actions distinguish objectives related to the creation of knowledge, termed Research Coordination (RC) Objectives, and objectives related to creating and empowering a community, termed Capacity-Building (CB) Objectives.

Regarding the RC objectives, enetCollect aims to review the state-of-the-art in order to gather and compile an overview of relevant approaches and techniques for crowdsourcing in order to obtain ashared understanding by creating a theoretical framework defining its terminology, key concepts, objectives and opportunities.

With respect to the CB objectives, the Action will create a community of active stakeholders and communication means allowing the easy exchange of information for, among other things, pursuing research experimentations and targeting new funded initiatives.

In pursuing these objectives, it is expected to create the following short and long term impacts: In the short term, the Action will build a balanced interdisciplinary community of experts that will initiate the R&I trend by creating a theoretical framework and evaluation data to complement it. Current members mostly come from the areas of Crowdsourcing, Computer Assisted Language Learning, Natural Language Processing, E-Lexicography, Learner Corpora, Corpus Linguistics and Learning Management Systems. In the short-to-long term, major impacts will consist of the enhanced creation of language learning material and of language-related data.

With regards to language learning material, the Action will foster the continuous creation and improvement of materials. This implies the participation of language learning participants, which will build an enormous crowdsourcing potential. Achievements even remotely comparable to the ones of Zooniverse[4] would significantly impact the domain. As a side effect, enetCollect will help balancing the coverage across languages by benefiting also less-resourced languages.

[2] `https://www.wikipedia.org/`

[3] `https://www.google.com/recaptcha`

[4] `https://www.zooniverse.org/`

## 3 Structure

The Grant Holder of enetCollect is EURAC, the European Academy of Bolzano[5]. The Management Committee includes 36 European countries[6]. enetCollect is structured in five Working Groups (WG) and three supportive coordination groups.

**WG1: Explicit Crowdsourcing** Its objective is developing or adapting explicit crowdsourcing approaches. For example, WG1 will research the most effective ways to collaboratively devise lesson content (e.g. grammar) and to assess its effectiveness by observing how different samples of users confronted to accordingly generated content perform subsequently.

**WG2 Implicit Crowdsourcing** WG2 aims at developing or adapting implicit crowdsourcing approaches. For instance, WG2 will research ways to generate exercises from language resources (e.g. lexica) and to crowdsource manual validation of automatically generated new entries (e.g neologisms) by cross-matching the learners' answers for exercises generated from such new entries.

**WG3: User-oriented design strategies for a competitive solution.** WG3 will create design strategies fostering the user-orientation of a language learning solution and ensuring its capacity to attract and retain a crowd. For example, WG3 will study the relevance and attractiveness of learner profiling for vocabulary training.

**WG4 Technology-oriented specifications for a flexible and robust solution.** WG4 targets the creation of technical specifications to support the functional demands of WG1, WG2, and WG3. For example, WG4 will study technical solutions for the scalability of crowdsourcing methods.

**WG5: Application-oriented specifications for an ethical, legal and profitable solution.** It will devise the application-oriented part of the theoretical framework related to (1) ethical questions regarding the user involvement and data collection, (2) legal regulations, and (3) opportunities and models for commercialization.

**Dissemination, Exploitation and Outreach coordinations** are overarching groups supporting the WGs with transversal tasks and standardizing the ways such transversal tasks are tackled.

Note that in many cases this structure responds to operational reasons. Thus, the five WGs each have their individual challenging tasks to address but they are interdependent in some aspects. For example, the boundary between explicit and implicit crowdsourcing (WG1 and WG2) can be fairly fuzzy for several approaches. Also, any crowdsourcing is meaningless if there is no crowd to rely on (WG3), no scalable solution to implement it (WG4), and no appropriate ethical or legal contexts (WG5).

## 4 enetCollect and NLP

The enetCollect action can be interesting for the Natural Language Processing (NLP) community to explore new avenues for crowdsourcing language resources (LR) which may be appropriate for language learning but also for domain specific learning (Science, History, etc). For example, implicit crowdsourcing could be done by means of simple language games where learners have to complete some challenges in order to improve their levels on the game. Similarly, explicit crowdsourcing could be implemented in collaborative applications where learners of different language levels or language communities help to each other by means of a peer-learning approach. Then, learners could tag or interpret the answers of lower level colleague mates as a peer collaboration, acting somehow as expert taggers of the data during the learning process.

enetCollect member statistics show that out of the five WGs, WG2 is the WG most followed by NLP-oriented Action members up to now. However, as there is a strong collaboration between WG1 and WG2, it is viable to organize NLP oriented Crowdsourcing challenges from both perspectives. This way language resources could be created to be used on learning approaches based on NLP.

## 5 Related Work

The online language learning platform Duolingo[7] follows a logic that is similar in many points to the one of enetCollect. It offers free language learning services for numer-

[5] http://www.eurac.edu

[6] http://www.cost.eu/COST_Actions/ca/CA16105

[7] www.duolingo.com

ous languages while explicitly crowdsourcing lessons from pro-active users and implicitly crowdsourcing translations through well-blended exercises.

The state-of-the-art regarding implicit crowdsourcing and NLP is mainly defined by "Games With A Purpose" (GWAPs) (Lafourcade, Brun, and Joubert, 2015). Some of the most well-known gamified interface for language resource production are JeuxDeMots (Lafourcade, Brun, and Joubert, 2015), Phrase Detectives (Poesio et al., 2012), ZombiLingo (Guillaume, Fort, and Lefebvre, 2016) and Wordrobe (Bos et al., 2017). Finally, two tools were designed for teaching and allow to crowdsource POS corpora (Sangati, Merlo, and Moretti, 2015) and syntactic dependencies (Hladká, Hana, and Lukšová, 2014).

## 6 Concluding Remarks

We have presented enetCollect, the European Network for Combining Language Learning with Crowdsourcing Techniques. The main objective of the action is to create a theoretical framework for achieving a shared understanding of Language Learning and Crowdsourcing. The network is well-balanced in terms of gender and includes both experienced as well as early-stage researchers (PhD students and post-docs), which enables the action to facilitate knowledge transfer and training of new generations of Crowdsourcing-focused researchers. The involvement of new and current members will be pursued through promotion of the Action through relevant communication channels of the research domains concerned. In addition, opportunities for short-term research stays will be advertised through open calls and workshops, and training schools related to relevant topics of the individual working groups will be organized in regular intervals.

## Acknowledgements

The authors have been funded by the Horizon 2020 Framework Programme of the European Union under the enetCollect CA16105 COST action.

## References

Bos, J., V. Basile, K. Evang, N. Venhuizen, and J. Bjerva. 2017. The groningen meaning bank. In N. Ide and J. Pustejovsky, editors, *Handbook of Linguistic Annotation*, volume 2. Springer, pages 463–496.

Guillaume, B., K. Fort, and N. Lefebvre. 2016. Crowdsourcing complex language resources: Playing to annotate dependency syntax. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, Osaka, Japan.

Hladká, B., J. Hana, and I. Lukšová. 2014. Crowdsourcing in language classes can help natural language processing. In *Second AAAI Conference on Human Computation and Crowdsourcing*.

Lafourcade, M., N. L. Brun, and A. Joubert. 2015. *Games with a Purpose (GWAPS)*. Wiley-ISTWiley-ISTE, July.

Poesio, M., J. Chamberlain, U. Kruschwitz, L. Robaldo, and L. Ducceschi. 2012. The phrase detective multilingual corpus, release 0.1. In *Collaborative Resource Development and Delivery Workshop Programme*, page 34.

Sangati, F., S. Merlo, and G. Moretti. 2015. School-tagging: interactive language exercises in classrooms. In *LTLT@ SLaTE*, pages 16–19.

# *Demostraciones*

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 177-180 recibido 21-03-2018 revisado 13-04-2018 aceptado 03-05-2018

# Monitorización de Social Media

## *Social Media Monitoring*

**Rosa Montañés, Rocío Aznar, Saúl Nogueras, Paula Segura, Rubén Langarita, Enrique Meléndez, Paula Peña, Rafael del Hoyo**
Grupo de Big Data y Sistemas Cognitivos
ITAINNOVA (Instituto Tecnológico de Aragón)
C/ María de Luna, nº 7. 50018 Zaragoza
{rmontanes,raznar,snogueras,psegura,rlangarita,emelendez,ppena,rdelhoyo}@itainnova.es

**Resumen:** El sistema desarrollado tiene como objetivo la integración y monitorización de la información en castellano de las redes sociales de un usuario (Facebook, Twitter y noticias web de interés) a través de una única aplicación web. El sistema se sustenta en tres componentes principales: un módulo que implementa una gran variedad de tareas de Procesamiento del Lenguaje Natural (PLN), un módulo software de recuperación de datos de redes sociales mediante crawlers y almacenamiento de resultados, y una aplicación web que presenta una interfaz de usuario para la visualización de la información de forma sugestiva e interactiva. De esta forma, la solución propuesta permite a los usuarios estar actualizados y tener un control de sus redes sociales, pudiendo estar al día de la información, tanto de sus publicaciones como de sus intereses, en una única interfaz sencilla e intuitiva.
**Palabras clave:** Redes sociales, crawlers, PLN, interfaz de usuario

**Abstract:** The developed system aims to integrate and monitor information in Spanish of a user's social networks (Facebook, Twitter and web news of interest) through a single web application. It is based on three main components: a module that implements a wide variety of Natural Language Processing tasks (NLP), an information retrieval module which capture social networks data by means of crawling and stores processing results, and an application web that presents a user interface through which visualizing the information obtained in a suggestive and interactive way. Therefore, the proposed solution allows users to be updated and control their social media networks, being to able to be up-to-date about the information of their publications and their interests, in a single, simple and intuitive graphical interface.
**Keywords:** Social networks, crawlers, NLP, user interface

### *1 Introducción*

Hoy en día, el uso de redes sociales (Twitter, Facebook, blogs, etc.) está ampliamente extendido en todos los ámbitos de la sociedad. Los usuarios, sea cual sea su perfil, comparten gran cantidad de información multimedia en la red, especialmente información escrita, lo que implica la generación de datos textuales de forma masiva. Esto ha llevado en los últimos años al estudio y desarrollo de aplicaciones capaces de explotar estos datos disponibles para extraer analíticas y conocimiento implícito de gran valor (He et al., 2015; Batrinca y Treleaven, 2015; Chang, 2017; Stieglitz et al., 2018).

No obstante, esta generación masiva de información conlleva a su vez a que los usuarios, en su motivación por permanecer informados, encuentren problemas a la hora de entender, clasificar y reconocer la información más relevante de su entorno.

Este prototipo pretende solucionar el problema anterior por medio de la monitorización de la actividad en castellano de las redes sociales del usuario, en concreto Facebook, Twitter y noticias de su interés (publicadas a través de RSS, *Really Simple Sindication* en inglés), mediante el desarrollo de una aplicación web interactiva e intuitiva que permita al usuario la visualización y filtrado de la información más relevante de su entorno, apoyándose en el uso de un amplio abanico de técnicas de procesamiento del lenguaje natural (PLN).

En las siguientes secciones se describe en

ISSN 1135-5948. DOI 10.26342/2018-61-26

detalle la metodología seguida, así como las conclusiones del trabajo realizado y posibles líneas de trabajo futuro.

## *2 Sistema de monitorización de Social Media*

El sistema propuesto se estructura sobre tres módulos funcionales: ingesta y almacenamiento de datos (crawler), procesamiento del lenguaje natural y aplicación web. En la

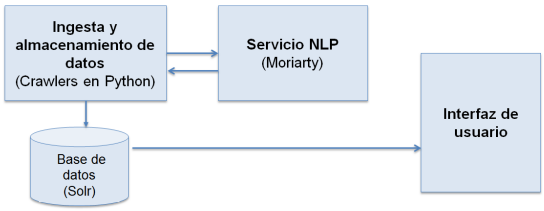


Figura 1: Arquitectura del sistema

Figura 1 se muestra la arquitectura del sistema. En primer lugar, se captura la información mediante crawlers en Python, después se invoca un servicio web a través de su interfaz REST, encargado de realizar el procesado y explotación de la información con la aplicación de técnicas de PLN y, por último, se almacena la información relevante en la base de datos, que es sobre la que se alimenta la interfaz de usuario. El módulo de PLN ha sido implementado en *Moriarty*[1], que es una herramienta de diseño e implementación de soluciones avanzadas de software de Big Data e Inteligencia Artificial desarrollada por ITAINNOVA.

### 2.1 Ingesta y almacenamiento de información

El proceso de ingesta y almacenamiento de información se realiza según una periodicidad configurable y de forma altamente eficiente mediante paralelización. La implementación se compone de tres módulos implementados en Python:

1. **Recuperación de la información**
   Se han creado sendas cuentas de usuario en Facebook y Twitter sobre las que se ha generado actividad variada y se ha recopilado un listado de RSS de interés. La recuperación de la información de las publicaciones realizadas en dichas fuentes se ha realizado mediante crawling. En el caso de Twitter y Facebook haciendo uso de sus APIs públicas (Twitter, 2018; Facebook, 2018), y en el de los RSS mediante consulta directa a la lista configurada. Si las publicaciones recuperadas contienen links de páginas a terceros, el proceso de crawling también extrae el contenido de dichas páginas.

   Como texto origen a analizar, se distingue entre lo que se denomina *perfil* e *interés*. El *interés* hace referencia a las publicaciones de los usuarios a los que se sigue (páginas en caso de Facebook) y el *perfil* a las propias publicaciones del usuario. En el caso de las RSS, todos los documentos se consideran de *interés*.
2. **Procesamiento de la información**
   Una vez se tiene extraído el texto, se preprocesa para eliminar metadatos, como cabeceras, títulos o pies de página para la posterior aplicación de técnicas de PLN. Asímismo, debido a la diversidad de fuentes de información, es necesario realizar un proceso de normalización y estructuración de la información que se recupera.
3. **Almacenamiento de la información**
   Por último, la información resultante de la aplicación de técnicas de PLN se incorpora a la información origen extraída de las fuentes de información y se almacena en una base de datos NoSQL. Se ha elegido el uso de *Apache Solr*, por ser una base de datos orientada a documentos que proporciona funcionalidades avanzadas de indexación y búsqueda, y se ha diseñado un esquema que permite unificar todos los datos de forma sencilla y eficiente.

### 2.2 Procesamiento del Lenguaje Natural

Recuperados y procesados los datos textuales de las publicaciones se hace una invocación a los servicios de PLN, integrados en el framework *Moriarty*.

Puesto que la información de las redes sociales puede aparecer en cualquier idioma y el prototipo se ha enfocado en el castellano, se aplica en primera instancia un algoritmo de detección del languaje que permite filtrar la información.

El servicio de PLN integra una gran variedad de técnicas de PLN dándole un gran po-

[1]"Moriarty". Información disponible en: `http://www.ita.es/moriarty/`


Figura 2: Pantalla principal de la aplicación web

tencial al sistema. En las siguientes secciones se explican las diferentes técnicas utilizadas.

### 2.2.1 Síntesis de la información

Una de las funcionalidades del sistema es la representacion visual de los conceptos más relevantes. Puesto que el dominio de trabajo son las redes sociales, se considera tanto la generación de nube de palabras como la de hashtags. Para la generación de la **nube de palabras** se realiza un preprocesado del texto de forma que se normalice la información y se eliminen conceptos no relevantes. Eliminación de stopwords o lematización son algunas de las técnicas de PLN que son aplicadas para ese propósito.

Además, para ofrecer mayor grado de detalle, el sistema es capaz de generar **resúmenes** de las publicaciones mediante la aplicación de un algoritmo de ranking basado en grafos (Erkan y Radev, 2004) que permite obtener sus frases más significativas.

### 2.2.2 Reconocimiento y clasificación de entidades nombradas

Otra de las funcionalidades es el reconocimiento y clasificación de entidades nombradas (NERC, siglas en inglés) mediante la aplicación de algoritmos basados en redes neuronales (Chiu y Nichols, 2015). Esta tarea permite al usuario tener un conocimiento acerca de las personas, organizaciones y localizaciones a las que hace referencia la publicación.

### 2.2.3 Análisis de sentimiento

Otra de las tareas de PLN que se aplican es el análisis de sentimientos que permite una clasificación de los documentos en diferentes categorías según la opinión que se expone en ellos. Tal como se presenta en la competición TASS (Martınez-Cámara et al., 2017), se diferencian hasta cinco categorías: muy malo, malo, neutro, bueno, muy bueno.

Previo a la aplicación del modelo entrenado, se aplica un preprocesamiento del texto que facilita la clasificación. Algunas de las técnicas que se usan son el reemplazamiento por sinónimos, la eliminación de stopwords o la lematización.

### 2.2.4 Categorización semántica

Además de la clasificación de las publicaciones según la opinión que manifiestan, el servicio PLN integra también una clasificación semántica de las publicaciones según su contenido. Esta categorización se realiza mediante un tesauro. Partiendo de las categorías genéricas que ofrece el estándar de "*Iptc newscodes*"[2], se ha poblado y creado un diccionario propio.

## 2.3 Aplicación web

Extraída la información, procesada y almacenada en la base de datos de Solr se visualizan los resultados en una interfaz gráfica mediante conexión a la base de datos. El desarrollo de la interfaz gráfica se ha realizado pensando en la usabilidad de cualquier usuario, ofreciendo la información de forma sencilla y atractiva. Además, los módulos implementados en la interfaz son interactivos, lo que permite al usuario navegar a través de la infor-

[2] https://iptc.org/standards/newscodes/

mación mostrada. En la Figura 2 se muestra la pantalla principal de la aplicación web.

En la parte de la izquierda de la interfaz se ofrece al usuario la posibilidad de filtrar por diferentes campos, como por ejemplo el tipo de fuente o el emisor (usuario o página a las que se sigue).

En el resto de la interfaz se visualiza la información más relevante de las redes sociales del usuario en diferentes formatos. En el primer bloque se muestra un resumen de cada publicación, con la opción de visualizar la publicación completa en la propia interfaz o incluso pudiendo navegar hasta la publicación original de la red social. La interfaz permite navegar por diferentes pestañas ofreciendo la siguiente información: una nube de palabras y de hastaghs, unos diagramas de sectores de las personas, organizaciones y localizaciones nombradas, un diagrama de sectores que muestra la frecuencia de las categorías de opinión, una evolución de dichas opiniones a través de un gráfico temporal y dos tipos de visualización de árbol de la distribución de las categorías semánticas en las que se han clasificado las publicaciones.

## 3 Conclusiones y trabajo futuro

La aplicación de diferentes técnicas de PLN ha permitido construir un sistema complejo que permite al usuario estar al día de la información más relevante de sus redes sociales.

Aunque se trabaja con diversidad de fuentes, dominios y registros, el sistema desarrollado presenta, en general, una buena precisión de sus resultados. Además, el uso de diferentes filtros de información y gráficas de visualización a través de la interfaz potencia un mayor grado de usabilidad y utilidad para el usuario final.

Además de su usabilidad y precisión, el sistema desarrollado es altamente escalable. En este sentido, el sistema podría extenderse y adaptarse al uso de otros lenguajes mediante el entrenamiento e integración de nuevos modelos de lenguaje, así como dar apoyo a varios perfiles de usuario independientes añadiendo funcionalidades de log-in.

## Agradecimientos

Este trabajo ha sido patrocinado en parte por el Grupo de Big Data y Sistemas Cognitivos del Instituto Tecnológico de Aragón. La difusión de este trabajo ha sido parcialmente financiada por el Programa Operativo FSE para Aragón (2014-2020).

## Bibliografía

Batrinca, B. y P. C. Treleaven. 2015. Social media analytics: a survey of techniques, tools and platforms. *Ai & Society*, 30(1):89–116.

Chang, V. 2017. A proposed social network analysis platform for big data analytics. *Technological Forecasting and Social Change*.

Chiu, J. P. y E. Nichols. 2015. Named entity recognition with bidirectional lstm-cnns. *arXiv preprint arXiv:1511.08308*.

Erkan, G. y D. R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Facebook. 2018. Api overview. facebook developers. Disponible en: https://developers.facebook.com/docs/graph-api/overview/. Recuperado en 2018.

He, W., H. Wu, G. Yan, V. Akula, y J. Shen. 2015. A novel social media competitive analytics framework with sentiment benchmarks. *Information & Management*, 52(7):801–812.

Martınez-Cámara, E., M. Dıaz-Galiano, M. Garcıa-Cumbreras, M. Garcıa-Vega, y J. Villena-Román. 2017. Overview of tass 2017. En *Proceedings of TASS 2017: Workshop on Semantic Analysis at SEPLN (TASS 2017)*, volumen 1896, páginas 13–21.

Stieglitz, S., M. Mirbabaie, B. Ross, y C. Neuberger. 2018. Social media analytics–challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, 39:156–168.

Twitter. 2018. Api overview. twitter developers. Disponible en: https://developer.twitter.com/en/docs. Recuperado en 2018.

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 181-184 recibido 21-03-2018 revisado 18-04-2018 aceptado 03-05-2018

# Advanced analytics tool for criminological research of terrorist attacks

## *Herramienta analítica avanzada para la investigación criminológica de ataques terroristas*

**Marta Romero Hernández**
Pragsis Technologies
Calle de Manuel Tovar 49, Madrid
mromerohdez@gmail.com

**Abstract:** This paper describes a multilingual text summarization and visual analytics tool that provides a searchable database of over 5,000 different online sources, from news portals to social media links and online documents, related to 10 case studies of terrorism for the period 2013 – 2017.
**Keywords:** Multilingual text summarization, DANTE, natural language processing, terrorism, visual analytics

**Resumen:** Este documento describe una herramienta de resúmenes de texto multilingües y visual analytics que proporciona un buscador sobre una base de datos de más de 5.000 fuentes diferentes, desde portales de noticias hasta enlaces de redes sociales y documentos online, relacionados con 10 casos de terrorismo para el período 2013 - 2017.
**Palabras clave:** Generación de resúmenes multilingües, DANTE, procesamiento del lenguaje natural, terrorismo, visual analytics

## *1 Introduction and Motivation*

The analysis of online resources has become a key issue in the monitoring of terrorist activities. Both a theoretical perspective and a set of methodological tools allow to understand and evaluate terrorist organizations, and to develop anti-terrorism policies and practices to detect and interrupt terrorist attacks.

Total Internet traffic has experienced a dramatic growth in the past two decades. As a result, massive amounts of data are generated every day and shared across different social networks. In particular, content related to terrorism also increases, so the use of powerful filtering tools becomes essential.

This framework has been developed in order to support the criminological analysis of terrorist activities that will be performed in DANTE project. The EC project DANTE "*Detecting and analysing terrorist-related online contents and financing activities*" aims to research and develop technologies for more effective, efficient, automated data mining and analytics. These activities will result in an integrated system to detect, retrieve, collect and analyse huge amounts of heterogeneous and complex multimedia and multi-language terrorist-related contents, from both the Surface and the Deep Web, including Dark nets.

Text summarization represents an essential component of the text analysis services in DANTE. Law Enforcement Agencies (LEAs) spend a lot of time reviewing all the information from internet. By summarizing documents, audio and video transcriptions, officers can easily classify all this content and determine whether they want to explore it in more detail or not.

Given this context, this paper presents a valuable tool that provides visual analytic services to show and trace terrorist activities, terrorist group profiles, etc., in order to reduce the information overload on web intelligence experts due to automated summarization of the relevant content. The proposed application works for English, Italian, Spanish, Portuguese and French and is based on Named Entity Recognition and extractive summarization. The

ISSN 1135-5948. DOI 10.26342/2018-61-27 

results of the analysis are presented in a visual way in order to establish complex relationships.

## *2 Corpus extraction*

The domains of interest for the research were the ones already selected by the DANTE Project, namely: i) online financing; ii) online propaganda and iii) online training and information sharing. Comprehensive desk research was carried out along with an in-depth analysis of 10 case studies of terrorist-related activities which have occurred in Europe between 2013 and 2016 (2 cases of foreign fighters; 1 case of failed terrorist attacks; 6 cases of successful terrorist attacks).

From a list of keywords in different languages corresponding to each of these use cases, articles and papers from multiple sources have been retrieved using web scrapping techniques.

The data contains the text for a total of 5024 documents, obtained by public web scraping from the following sources indicated in Table 1:

| Source | Language | N. of documents |
|---|---|---|
| Reuters | English | 716 |
| New York Times | English | 119 |
| Perspectives on Terrorism Journal | English | 36 |
| La Repubblica | Italian | 358 |
| Corriere della Sera | Italian | 265 |
| Itistime Universita Cattolica di Milano | Italian | 16 |
| El Pais | Spanish | 438 |
| El Mundo | Spanish | 184 |
| Agencia EFE | Spanish | 133 |
| Le Figaro | French | 1937 |
| Le Monde | French | 617 |
| Le Parisien | French | 205 |

Table 1: Sources from where the data was obtained.

## *3 Text preprocessing*

### 3.1 Extractive Summarizer

#### 3.1.1 About Text Summarization

There are various kinds of summaries. One distinguishes extractive summaries from abstractive summaries. Extractive summarizers are quite robust since they use existing natural-language phrases that are taken straight from the input. They give a general idea of the text. Abstractive summarizers enable to make more fluent and natural summaries but are more complex to generate and domain-dependent.

There is another classification of automatic summaries that is based on content. An indicative summary is one that informs the reader what a given text or set of texts is about. An informative summary, on the other hand, is one that reproduces the main information of the original and can be used as a replacement of it. Indicative summary contains the main topics of the original text, in contrast to the informative summary that includes full information of the document.

The aim of the summarization task in DANTE is to significantly reduce the text length, without missing the key points of the overall meaning. Since DANTE must deal with arbitrary domain documents (from blogs to propaganda manuals), the indicative-extractive approach, based on term frequency, is selected in this work. This approach is preferred, due to its relative implementation simplicity that is, however, quite robust, unsupervised and fast.

#### 3.1.2 Techniques

Due to the nature of the project, the extractive summarizer is based on term frequency, which has been demonstrated to perform well across several domains. This approach is quite robust since it uses existing natural-language phrases that are taken straight from the input. In addition, it is fast, unsupervised and simple to implement.

The approach presented in this paper consists of three phases:

**Phase I: Preprocessing and weighting**. The pre-processing step consists in a form of dimensionality reduction by removing noise (e.g. uninformative words or conjugation). This phase involves: splitting the text into segments (phrases, sentences, paragraphs); splitting segments into words (tokenization); word normalization (stemming); stop word filtering and redundancy removal.

After the pre-processing stage, each sentence is scored by the frequency of all of its words, that is, the number of times a word appears in the document. If a word's frequency in a document is high, then it can be assumed that this word has a significant effect on the content of the document. Words occurring in a frequent basis increase the score of their

belonging sentences. The total frequency value of a sentence is calculated by summing up the frequency of every word in the document. Thanks to the pre-processing done previously (elimination of redundant phrases, stop words removal and stemming) we can affirm that more frequent words are indeed significant.

**Phase 2: Extraction**. After each sentence is scored, they are arranged in descending order of their score value i.e. the sentence whose score value is highest is in top position and the sentence whose score value is lowest is in bottom position. After ranking the sentences based on their total score the summary is produced selecting certain number of top ranked sentences where the required number of sentences is provided by the user.

**Phase III: Generation.** The algorithm's output was the list of important sentences sorted by score in descending order. The method shows better results if the target number of top sentences is low (2-3 sentences).

## 3.2 Factual Summarizer

### 3.2.1 About Named Entity Recognition

Factual summarizer is based on Named Entity Recognition (NER) answers who, where and when of each event in order to generate a global database of events and involved entities.

This summarizer extracts the most significant entities (person, organization, location) from unstructured textual data, to fill event templates and allow browsing document collections according to them.

### 3.2.2 Techniques

In a preliminary stage a Knowledge Base that contains the known Named Entities is built. This will be used as a dictionary in order to normalize and disambiguate all the different Named Entities recognized by the algorithm.

This database consists of multiple entities classified as LOCATION, PERSON and ORGANIZATION. There is a number of knowledge bases that provide such a background repository for entity classification, predominantly DBpedia, YAGO, and Wikidata. However, there are several reasons to choose Wikidata over other KBs. First, especially when dealing with news articles and social media data streams, it is crucial to have an up-to-date repository of persons and organizations. To the best of our knowledge, Wikidata was the most recent of all, as it provides a weekly data dump. Even though all three KBs (Wikidata, DBpedia, and YAGO3) are based on Wikipedia, Wikidata also contains information about entities and relationships that have not been simply extracted from Wikipedia (YAGO and DBpedia extract data predominantly from infoboxes) but collaboratively added by users.

After the knowledge base is built a process of two phases is performed:

**Phase I: Preprocessing**. This phase is pretty similar to the extractive summarizer, where the text is splitted in sentences and tokenized. The most important step in this phase is the part-of-speech tagger, or POS-tagger, which processes a sequence of words, and attaches a part of speech tag to each word.

**Phase II: Disambiguation rules**. The main idea of this point is to "standardize" the named entities using chunk-joining rules such as the location of the entities within the sentence, the positions of uppercase words, the number of repetitions in the text, etc. Then, candidate named entities are searched in the knowledge base where all the named entities are located and used as a dictionary where they are replaced by the "standardized" named entities.

## *4 Visual representation*

The multilingual text summarization and visual analytics tool will allow LEAs to quickly identify multiple sources of data in multiple languages related to an incident and filter the data down to the most relevant information.

The tool has been adapted to provide a separate dashboard for each case study, allowing the user to search through all online content related to a particular case study.

Users may then apply multiple filters using: keywords, news source, location of source, release date and type of source (Figure 1). For analytical purposes the data can then be displayed in graphs to identify peaks in user engagement with a particular news source (Figure 2).

The use of keywords and various filters makes the tool an ideal aid for the methodological approach adopted in DANTE. As indicated in Figure 2, the chronological ordering of articles and the use of keyword searches will make it possible for investigators and researchers similar to quickly construct a crime script of a particular event.

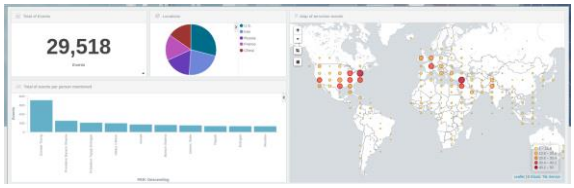


Figure 1: Screenshot of the tool: displaying geographical location of sources on the right and comparative data on the left

Figure 2: Screenshot of the tool: the graph indicates peaks in user engagement with source material and below the title of articles and the summary

## 5 Implementation

The proposed techniques and methodologies are written in python using the natural language processing toolkit (NLTK).

The last phase of data processing consists of the ingestion in the ELK stack. ELK stands for Elasticsearch, Logstash and Kibana which are technologies for creating visualizations from raw data.

Elasticsearch is an open source, distributed, RESTful search engine, usable by any language that speaks JSON and HTTP.

Kibana is a flexible analytics and visualization platform that lets you set up dashboards for real time insight into your Elasticsearch data.

So, data is ingested with Elastic Search and indexes are created to speed up searches. Then, with Kibana is used to design a dashboard that allows to understand and access information in a simple, easier and fast way.

## Acknowledgments

The work presented in this paper was supported by the European Commission under contract H2020-700367 DANTE.

## References

Chen, D., J. Bolton, and D.C. Manning. 2016. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2367.

Erkan, G., and D.R. Radev. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, 22(1):457–479.

Haghighi, A., and L. Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370.

Lin, C.Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, Vol. 8.

Luhn, P.H. 1958. Automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2): 159-165.

Mani, I., G. Klein, D. House, L. Hirschman, T. Firmin, and B. Sundheim. 2002. Summac: a text summarization evaluation. In *Natural Language Engineering*, 8(1):43–68.

Mihalcea, R., and P. Tarau. 2004. TextRank: Bringing Order into Texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 404-411.

Nallapati, R., B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *CoNLL*.

Nenkova, A., and K. McKeown. 2011. Automatic summarization. In *Foundations and Trends in Information Retrieval*, 5(2-3): 103–233.

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

Vanderwende, L., H. Suzuki, C. Brockett, and A. Nenkova. 2007. Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. In *Information Processing and Management*, 43(6): 1606-1618.

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 185-188 recibido 26-03-2018 revisado 19-04-2018 aceptado 03-05-2018

# TEITOK as a tool for Dependency Grammar

## *TEITOK y gramática de dependencia*

**Maarten Janssen**
CELGA-ILTEC, University of Coimbra, Largo da Porta Férrea, Coimbra, Portugal
maartenjanssen@uc.pt

**Abstract:** TEITOK is an online platform for visualizing, searching, and editing TEI-based corpora. TEITOK has a modular set-up, and amongst other things provides an interface to easily add dependency relations to a tokenized TEI document, and provides the option to visualize the dependency trees, edit them, and search through the corpus using dependency relations.
**Keywords:** TEI/XML, dependency grammar, CWB

**Resumen:** TEITOK es una herramienta *online* para visualizar, buscar y editar corpora en el formato TEI/XML. TEITOK utiliza un sistema modular y, entre otras funcionalidades, contiene la opción de añadir dependencias a un documento tokenizado en TEI. También permite visualizar las dependencias de árbole, editar los errores y buscar el corpus resultante utilizando las dependencias.
**Palabras clave:** TEI/XML, gramática de dependencia, CWB

## *1 Introduction*

TEITOK (Janssen, 2016) is an online platform for visualizing, searching, and editing TEI-based corpora. Other than text-based corpora, corpora encoded in the TEI/XML language (Consortium, ) can keep all the typographic information that the original document contains, with a very rich set of annotations defined to account for the demands of very different types of corpora. This makes TEI the preferred representation format for corpora where detailed annotation is crucial, for instance for historical corpora (where detailed paleographic annotation is essential), and learner corpora (where annotation of corrections and errors by the student are of primary importance). But adding linguistic annotations to TEI documents is notoriously difficult. TEITOK attempts to remedy this by providing a user-friendly interface to manipulate TEI documents, tokenize them inline and run computational tools on the document behind the scenes.

TEITOK has a modular design that allows a different visualization of the documents in the corpus depending on their content. There are for instance specific visualization and editing modules for manuscript-based documents, time-aligned spoken documents, and interlinear glossed texts. All these modules work with the same base XML files, and can hence be combined in a single corpus, or even for a single file that contains, say, both manuscript data and a morphological analysis.

In this demo we will present the modules that TEITOK incorporates to work with dependency grammar: there is a script that allows you to add dependency relations to tokenized TEI/XML documents, and then visualize the resulting dependency trees as an SVG image. Incorrect trees can be edited directly in the interface. And you can search through a dependency-parsed corpus in a modified version of the CWB Corpus Query Language that allows you to access the head of a token directly. These different modules make TEITOK a powerful online platform for working with dependency grammars, and do so in an interface that can combine the use of dependency grammar with corpora containing rich typographic information, sound data or a morphological analysis.

### 1.1 File format

A corpus in TEITOK is a collection of files in a TEI/XML compliant format, with some minor adaptations to make it suitable for use in corpus linguistics. The documents can contain virtually any type of information TEI allows you to use, with few restrictions.

The way in which TEITOK adds linguis-

ISSN 1135-5948. DOI 10.26342/2018-61-28

tic annotations to existing documents in the TEI/XML format is by adding inline tokenization nodes, similar to the ⟨w⟩ nodes in standard TEI, and then add linguistic annotations as attributes over those nodes. Since TEITOK takes a linguistic perspective in which punctuation marks are considered tokens, token nodes are named ⟨tok⟩ instead of the standard ⟨w⟩ in TEI/XML, and the possible attributes of ⟨tok⟩ are much more liberal than what standard TEI/XML allows.

For dependency relations, TEITOK uses the columns and names of the CoNLL-U format, converted into attributes over the ⟨tok⟩ elements. An example of a simple two-word dependency-parsed sentence *Good day* in the TEITOK format is given in Figure 1.

Working with the TEITOK format is not that different from working with a line-based format like CoNLL, except that tokens are sequences of XML nodes rather than sequences of lines. However, a TEI-based corpus maintain much more of the information from the source document: in TEI/XML it is easy to keep inter-token spacing, which has to be done in a circumvent way in a line-based format. And it is possible to keep any typesetting information, such as paragraphs, bold and italic text, deleted and added text, color, footnotes, etc. interspersed among the ⟨tok⟩ elements.

It is easy to convert the output of a dependency parser to the TEITOK format, but that would not keep any typesetting data. Therefore, TEITOK assumes you create TEI/XML documents using any of the dedicated tools out there, including some provided by TEITOK itself, and then tokenizes the resulting document. For the tokenized document, it then provides the option to run a parser on the text automatically. This will export the list of tokens to a verticalized format, in which it keeps the ID of each word together with the word itself. It then runs the verticalized text through the web-service of the UDPIPE (Straka and Straková, 2017) multilingual dependency parser. And finally it imports the resulting parsed data back into the XML using the ID of each resulting token. Although this process by default uses UDPIPE, the process can be easily adapted to use any dependency parser using the CoNLL-U format.

## 2 Visualization and Editing

Once dependency relations have been added to the XML document, TEITOK can produce a dependency graph for each sentence in the document, either as a sentence with arches or as a dependency tree. The images are drawn as SVG images, which can be downloaded as SVG, or as rasterized PNG images for easy inclusion in publications.

Since a sentence in TEITOK can contain not only the dependency information but also a lot of typographic information, the sentence itself is visualized above the graph. In TEITOK, this visualization is done by simply reproducing the raw XML of the ⟨s⟩ node in the HTML document, and use CSS to display the TEI mark-up in a visually intuitive way. And since each token in TEITOK can contain much more information than just the dependency relations, all additional attributes on the ⟨tok⟩ (such as the POS tag and the lemma) are shown in a roll-over popup when the mouse is moved over a token in the sentence or in the tree. And to make it easy to see which word in the sentence corresponds to a given node in the tree, the word is highlighted when moving your mouse over a node in the tree. An example of the interface is shown in Figure 2, showing a sentence with the mouse hovering over the node *por* in the tree.

For corpus administrators, the same tree visualization can also be used to correct errors in the automatically produced dependency parses. This works in a very simple and intuitive way: while in edit mode, just click on the wrongly attached node, and then select where to reattach it. Doing so will redraw the graph and allow you to save the resulting tree back into the corpus. To correct an edge label, just click on the label and select the corrected label from a pop-up list, which will by default show the universal dependency labels, but can be customized to work with any other tagset as well. With these options, it is relatively quick to turn an automatically parsed corpus into a manually verified gold-standard corpus.

## 3 Searching

To make corpora searchable, TEITOK exports all tokens of all XML documents in the corpus to a CWB corpus (Evert and Hardie, 2011) using a custom encoding program called TT-CWB-ENCODE that di-

```
<TEI>
<teiHeader/>
<text id="test.txt">
<p id="p-1"><s id="s-1">
  <tok id="w-1" lemma="good" upos="ADJ" xpos="JJ"
    feats="Degree=Pos" head="w-2" deprel="amod">Good</tok>
  <tok id="w-2" lemma="day" upos="NOUN" xpos="NN"
    feats="Number=Sing" deprel="root">day</tok>
</s></p>
</text></TEI>
```

Figure 1: Example in TEITOK format

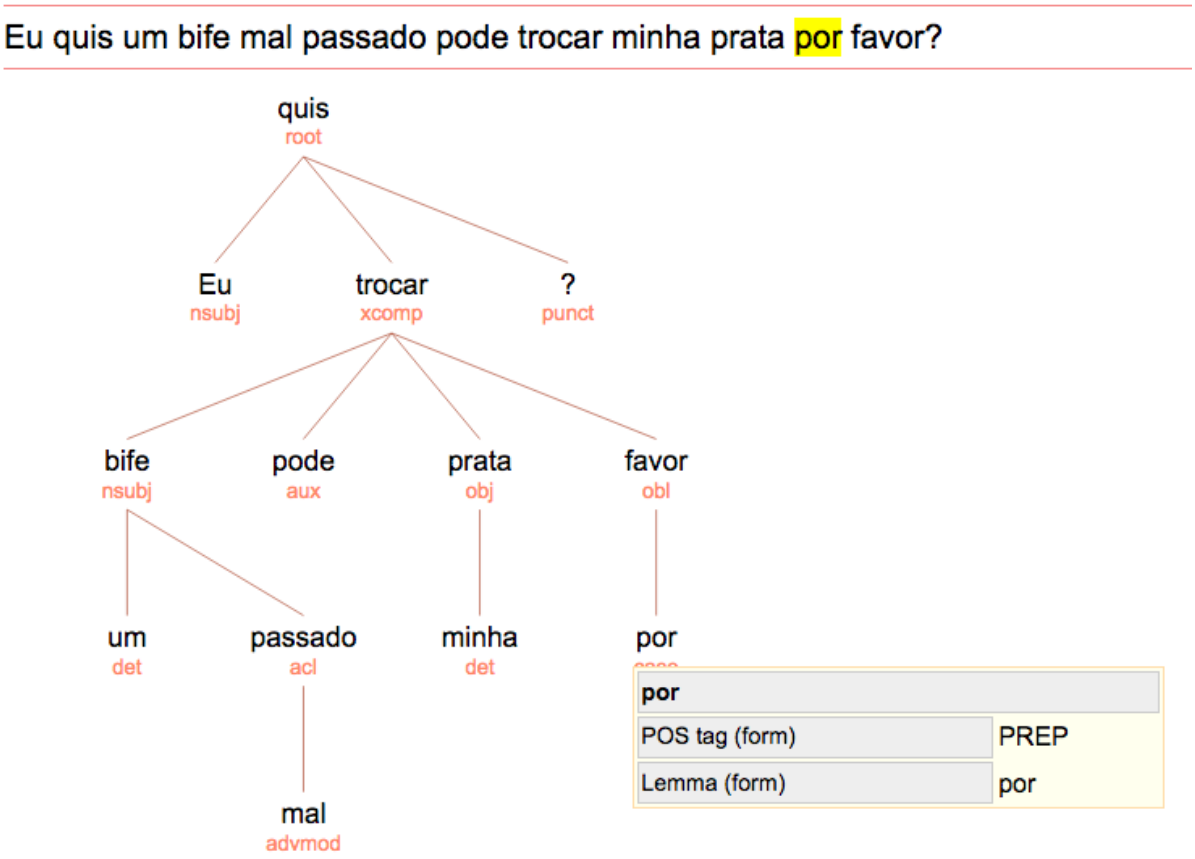


Figure 2: A screenshot of the TEITOK dependency tree view

rectly reads XML files and writes CWB files. Since it is typically not desirable or possible to export everything in the XML file to the searchable corpus, the corpus administrator can define what from the XML files is exported to the CWB corpus, exporting XML regions as sattributes and token-attributes as pattributes.

TEITOK keeps a deep integration between the XML files and the CWB corpus: rather than showing the search results from the CWB query, TEITOK displays the corresponding XML fragments from the XML files used to build the CWB corpus. And the CQL query can also be used to modify selected tokens in the underlying XML files in an efficient manner.

Although CQL can be used to query a dependency-parsed corpus and allows you to search by word, lemma, POS, or edge label, it does not really allow you to search using the dependency relations themselves. TEITOK comes with a custom version of the query language called TT-CQP that does allow this.

```
(2) a:[word="prata" & deprel="obj"] :: head(a).upos="VERB";
    sort head(a).lemma; tabulate a[-1].word, head(a).lemma, head(a).upos
```

Figure 3: Search query in TT-CQP

TT-CQP uses the same files as CWB (and can hence be used with existing CWB corpora), and partially implements the CQL language, while adding a set of additional options to the language.

One added option is that in TT-CQP you can define restrictions on the head of a token, or access the head of a token directly in the output. An example of a query in TT-CQP doing this is given in Figure 3, which looks for all occurrences of *prata* (plate in Portuguese) in the corpus that have a verbal head. It then sorts the results on the lemma of the head, and outputs a list of the first word to the left of each occurrence, as well as the lemma and the POS of the head. For a corpus containing the sentence in Figure 2, that would hence include the result "minha trocar VERB".

Using a query very similar to that in Figure 3, TEITOK can produce an output close to a word sketch from the SketchEngine (Kilgarriff et al., 2014): the list of most typical heads and daughters of a word, ordered by their mutual information score. Such a list gives you the verbs that *prata* is most typically an argument of, or the adjectives that are most frequently used with *prata*.

## 4 Conclusion

TEITOK makes it easy to add dependency relations to existing corpora in the TEI/XML format, taking care of tokenization and parsing behind the scenes. It also makes it easy to correct trees in the interface, and does all this without requiring much understanding of the underlying computational processes from the corpus administrator. This should allow a lot more people to include dependency relations to their corpora, including corpora typically not treated this way such as historical corpora, spoken corpora, and learner corpora.

Once provided with dependency relations, TEITOK allows users to visualize and search the resulting dependencies parses in an intuitive way to make sure the resulting dependency parses are also usable for a wider audience. And it does all this in an interface that also provides a number of tools of a very different nature, such as tools specific to manuscript transcriptions, time-aligned spoken corpora, and morphologically parsed corpora. Hopefully this will lead to an increase in the use of dependency relations in a larger range of corpora.

## References

Consortium, T. Guidelines for electronic text encoding and interchange. http://www.tei-c.org/P5/.

Evert, S. and A. Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Corpus Linguistics 2011*.

Janssen, M. 2016. TEITOK: Text-faithful annotated corpora. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 4037–4043.

Kilgarriff, A., V. Baisa, J. Bušta, M. Jakubíček, V. Kovvář, J. Michelfeit, P. Rychlý, and V. Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, pages 7–36.

Straka, M. and J. Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 189-192 recibido 02-04-2018 revisado 26-04-2018 aceptado 04-05-2018

# Buscador Semántico Biomédico

## *Biomedical Semantic Information Retrieval*

**Pilar López-Úbeda, Manuel Carlos Díaz-Galiano, Arturo Montejo-Ráez, Fernando Martínez-Santiago, Alberto Andreu-Marín, M. Teresa Martín-Valdivia, L. Alfonso Ureña López**
SINAI Group - CEATIC - Universidad de Jaén
Campus Las Lagunillas s/n. E-23071
{plubeda, mcdiaz, amontejo, dofer}@ujaen.es
{aandreu, maite, laurena}@ujaen.es

**Resumen:** El Buscador Semántico Biomédico propone una herramienta web de sencilla utilización para la identificación de terminología médica, la recuperación de literatura especializada y la exploración semántica del contenido gracias a la integración, con requisitos de tiempo de respuesta y alta disponibilidad, de ontologías médicas, técnicas de análisis de texto, reconocimiento de entidades y búsqueda de información sobre fuentes diversas externas. El resultado es una aplicación intuitiva y a la vez potente, que permite identificar la terminología médica sobre cualquier texto con un solo "click". Sobre ese reconocimiento, se permite filtrado de sentidos y subconceptos y la recuperación de información sobre recursos como SciELO, Google Scholar y Medline. Además, el sistema genera un grafo conceptual de manera automática, que permite relacionar semánticamente los términos que aparecen en el texto.
**Palabras clave:** reconocimiento automático de entidades, ontologías, terminología médica, informes médicos, recuperación de información, UMLS

**Abstract:** The Biomedical Semantic Information Retrieval system is an easy web solution to medical term identification, retrieval of specialized literature and semantic concept browsing thanks to the integration, with constraints in speed and high availability, of medical ontologies, text analysis, entity recognition and information retrieval from multiple sources. The result is an intuitive application, yet powerful, that performs term identification of medical concepts over any text with a simple click. Over identified terms the user is able conduct sub-concept selection to fine-tune the retrieval process over resources like SciELO, Google Scholar and Medline. Besides, the system generates a conceptual graph automatically which semantically relates all the terms found in the text.
**Keywords:** automatic entity recognition, term identification, ontologies, biomedical terminology, medical reports, information retrieval, UMLS

## *1 Introducción*

Las herramientas para la comprensión de terminología especializada permiten, por un lado, ayudar a los iniciados en una mejor legibilidad de los textos y, para los expertos, acceder a nivel mayor de análisis cuando esas herramientas proveen de referencias adicionales sobre los términos detectados. Los sistema de detección de terminología médica han despertado siempre mucho interés (Krauthammer and Nenadic, 2004), dado el elevado número de recursos relacionados y la calidad de éstos, como las ontologías especializadas UMLS (Bodenreider, 2004) y MeSH (Díaz-Galiano et al., 2007). La identificación de términos no es sólo interesante sobre textos especializados, sino que puede resultar muy útil también sobre los textos escritos por pacientes (MacLean and Heer, 2013). En ambos casos, los sistemas de identificación de términos se convierten en la clave para acceder a documentación bibliográfica y literatura relacionada, pues estos términos y sus relaciones, toda vez identificados, permiten conectar el conocimiento entre

ISSN 1135-5948. DOI 10.26342/2018-61-29

distintas fuentes.

La plataforma propuesta es lo suficientemente flexible como para que puedan beneficiarse de ella tanto profesionales de medicina como usuarios no especializados (por ejemplo, pacientes).

La búsqueda semántica busca mejorar la precisión de la búsqueda al comprender la intención del usuario y el significado contextual de los términos tal y como aparecen en el espacio de búsqueda, ya sea en la Web o dentro de un sistema cerrado y así generar al lector los resultados más relevantes.

## 2 *Desarrollo del prototipo*

El prototipo es un sistema de recuperación de información sobre textos biomédicos. Identifica terminología especializada automáticamente y la utiliza en un proceso de meta-búsqueda sobre varias bases documentales, enriqueciendo semánticamente los resultados para llegar a obtener mayor precisión. La figura 1 muestra un esquema de su funcionamiento.

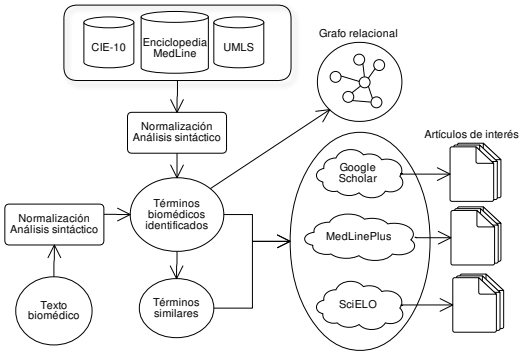


Figura 1: Diagrama de funcionamiento del prototipo

La aplicación recibe como entrada un texto, bien mediante un formulario, bien mediante una extensión para el navegador Google Chrome que permite lanzar el procesamiento desde una simple selección de texto sobre cualquier página web. Se reconocen entidades tales como enfermedades, síntomas y tratamientos, entre otros. Además, el prototipo permite incluir bases de conocimientos de forma rápida y sencilla. El sistema se ha construido sobre tecnologías como Elasticsearch, Google Scholar y el *webservice* de Medline, que pasamos a detallar a continuación.

Elasticsearch[1] es una potente herramienta que nos permite indexar una gran volumen de datos y posteriormente hacer consultas sobre ellos soportando entre otras muchas cosas búsquedas aproximadas, facetas y resaltado.

Google Scholar[2] es un buscador de Google enfocado en el mundo académico que se especializa en literatura científico-académica en una variedad de disciplinas y formatos de publicación. El índice de Google Scholar incluye la mayoría de las revistas y libros académicos en línea revisados por pares, documentos de conferencias, tesis y disertaciones, preimpresiones, resúmene, informes técnicos y otra literatura académica, incluyendo opiniones judiciales y patentes.

MedlinePlus[3] es un servicio de información en línea provisto por la Biblioteca Nacional de Medicina de los Estados Unidos. MedlinePlus contienen enlaces a portales de Internet con información de alrededor de más de 1.000 temas de salud, además, estos temas de salud incluyen enlaces a noticias actualizadas diariamente. Este recurso es de interés para usuarios menos expertos ya que ofrece un vocabulario más informal y familiar para el lector.

Gracias al desarrollo modular del prototipo, es posible independizar diversos aspectos en diferentes servicios incluidos en sistemas potentes que permitan mayor rendimiento sobre bases de conocimiento más amplias. La figura 2 muestra el aspecto final de este prototipo.

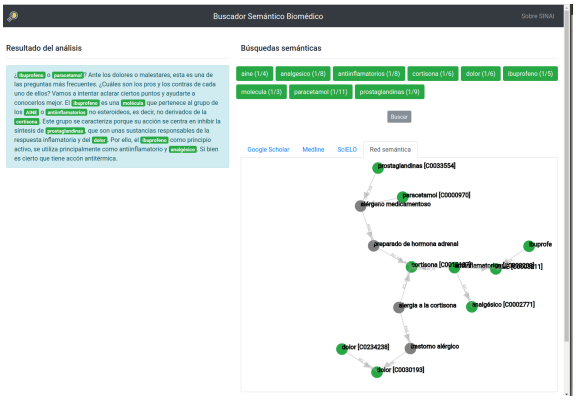


Figura 2: Captura de pantalla del prototipo

### 2.1 Reconocimiento de entidades

Existen herramientas que reconocen entidades médicas en inglés, como MetaMAP (Aronson, 2001) o cTakes (Savova et al., 2010), también para el español como la versión en español de

[1] https://www.elastic.co/

[2] https://scholar.google.es/

[3] https://medlineplus.gov/

MMTx (Carrero, Cortizo, and Gómez, 2008) y Freeling-Med (Oronoz et al., 2013), si bien el uso de estas tecnologías en un herramienta con requisitos de respuesta se tornó poco conveniente. Por lo tanto, se ha diseñado una herramienta propia para la detección de términos biomédicos dentro del texto. Los recursos y algoritmos que utiliza el reconocedor son:

- **Bases de conocimiento**: UMLS (*Unified Medical Language System*), es un compendio de diversos vocabularios y estándares, tanto de salud como biomédicos, para permitir la interoperabilidad entre sistemas informáticos. UMLS permite, entre otros usos, vincular información de salud, términos médicos, nombres de medicamentos y códigos de facturación a través de diferentes sistemas informáticos. Un ejemplo de uso es la vinculación de términos y códigos entre médico, farmacia y compañía de seguros a través de la enciclopedia de Medline y CIE-10.
- **Procesamiento del lenguaje natural**: Para la detección de entidades se ha llevado a cabo una normalización del texto, tanto en los diccionarios utilizados como en el texto introducido por el usuario. La herramienta utilizada en este caso es la biblioteca NLTK (*Natural Language Toolkit*) desarrollada en el lenguaje de programación Python. Además, para obtener una mayor precisión a la hora de identificar terminología (lematización, desambiguación, palabras compuetas), se utiliza el analizador sintáctico incluido en la herramienta CoreNLP desarrollada por la Universidad de Stanford para el español (Manning et al., 2014).
- **Referencias enlazadas**: Dentro del dominio clínico, se hace notar la importancia acerca del intercambio de información debido a los nuevos requisitos de los servicios sanitarios. Para poder seguir prestando servicios como cambios demográficos, movilidad o equidad en el acceso a información de forma efectiva y eficiente es necesario el uso de catálogos estandarizados internacionalmente que unifiquen los datos empleados en las distintas instituciones (Hammond and Cimino, 2006).

  El prototipo muestra los conceptos médicos detectados junto con su código identificador según el diccionario del que se haya extraído la información, este código, además incluye una referencia a una web. Así por ejemplo, el término *'cólera'* tiene asociado el código *C0008354*, *A00* y *000303* en UMLS, CIE-10 y la enciclopedia de MedLine respectivamente.

## 2.2 Búsqueda semántica

La ontología UMLS proporciona para cada concepto detectado otros términos con significado similar. Todos los términos similares contienen el mismo CUI (*Concept Unique Identifier*), por ejemplo, para el código C0004057 se obtienen los conceptos 'Aspirina', 'Ácido acetilsalicílico', 'AAS' o 'aspirina como antiplaquetario', entre otros. Con ello, el usuario puede filtrar su consulta con los términos que él considere apropiados para su búsqueda.

Partiendo de esta terminología seleccionada, se lanzan consultas sobre varias bases documentales como, Google Scholar, Medline o SciELO. Los términos origen de la consulta pueden ser en todo momento modificados mediante la selección de terminología más específica o sinónimos gracias a la posibilidad de personalizar cada término a partir del concepto detectado, como se ha explicado más arriba. Inmediatamente después es posible relanzar las consultas a las distintas fuentes, obteniendo así nuevos resultados.

## 2.3 Grafo relacional

Gracias a que la información que manejamos es muy rica en contenido semántico, una opción añadida al prototipo BSB es modelar dicha información como una red semántica. En UMLS, los términos sinónimos se agrupan para formar un mismo concepto y los conceptos se vinculan unos a otros por medio de varios tipos de relaciones, lo que da como resultado un gráfico enriquecido, etiquetado y dirigido.

Para la generación, se hace uso del algoritmo de Dijkstra también llamado algoritmo de caminos mínimos sobre grafos. Es un algoritmo para la determinación del camino más corto desde un vértice inicial al resto de vértices en un grafo. El vértice origen es el llamado nodo central o centroide, este centroide se elige según la frecuencia de aparición en el texto.

El grafo es interactivo, por lo que el usuario puede elegir cualquier otro nodo como centroide lo cual disparará la regeneración del grafo de forma distinta a la anterior. De esta manera es posible explorar la ontología UMLS

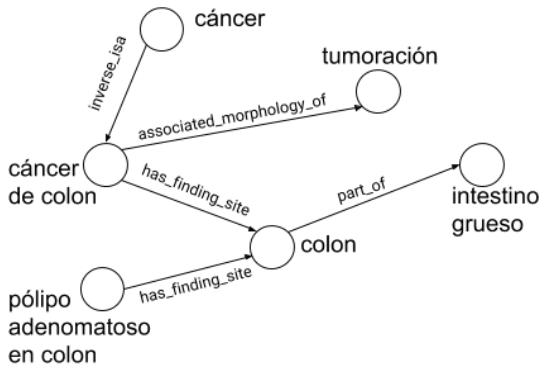


Figura 3: Grafo semántico sobre seis conceptos identificados a partir de un texto

de forma visual e interactiva.

Dentro del metatesauro UMLS, las relaciones simbólicas pueden ser jerárquicas (por ejemplo, '*is a kind of*', '*isa*', '*part of*') o asociativa (por ejemplo, '*location of*', '*caused by*'). En la Figura 3 podemos observar cómo una búsqueda simple de cáncer de colon nos lleva a obtener información relevante acerca de los nodos más cercanos a través de sus relaciones como *has finding site* o *part of*.

## *3 Conclusiones*

El objetivo de la aplicación desarrollada es ofrecer una interfaz sencilla que permita, tanto al experto en medicina como a un usuario no experto, acceso a conocimiento adicional y potencialmente útil a partir de un texto biomédico. El sistema presenta el resultado del análisis (detección de entidades), los resultados de búsqueda en varias fuentes documentales y la red semántica de conceptos identificados, todo esto con respuesta en tiempo real. El sistema es interactivo, pues las búsquedas pueden refinarse a partir de los términos asociados a los conceptos identificados y sobre el grafo es posible explorar las relaciones entre términos de una forma visual y modificar los términos centrales para la construcción del mismo.

## *Agradecimientos*

Este trabajo está parcialmente subvencionado por el proyecto REDES (TIN2015-65136-C2-1-R) del MICINN del Gobierno de España.

## *Bibliografía*

Aronson, A. R. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of AMIA*, page 17. American Medical Informatics Association.

Bodenreider, O. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Carrero, F., J. C. Cortizo, y J. M. Gómez. 2008. Building a spanish mmtx by using automatic translation and biomedical ontologies. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 346–353. Springer.

Díaz-Galiano, M. C., M. García-Cumbreras, M. T. Martín-Valdivia, A. Montejo-Ráez, y L. Urena-López. 2007. Integrating mesh ontology to improve medical information retrieval. In *Workshop of the CLEF*, pages 601–606. Springer.

Hammond, W. E. y J. J. Cimino. 2006. Standards in biomedical informatics. In *Biomedical Informatics*. Springer, pages 265–311.

Krauthammer, M. y G. Nenadic. 2004. Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37(6):512–526.

MacLean, D. L. y J. Heer. 2013. Identifying medical terms in patient-authored text: a crowdsourcing-based approach. *Journal of AMIA*, 20(6):1120–1127.

Manning, C., M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, y D. McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of ACL: system demonstrations*, pages 55–60.

Oronoz, M., A. Casillas, K. Gojenola, y A. Perez. 2013. Automatic annotation of medical records in spanish with disease, drug and substance names. In *Iberoamerican Congress on Pattern Recognition*, pages 536–543. Springer.

Savova, G. K., J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, y C. G. Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of AMIA*, 17(5):507–513.

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 193-196 recibido 03-04-2018 revisado 26-04-2018 aceptado 04-05-2018

# Monge: Geographic Monitor of Diseases

## *Monge: Monitor Geográfico de Enfermedades*

**Salud María Jiménez-Zafra, Flor Miriam Plaza-del-Arco, Miguel Ángel García-Cumbreras, María Dolores Molina-González, L. Alfonso Ureña-López, M. Teresa Martín-Valdivia**
Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{sjzafra, fmplaza, magc, mdmolina, laurena, maite}@ujaen.es

**Abstract:** Monge is a prototype of a geographic monitor of diseases, based on tweets. After the recovering phase of tweets, located in different Spanish cities, these tweets are processed and filtered with techniques and tools of Human Language Technologies. Tweets are filtered with three criteria: location, language (Spanish and Catalan) and bag of words of diseases (generated using synonyms of WordReference and embeddings). The processed information is presented in an interactive way allowing to predict possible epidemic outbreaks of different diseases (e.g. flu, asthma). This demo could be very useful because the Centers for Disease Control and Prevention take between 1-2 weeks from the moment the patient is diagnosed until the data is available, while with this prototype a real-time monitoring of diseases is offered.
**Keywords:** Natural language processing, web application, social monitoring, Twitter, word embeddings

**Resumen:** Monge es un prototipo de un monitor geográfico de enfermedades basado en tweets. Recuperando tweets localizados en distintas ciudades españolas, tanto en español como en catalán, y procesando y analizando la información con técnicas y herramientas de Tecnologías del Lenguaje Humano, permite predecir posibles brotes epidémicos de distintas enfermedades de interés general (gripe, asma, etc.). Los tweets son filtrados utilizando tres criterios: localización, idioma y bolsas de palabras de enfermedades que han sido generadas utilizando sinónimos de WordReference y embeddings. Esta demo podría ser de gran utilidad porque los Centros para el Control y la Prevención de enfermedades tardan entre 1-2 semanas desde que se diagnostica al paciente hasta que los datos están disponibles, mientras que con este prototipo se ofrece una monitorización en tiempo real.
**Palabras clave:** Procesamiento del lenguaje natural, aplicación web, monitorización social, Twitter, word embeddings

## *1 Introduction and Motivation*

Social media has clearly changed how we interact and communicate with each other. There are different mass media in which people published content, such as blogs, social networks, wikis or forums but, currently, social networks are the main one where people express their opinions and experiences. The web has been transformed from a static container of information into a dynamic environment in which users publish any type of information, including ailments and diseases.

In this work, we present *Monge*, a prototype of a geographic monitor of diseases that retrieves tweets located in different Spanish cities, written in Spanish or in Catalan, and allows predict possible epidemic outbreaks of different diseases (e.g. flu, asthma), making use of Human Language Technologies (HLT). This prototype was developed to participate in the *II Hackathon of HLT*[1] that was held on February 26, 2018 in Barcelona, as part of the *Four Years From Now* of the *Mobile World Congress*. It was awarded the second prize in the "General Corpora" category[2].

[1] http://www.agendadigital.gob.es/tecnologias-lenguaje/

[2] https://goo.gl/bSqTcz

ISSN 1135-5948. DOI 10.26342/2018-61-30

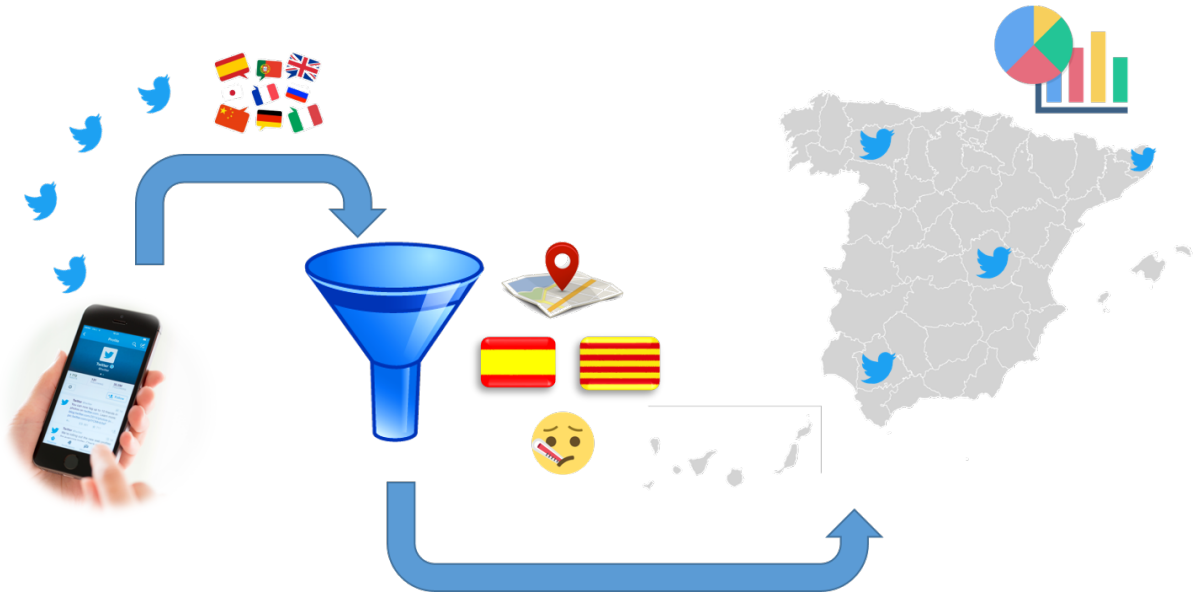

Figure 1: System flow

The reasons that led us to the development of this idea are the fact that most of the Spanish population is connected to the Internet (85% of the Spanish population[3]), performs searches related to health, being diseases one of the main concerns (60% of Spanish Internet users[4]), and uses social networks as way of expression. Therefore, we decided to recover and analyze the publications made in Twitter related to ailments and diseases, because it is one of the main platforms in which people share opinions and experiences (Vilares et al., 2017).

We think that this prototype could be useful because reducing the impact of seasonal epidemics is of vital importance to public health authorities. Studies have shown that effective interventions can be carried out to contain epidemics if there is an early detection. However, the traditional approach used by the Centers for Disease Control and Prevention usually has a delay of 1-2 weeks between the time the patient is diagnosed and the time when the data are available (Achrekar et al., 2011). With this prototype, intoxications, an epidemic or any illness could be detected and monitored in real time at a specific location.

Twitter has been used for real-time notifications such as large-scale fire emergencies, downtime on services provided by content providers (M. Motoyama and Savage, 2010) and live traffic updates. Moreover, there have been efforts in utilizing Twitter data for predicting national mood (Mislove, 2010), currency tracing and performing market and risk analysis.

## 2 *System description*

This prototype is composed by a back-end module, that deals with the retrieval, filtering and processing of tweets, and the front-end monitor, that shows the analyzed data with different interactive elements, such as a map, a line graph, etc.

### 2.1 Back-end

The back-end module was developed in Python, and it consists of the following steps (Figure 1):

1. *Seed words selection.* For this prototype we have selected some diseases as seed words, and for each one we have built different Bag of Words (BoW) to filter tweets. These are the 16 diseases selected: ebola, flu, cold, cancer, asthma, hepatitis, otitis, diabetes, caries, anorexia, obesity, Alzheimer's, AIDS, varicella, measles and appendicitis. The system works with two lists of diseases, one with Spanish words and another one with Catalan words.

2. *Tweets retrieval.* This module uses the streaming API of Twitter to recover

[3] https://goo.gl/9AcJtF
[4] https://goo.gl/CJjgmF

tweets in real time that satisfy the three filtering criteria. For this, the tweepy Python library[5] has been used.

3. *Filtering.* Tweets are filtered following these criteria:

   - *Location.* We have defined geoboxes of specific locations (seven cities from Spain) to analyze and locate only the tweets published in these cities.
   - *Language detection.* The system detects the language of each tweet and processes it only if is in Spanish or Catalan. We have used the langdetect Python library[6] that ports of Google's language detection module.
   - *Filtering by disease.* In this final step, tweets are filtered using different BoW that have been generated from the seed words and have been enriched following two approaches:
     - *Using WordReference.* We have created a BoW using the WordReference API[7] to extract synonyms of each disease selected as seed word in the first step. This BoW has been revised manually in order to filter the words related to the human disease. For instance, the synonyms related to the seed word cold are cold, congestion, flu and catarrh.
     - *Using word embeddings.* In this case, the initial list of diseases has been enriched with the 30 most similar words to each of the seed words using two models based on word embeddings. On the one hand, it was used a model generated using a dump of the Spanish Wikipedia (Montejo-Ráez and Díaz-Galiano, 2016). On the other hand, we used the model developed by Cardellino (Cardellino, 2016) with the Spanish Billion Words Corpus and Embeddings[8], which consists of a collection of texts from various corpora and sources (e.g. Wikipedia, Ancora, OPUS Project) with a total of almost 1.5 billion words.

4. *Tweets preprocessing.* Tweets that meet the filtering criteria are preprocessed as follows: they are tokenized using the TweetTokenizer of NLTK[9], all letters are converted to lower-case, and stopwords and punctuations are removed.
5. *Indexing.* At last, the final set of tweets is indexed using ElasticSearch[10].

## 2.2 Front-end

We have used Kibana[11] to implement the monitor. It is an open-source tool belonging to Elastic, which allows us to visualize, explore and analyze data in real-time that are indexed in ElasticSearch. Kibana is also known for the ELK stack[12] (Elasticsearch, Logstash, Kibana). In this tool, users can create visualizations in the form of tables, charts, maps, histograms, among others. It is useful to create dashboards and helps query data in real time. Dashboards are nothing but an interface for underlying JSON documents. They are used for saving, templating, and exporting. They are simple to set up and use, which helps us play with data stored in ElasticSearch in minutes (Gupta, 2015). In this case, we have created our own dashboard, modifying the default configuration of Kibana.

Our dashboard, shown in Figure 2, is composed of the following elements:

- Spanish map of geographical dispersion of diseases.
- Bar chart with the distribution of diseases by city.
- Table with all the tweets recovered.
- Bar chart with the distribution of diseases by date.
- Line chart with the number of tweets by date and illness.
- Cloud of words most used in tweets.

[5] http://www.tweepy.org/
[6] https://pypi.python.org/pypi/langdetect
[7] http://api.wordreference.com/
[8] http://crscardellino.me/SBWCE/
[9] http://www.nltk.org/api/nltk.tokenize.html
[10] http://elastic.com/
[11] https://www.elastic.co/products/kibana
[12] https://www.elastic.co/elk-stack

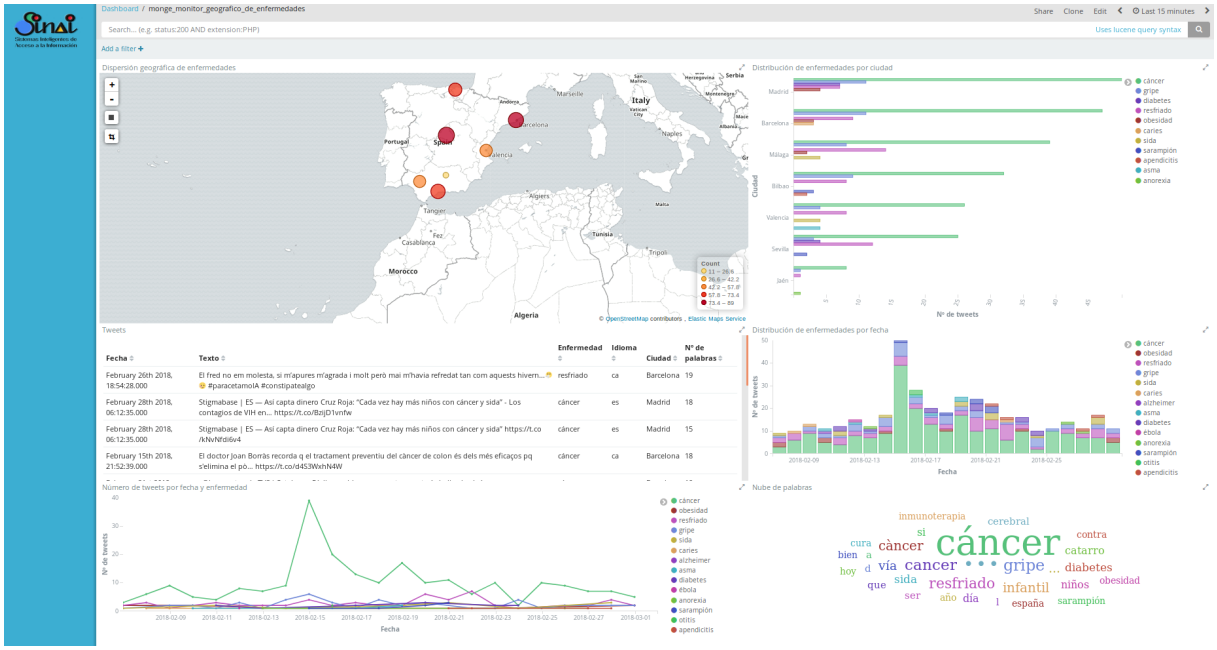


Figure 2: Dashboard

## 3 Conclusions and Future Work

In this paper, we have presented Monge, a prototype of a geographic monitor of diseases, based on tweets.

With the proposed method we can obtain in real time data to track and predict the appearance and spread of an epidemic in a population. In addition, some public health problems can be monitored, such as intoxications, complaints or, in short, any illness that affects a population in a specific location.

As future work, we want to improve the filtering of diseases in tweets and include the analysis of the presence of negation to improve our results (Martí et al., 2016). Moreover, we want to adapt our system to other different languages.

## Acknowledgments

This work has been partially supported by a grant from the Ministerio de Educación Cultura y Deporte (MECD - scholarship FPU014/00983), Fondo Europeo de Desarrollo Regional (FEDER) and REDES project (TIN2015-65136-C2-1-R) from the Spanish Government.

## References

Achrekar, H., A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu. 2011. Predicting flu trends using twitter data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, pages 702–707. IEEE.

Cardellino, C. 2016. Spanish Billion Words Corpus and Embeddings, March.

Gupta, Y. 2015. *Kibana Essentials*. Packt Publishing Ltd.

M. Motoyama, B. Meeder, K. L. G. M. V. and S. Savage. 2010. Measuring online service availability using twitter. In *Workshop on online social networks*.

Martí, M. A., M. T. Martín-Valdivia, M. Taulé, S. M. Jiménez-Zafra, M. Nofre, and L. Marsó. 2016. La negación en español: análisis y tipología de patrones de negación. *Procesamiento del Lenguaje Natural*, 57:41–48.

Mislove, A. 2010. *Pulse of the nation: U.S. mood throughout the day inferred from twitter*.

Montejo-Ráez, A. and M. C. Díaz-Galiano. 2016. Participación de sinai en tass 2016. In *TASS@ SEPLN*, pages 41–45.

Vilares, M., E. S. Trigo, C. Gómez-Rodríguez, and M. A. Alonso. 2017. Tecnologías de la lengua para análisis de opiniones en redes sociales. *Procesamiento del Lenguaje Natural*, 59:125–128.

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 197-200 recibido 05-04-2018 revisado 26-04-2018 aceptado 04-05-2018

# QuarryMeaning: Una aplicación para el modelado de tópicos enfocado a documentos en español

## *QuarryMeaning: A Topic Model Application focused on Spanish Documents*

**Olga Acosta, César Aguilar y Fabiola Araya**
Facultad de Letras de la Pontificia Universidad Católica de Chile
Campus San Joaquín, Santiago de Chile
{oacostal;caguilara;fbaraya}@uc.cl

**Resumen:** Esta demostración presenta una aplicación *standalone* que permite entrenar y probar un modelo de tópicos. Tal aplicación considera filtros para reducir *ruido* en los resultados. Así, por una parte, se incluye una lista de palabras base no relevantes que se puede complementar con otros vocabularios, ya sean propuestos por el usuario, o bien obtenidos mediante un enfoque comparativo usando un corpus de referencia. Por otro lado, es posible considerar únicamente las palabras que tienen un valor semántico alto usando etiquetas de partes de la oración. Además, se incluye un despliegue visual de nubes de palabras que muestra los primeros 10 tópicos derivados del entrenamiento, con el objetivo de explorar visualmente los resultados. Finalmente, se realizó la evaluación de la herramienta considerando una tarea de clasificación de documentos. El modelo logró niveles de precisión superiores al 95% en el conjunto de prueba.
**Palabras clave**: Procesamiento de lenguaje natural, minería de textos, modelación de temas, enfoque contrastivo, clasificación de textos

**Abstract**: This demo shows a standalone application that allows to easily train and test a topic model. The application includes filters for reducing noise in the results. On the one hand, a base stop-list is included, but it can be complemented with a non-relevant word list proposed by user, or obtained it by means of a contrastive approach using a reference corpus. On the other hand, words having a high semantic value can be considered using POS tags. We also include a visualization in word-clouds way, where ten topics can be shown, in order to analyze in detail the results. Finally, evaluation was carried out focusing topic model for classifying documents. Our model achieved levels of precision above 95% in the test set.
**Keywords**: Natural language processing, text mining, topic modeling, contrastive approach, text classification

## *1 Introducción*

En el análisis de datos, tradicionalmente, la reducción de dimensionalidad ha sido una etapa importante debido a que es posible obtener representaciones de menor dimensión que sean más robustas e interpretables. En el escenario actual de abundancia de datos, esta fase cobra mayor relevancia porque nos permite enfocar el análisis en aquellos que sean más importantes, optimizando así los recursos disponibles. Un buen ejemplo de esto es el análisis de fuentes textuales, la indexación de semántica latente (en inglés,

ISSN 1135-5948. DOI 10.26342/2018-61-31 

LSI) y la modelación probabilística de tópicos.

El modelado probabilístico de tópicos es una técnica que ha demostrado ser útil para la extracción de información semántica de fuentes textuales. Este tipo de modelos se conciben como mecanismos para abstraer del discurso real una representación más compacta que capture el contenido de los documentos analizados.

En el caso particular de esta aplicación consideramos el modelado de tópicos basado en la técnica de **asignación de Dirichlet latente** (en inglés, LDA), que describe una clase de modelos donde las propiedades semánticas de las palabras y los documentos se expresan en términos de tópicos probabilísticos (Blei 2012, Steyvers y Griffiths, 2007).

Existen escasas o nulas aplicaciones con una interfaz gráfica que usuarios interesados en realizar este tipo de análisis puedan usar directamente. Un ejemplo de aplicación del tipo anterior es **Stanford Topic Modeling Toolbox** (TMT)[1]; sin embargo, actualmente ya no es mantenido. Por otro lado, existen paquetes bastante eficientes para el procesamiento estadístico de lenguaje natural que incluyen esta técnica, como es el caso de **Mallet**[2] (McCallum, 2002); sin embargo, la interacción para realizar cualquier análisis es vía comandos, lo que puede resultar sumamente complejo para el usuario común. Finalmente, existen librerías que pueden ser importadas en programas, como es el caso del mismo **Mallet**, **lda**[3], **Gensim**[4], por mencionar algunas de las más conocidas. Desafortunadamente, estas librerías asumen conocimientos de programación avanzados.

Dado el escenario anterior, nuestra propuesta con **QuarryMeaning** es proporcionar una interfaz transparente para el usuario común que desee realizar modelado de tópicos. Para lograr nuestro cometido, aunado a proveer de la funcionalidad mencionada, también incluimos opciones para filtrar palabras no relevantes con la finalidad de mejorar los resultados obtenidos en el proceso iterativo de entrenar un modelo. Dichas palabras pueden ser propuestas por el usuario, o bien obtenidas de forma automática. Por otro lado, consideramos también el filtrado de palabras condicionando por la etiqueta de partes de la oración que le corresponda en el discurso real.

Este tipo de aplicación, finalmente, puede beneficiar a aquellos usuarios que posean conocimientos básicos sobre la técnica de modelación de tópicos y nulos conocimientos de programación porque no requerirán construir ningún tipo de comando y contarán además con mecanismos para reducir *ruido* en los resultados cuando asumen la tarea de realizar un modelado de tópicos a partir de textos.

## *2 QuarryMeaning: una aplicación stand-alone*

**QuarryMeaning** es una aplicación stand-alone desarrollada en Python, concretamente con el módulo **Tkinter** y se divide en dos fases: entrenamiento y prueba (véase la Figura 1). En la fase de entrenamiento se define el conjunto de documentos que se usará para construir el modelo, así como los parámetros obligatorios y opcionales para llevar a cabo el análisis. Por otro lado, la fase de prueba requiere también de especificar el conjunto de documentos con el que se probará el desempeño del modelo. A continuación, en las siguientes secciones se ofrece más detalle respecto a cada una de las etapas.

[1] https://nlp.stanford.edu/software/tmt/tmt-0.4/
[2] http://mallet.cs.umass.edu/
[3] https://pypi.python.org/pypi/lda
[4] https://radimrehurek.com/gensim/models/ldamodel.html

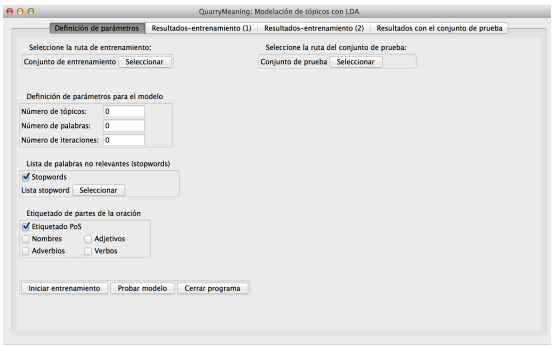


Figura 1. Interfaz del programa

## 2.1 Entrenamiento

En esta fase, el usuario debe proporcionar la ruta donde se encuentran los documentos que se usarán para la etapa de entrenamiento. Además, deberá especificar los parámetros para realizar el proceso (número de tópicos deseados, número de palabras por tópico y número de iteraciones). Asimismo, si desea filtrar la lista base de palabras no relevantes[5] (en este caso, palabras funcionales) deberá activar la casilla de verificación *stoplist* y, posteriormente, seleccionar el archivo. Del mismo modo, si desea incluir palabras de forma manual en la lista, podrá hacerlo directamente editando el archivo correspondiente. Esta opción es útil cuando una palabra corresponde al dominio analizado, pero aparece en más de un tópico y en todos ellos tiene el mismo significado (palabra no polisémica). Por ejemplo, palabras como *enfermedad, paciente, síndrome* en diferentes subáreas de la medicina tienen el mismo significado, así, si aparecen en varios tópicos correspondientes a sub-áreas como la oftalmología, ginecología, etc., pueden ser filtradas.

Existe también la opción de contrastar el conjunto de entrenamiento con un corpus de referencia y todas aquellas palabras que resulten no relevantes podrán ser agregadas a la lista (Acosta, Aguilar e Infante, 2015). Otra de las opciones consideradas es la inclusión de determinadas categorías de palabras, por ejemplo, nombres y adjetivos. Este tipo de categorías gramaticales son las que más se utilizan para construir términos en un dominio específico por lo que esperamos sean las que tengan el valor semántico más alto. Aunado a estas dos categorías, también se pueden seleccionar verbos y adverbios. Cabe mencionar que cuando se activa la casilla del etiquetado de partes de la oración, éste se realiza con el etiquetador TreeTagger para el español (Schmid, 1995).

Una vez especificados los parámetros y activadas las opciones de análisis deseadas, podemos iniciar el entrenamiento del modelo, para ello se utiliza la librería Python lda[6]. Esto generará un despliegue visual para los primeros 10 tópicos en la forma de nube de palabras[7], lo que facilitará el análisis de los mismos (Figura 2):

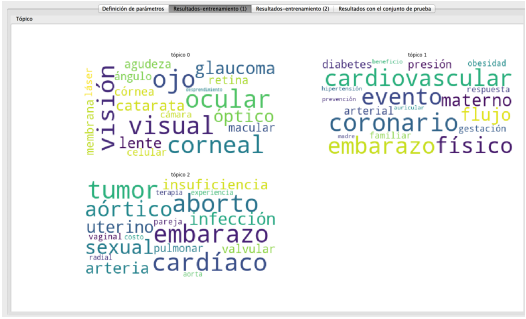


Figura 2. Despliegue visual de tópicos

Si el número de tópicos definido es mayor que 10, tendrá la opción de grabar las imágenes en la carpeta que seleccione para su revisión posterior. Finalmente, podrá ver la distribución de tópicos y el tópico más probable para cada documento dentro de la aplicación, sin embargo, si se han definido muchos tópicos será más apropiado enviar los resultados a un archivo con formato *.csv*, opción disponible también en el programa.

## 2.2 Prueba

La fase de prueba consiste en procesar los documentos correspondientes al conjunto de prueba con el modelo ajustado y explorar el desempeño observando la categoría predicha

[5] La stoplist base fue seleccionada del módulo NLTK.

[6] https://pypi.python.org/pypi/lda

[7] https://pypi.python.org/pypi/wordcloud

como la más probable (tópico principal). Del mismo modo que la etapa de entrenamiento, es posible grabar como .csv la distribución de tópicos y el tópico principal para cada uno de los documentos de prueba.

## *3 Evaluación y resultados*

Cuando se analiza un método estadístico, es útil probarlo con un caso muy simple donde *a priori* se conozca la respuesta correcta, en este caso, el número de tópicos y así observar si el algoritmo puede distinguir correctamente los tópicos. Lo anterior nos sirve para probar la eficiencia del modelo. En este sentido, probamos con un conjunto de 245 artículos en español de tres sub-áreas médicas: oftalmología, ginecología y cardiología, obtenidas de **Scielo**[8] en español.

De ese conjunto, 221 artículos se utilizaron para entrenar el modelo y el resto para probar su desempeño. Los resultados obtenidos considerando el filtrado de palabras no relevantes obtenidas con el enfoque de comparación de corpus y la lista base, considerando además, sólo nombres y adjetivos, así como el filtrado de palabras no polisémicas pero presentes en más de un tópico. Considerando todo esto, se logra una precisión del 100% en la clasificación de los artículos del conjunto de prueba.

## *4 Conclusiones y trabajo a futuro*

Presentamos **QuarryMeaning**, una aplicación *stand-alone* que facilita el entrenamiento y prueba de un modelo de tópicos a usuarios que requieran realizar este tipo de análisis y no cuenten con el conocimiento ni la habilidad en programación.

Los resultados de la evaluación realizada a la herramienta mostraron que es posible generar un modelo de tópicos con un buen desempeño a partir de seleccionar conjuntos de documentos representativos para entrenarlo, no necesariamente enormes cantidades de ellos. Así, en un proceso de entrenamiento iterativo, las diferentes opciones disponibles permiten la obtención de mejores resultados.

Como trabajo futuro, deseamos adaptar la herramienta para trabajar con otros lenguajes, por ejemplo, el inglés. Además, de incluir un módulo para el análisis del número adecuado de tópicos que sirva de soporte para el usuario al momento de decidir cuántos tópicos considerar.

## *Bibliografía*

Acosta, O., C. Aguilar y T. Infante. 2015. Recognition of Terms in Spanish by Applying a Contrastive Approach. *Linguamatica*, *7*(2): 19-34.

Arun, R., V. Suresh, C. Madhavan y M. Murthy. 2010. On finding the natural number of topics with latent dirichlet allocation: Some observations. En Zaki, M., J. Xu Yu, B. Ravindran y V. Pudi (eds.), *Advances in Knowledge Discovery and Data Mining*. Berlin, Springer: 391-402.

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

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. University of Massachusetts, Amherst, Mass., USA: http://mallet.cs.umass.edu/.

Steyvers, M., y T. Griffiths. 2007. Probabilistic topic models. En Landauer, T., D. McNamara, S. Dennis y W. Kintsch (eds.), *Handbook of latent semantic analysis*, Routledge, Oxford, UK: 427-448.

Schmid, H. 1994. Treetagger a language independent part-of-speech tagger. En *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK: 44-49.

[8] www.scielo.org

# *Información General*

# Información para los Autores

**Formato de los Trabajos**

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

**Envío de los Trabajos**

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



## 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:

## 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/

**Proyectos**



**Demostraciones**



**Información General**