



**Sociedad Española para el  
Procesamiento del Lenguaje Natural**



ISSN: 1135-5948

## Artículos

Analysis of eHealth knowledge discovery systems in the TASS 2018 workshop <i>Alejandro Piad-Morffis, Yoan Gutiérrez, Suilán Estévez-Velarde, Yudián Almeida-Cruz, Andrés Montoyo, Rafael Muñoz</i> .....	13
NEGES 2018: Workshop on negation in Spanish <i>Salud María Jiménez-Zafra, Noa P. Cruz Díaz, Roser Morante, María Teresa Martín-Valdivia</i> .....	21
Creación de un corpus de noticias de gran tamaño en español para el análisis diacrónico y diatópico del uso del lenguaje <i>Pavel Razgovorov, David Tomás</i> .....	29
Deep learning approach for negation trigger and scope recognition <i>Hermenegildo Fabregat, Juan Martínez-Romo, Lourdes Araujo</i> .....	37
Hacia una generación de resúmenes sin sesgo a partir de contenido generado por el usuario: Un enfoque preliminar <i>Alejandro Reyes, Elena Lloret</i> .....	45
A different description of orientation in sign languages <i>Antonio F. G. Sevilla, Jose María Lahoz-Bengochea</i> .....	53
Detección de plagio translingüe con grafos semánticos: experimentando con recursos en abierto <i>Ana García Serrano, Antonio Menta Garuz</i> .....	61
Análisis comparativo de las características computacionales en los sistemas modernos de análisis de sentimiento para el español <i>Edgar Casasola-Murillo, Alejandro Pimentel-Alarcón, Gerardo Sierra-Martínez, Eugenio Martínez-Cámara, Gabriela Marín-Raventós</i> .....	69
TASS 2018: The strength of deep learning in language understanding tasks <i>M. Carlos Díaz Galiano, Miguel Á. García Cumbreiras, Manuel García Vega, Yoan Gutiérrez, Eugenio Martínez Cámara, Alejandro Piad-Morffis, Julio Villena Román</i> .....	77
La modelización de la morfología verbal bribri <i>Sofía Flores Solórzano</i> .....	85

## Tesis

La interfaz estructura informativa-prosodia: El rol de la tematicidad jerárquica basado en un modelo empírico <i>Mónica Domínguez Bajo</i> .....	95
Contribuciones a la comprensión lectora: Mecanismos de atención y alineamiento entre n-gramas para similitud e inferencia interpretable <i>Iñigo López-Gazpio</i> .....	99
Similitud entre palabras: Aportaciones de las técnicas basadas en bases de datos <i>Josu Goikoetxea Salutregi</i> .....	103
Irony and sarcasm detection in twitter: The role of affective content <i>Delia Irazú Hernández Farías</i> .....	107
A cross-domain and cross-language knowledge-based representation of text and its meaning <i>Marc Franco-Salvador</i> .....	111
Cross-view embeddings for information retrieval <i>Parth Gupta</i> .....	115

## Información General

XXXV Congreso Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural ..	121
Información para los autores .....	125
Información adicional.....	127





**Sociedad Española para el  
Procesamiento del Lenguaje Natural**



ISSN: 1135-5948

---

## Comité Editorial

### Consejo de redacción

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

**ISSN:** 1135-5948

**ISSN electrónico:** 1989-7553

**Depósito Legal:** B:3941-91

**Editado en:** Universidad de Jaén

**Año de edición:** 2019

**Editores:** Mariona Taulé Delor Universidad de Barcelona mtaule@ub.edu  
M. Teresa Martín Valdivia Universidad de Jaén maite@ujaen.es  
Eugenio Martínez Cámara Universidad de Granada emcamara@decsai.ugr.es

**Publicado por:** Sociedad Española para el Procesamiento del Lenguaje Natural  
Departamento de Informática. Universidad de Jaén  
Campus Las Lagunillas, EdificioA3. Despacho 127. 23071 Jaén  
secretaria.sepln@ujaen.es

### Consejo asesor

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

Raquel Martínez Unanue	Universidad Nacional de Educación a Distancia (España)
Ruslan Mitkov	University of Wolverhampton (Reino Unido)
Manuel Montes y Gómez	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)
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)
Juan-Manuel Torres-Moreno	Laboratoire Informatique d'Avignon / Université d'Avignon (Francia)
José Antonio Troyano Jiménez	Universidad de Sevilla (España)
L. Alfonso Ureña López	Universidad de Jaén (España)
Rafael Valencia García	Universidad de Murcia (España)
René Venegas Velásquez	Pontificia Universidad Católica de Valparaíso (Chile)
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)
<b>Revisores adicionales</b>	
Mario Ezra Aragón	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)
Delia Irazu Hernández Farias	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)
Jose Manuel Perea Ortega	Universidad de Extremadura (España)



**Sociedad Española para el  
Procesamiento del Lenguaje Natural**



ISSN: 1135-5948

---

## Preámbulo

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

Las áreas temáticas tratadas son las siguientes:

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

El ejemplar número 62 de la revista *Procesamiento del Lenguaje Natural* contiene trabajos correspondientes a dos apartados diferentes: comunicaciones científicas y resúmenes de tesis. Todos ellos han sido aceptados mediante el proceso de revisión tradicional en la revista.

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

Se recibieron 25 trabajos para este número, de los cuales 19 eran artículos científicos y 6 resúmenes de tesis. De entre los 19 artículos recibidos, 10 han sido finalmente seleccionados para su publicación, lo cual fija una tasa de aceptación del 52%.

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

Marzo de 2019  
Los editores.



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 62th issue of the *Procesamiento del Lenguaje Natural* journal contains scientific papers and doctoral dissertation summaries. All of these were accepted by a peer review process. We would like to thank the Advisory Committee members and additional reviewers for their work.

Twenty-five papers were submitted for this issue, from which nineteen were scientific papers and six doctoral dissertation summaries. From those nineteen papers, we selected ten (52%) for publication.

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

March 2019  
Editorial board.



**Sociedad Española para el  
Procesamiento del Lenguaje Natural**



ISSN: 1135-5948

## Artículos

Analysis of eHealth knowledge discovery systems in the TASS 2018 workshop <i>Alejandro Piad-Morffis, Yoan Gutiérrez, Suilán Estévez-Velarde, Yudivián Almeida-Cruz, Andrés Montoyo, Rafael Muñoz</i> .....	13
NEGES 2018: Workshop on negation in Spanish <i>Salud María Jiménez-Zafra, Noa P. Cruz Díaz, Roser Morante, María Teresa Martín-Valdivia</i> .....	21
Creación de un corpus de noticias de gran tamaño en español para el análisis diacrónico y diatópico del uso del lenguaje <i>Pavel Razgovorov, David Tomás</i> .....	29
Deep learning approach for negation trigger and scope recognition <i>Hermenegildo Fabregat, Juan Martínez-Romo, Lourdes Araujo</i> .....	37
Hacia una generación de resúmenes sin sesgo a partir de contenido generado por el usuario: Un enfoque preliminar <i>Alejandro Reyes, Elena Lloret</i> .....	45
A different description of orientation in sign languages <i>Antonio F. G. Sevilla, Jose María Lahoz-Bengoechea</i> .....	53
Detección de plagio translingüe con grafos semánticos: experimentando con recursos en abierto <i>Ana García Serrano, Antonio Menta Garuz</i> .....	61
Análisis comparativo de las características computacionales en los sistemas modernos de análisis de sentimiento para el español <i>Edgar Casasola-Murillo, Alejandro Pimentel-Alarcón, Gerardo Sierra-Martínez, Eugenio Martínez-Cámara, Gabriela Marín-Raventós</i> .....	69
TASS 2018: The strength of deep learning in language understanding tasks <i>M. Carlos Díaz Galiano, Miguel Á. García Cumbreiras, Manuel García Vega, Yoan Gutiérrez, Eugenio Martínez Cámara, Alejandro Piad-Morffis, Julio Villena Román</i> .....	77
La modelización de la morfología verbal bribri <i>Sofía Flores Solórzano</i> .....	85

## Tesis

La interfaz estructura informativa-prosodia: El rol de la tematicidad jerárquica basado en un modelo empírico <i>Mónica Domínguez Bajo</i> .....	95
Contribuciones a la comprensión lectora: Mecanismos de atención y alineamiento entre n-gramas para similitud e inferencia interpretable <i>Iñigo López-Gazpio</i> .....	99
Similitud entre palabras: Aportaciones de las técnicas basadas en bases de datos <i>Josu Goikoetxea Salutregi</i> .....	103
Irony and sarcasm detection in twitter: The role of affective content <i>Delia Irazú Hernández Farías</i> .....	107
A cross-domain and cross-language knowledge-based representation of text and its meaning <i>Marc Franco-Salvador</i> .....	111
Cross-view embeddings for information retrieval <i>Parth Gupta</i> .....	115

## Información General

XXXV Congreso Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural ..	121
Información para los autores .....	125
Información adicional.....	127



# *Artículos*



# Analysis of eHealth knowledge discovery systems in the TASS 2018 Workshop

## *Análisis de Sistemas de Descubrimiento de Conocimiento en Documentos de Salud en el Taller TASS 2018*

Alejandro Piad-Morffis<sup>1</sup>, Yoan Gutiérrez<sup>2</sup>, Suilan Estévez-Velarde<sup>1</sup>  
Yudiivián Almeida-Cruz<sup>1</sup>, Andrés Montoyo<sup>2</sup>, Rafael Muñoz<sup>2</sup>,

<sup>1</sup>Department of Artificial Intelligence, University of Havana

<sup>2</sup>Department of Software and Computing Systems, University of Alicante  
apiad@matcom.uh.cu

**Abstract:** This paper presents an analysis of Task 3 eHealth-KD challenge in the TASS 2018 Workshop. The challenge consisted of the extraction of concepts, actions, and their corresponding semantic relations from health-related documents written in the Spanish language. The documents were manually annotated with a schema based on triples (Subject, Action, Target) and an additional set of semantic relations. Several research teams presented computational systems, obtaining relevant results in different subtasks. In this paper, the approaches performed by each team are analyzed and the most promising lines for future development are highlighted and discussed. Moreover, an in-depth analysis of the results is presented focusing on the main characteristics of each subtask. The overall eHealth-KD analysis has indicated that the Knowledge Discovery (KD) task, specifically focused on concrete domains and languages, represents a rich area for further research. In addition, this study considers that the fusion of machine learning –especially deep learning techniques– and knowledge-based approaches will benefit the KD task.

**Keywords:** Machine learning, natural language processing, knowledge bases, knowledge discovery, eHealth

**Resumen:** Este artículo presenta un análisis de la Tarea 3 eHealth-KD in el Taller TASS 2018. La tarea consistió en la extracción de conceptos, acciones, y sus correspondientes relaciones semánticas a partir de documentos sobre temas de salud en idioma español. Los documentos fueron manualmente anotados con un esquema basado en tripletas (Sujeto, Acción, Objeto) y un conjunto adicional de relaciones semánticas. Varios investigadores presentaron sistemas computacionales para la tarea, obteniendo resultados relevantes en las diferentes subtareas definidas. Los enfoques presentados por cada equipo son analizados en este artículo, subrayando las líneas de investigación futura más prometedoras. Además, se presenta un análisis profundo de los resultados, enfocado en las características de cada subtarea. El análisis general de la tarea eHealth-KD indica que las tareas de descubrimiento de conocimiento en idioma español para dominios específicos es un área fructífera de investigación. El progreso en este campo podría beneficiarse considerablemente de la fusión de técnicas de aprendizaje automático –especialmente aprendizaje profundo– con enfoques basados en conocimiento.

**Palabras clave:** Aprendizaje automático, procesamiento de lenguaje natural, bases de conocimiento, descubrimiento de conocimiento, salud electrónica

## 1 Introduction

The automatic discovery and extraction of knowledge from unstructured health text is a growing research field. Recent advances in this area merge natural language pro-

cessing techniques with machine learning and knowledge-based approaches (Liu et al., 2013; Doing-Harris & Zeng-Treitler, 2011; Gonzalez-Hernandez, Sarker, O'Connor, & Savova, 2017). To allow for a fair comparison

of these distinct approaches, and encourage promising ideas, several knowledge discovery challenges have been organized over the years. Recently, the eHealth Knowledge Discovery Challenge (eHealth-KD) was proposed in the TASS 2018 Workshop, which consists of the extraction of (Subject,Action,Target) triples from health-related documents in natural language. The main results of this challenge were presented in the TASS 2018 Overview Report (Martínez-Cámara et al., 2018), where 6 teams of researchers presented widely different approaches with various degrees of success.

The purpose of this paper is to provide a deeper analysis of the characteristics of the participating systems and the difficulties the teams encountered in the different subtasks of the challenge. By identifying which parts of the knowledge discovery problem are more difficult to deal with, researchers can focus their resources and energy into solving these sub-problems. Also, by suggesting which of the current approaches have more potential, we expect to encourage development in these lines in future work.

The semantic structure is a characteristic of the eHealth-KD challenge that is different from similar initiatives. Most similar corpora and tasks are defined in terms of a domain-specific conceptualization, i.e., recognizing health-related concepts such as diseases, symptoms, genes, or treatments (Van Landeghem, Ginter, Van de Peer, & Salakoski, 2011). However, eHealth-KD is based on a general purpose conceptualization, inspired by the Teleologies framework (Giunchiglia & Fumagalli, 2017) and the recognition of (Subject,Action,Target) triplets. This provides a benefit in terms of generalization. The systems presented in this challenge (and other proposals within this framework) are thus easily applicable to different knowledge domains and to cross-domain tasks.

## 2 Task and Corpus Description

The eHealth-KD challenge proposes the identification of two types of elements: **Concepts** and **Actions**. Concepts are key phrases which represent actors relevant in the text domain, while Actions are key phrases that represent the interactions between these Concepts. Actions and Concepts can be linked by two types of roles: **Subject** and **Target**. Four additional semantic

relations between Concepts are defined: **is-a**, **property-of**, **part-of** and **same-as**. These elements are designed to capture the semantics of a broad range of documents without restricting to specific knowledge domains. Figure 1 shows an example.

The overall task is divided into three subtasks that simplify the whole process: Each subtask is aimed at solving a specific sub-problem, with different characteristics.

**Subtask A** Extraction of the relevant key phrases. It can be framed as a standard information extraction task, similar to entity tagging.

**Subtask B** Classifying the key phrases identified in Subtask A as either **Concept** or **Action**. It can be framed as a standard classification task.

**Subtask C** Discovering the semantic relations between pairs of entities. It can be framed as a multi-classification task, where for each possible relation there is an estimation as to whether that relation appears or not.

A more detailed explanation of the eHealth-KD Task is available in the TASS 2018 Overview Report (Martínez-Cámara et al., 2018) and the competition website<sup>1</sup>.

### 2.1 Corpus description

The eHealth-KD corpus consists of a selection of articles collected from the Medline-Plus<sup>2</sup> website. The Spanish entries were selected and pre-processed to remove all markup and leave only plain text. The final documents were manually tagged by a group of 15 annotators. After three stages of annotation and normalization, an average  $F_1$  agreement score of 0.79 was achieved. This  $F_1$  score is a micro-average across all concepts and relations that also considers partial agreement in annotations. The score is based on formulations designed for the Drug Semantics corpus (Moreno, Boldrini, Moreda, & Romá-Ferri, 2017), which presents similar annotation characteristics. This score is not directly comparable to the score obtained by participants, since it does not consider separately the keyword extraction phase and it is computed for the full corpus and not only for the test collection. The corpus has been split

<sup>1</sup><http://www.sepln.org/workshops/tass/2018/task-3>

<sup>2</sup><https://medlineplus.gov/xml.html>



Figure 1: Example annotation of a small set of sentences. The labels used in this annotation schema are explained in Section 2

Metric	Overall	Trial	Train	Dev	Test
<i>Files</i>	11	1	6	1	3
<i>Sentences</i>	1173	29	559	285	300
<i>Annotations</i>	13113	254	5976	3573	3310
<b>Key phrases</b>	7188	145	3280	1958	1805
- Concepts	5366	106	2431	1524	1305
- Actions	1822	39	849	434	500
<b>Roles</b>	3586	71	1684	843	988
- subject	1466	33	693	339	401
- target	2120	38	991	504	587
<b>Relations</b>	2339	38	1012	772	517
- is-a	1057	18	434	370	235
- part-of	393	3	149	145	96
- property-of	836	15	399	244	178
- same-as	53	2	30	13	8

Table 1: Statistics of the eHealth-KD v1.0 corpus

into three sets: a training set, a development set (e.g. for hyper-parameter tuning), and test set for blind evaluation. Table 1 summarizes the main statistics of the corpus.

## 2.2 Task Evaluation Metrics

For comparing different systems, a set of evaluation metrics and evaluation scenarios were designed. The evaluation metrics are based on comparing the output of a given system on a specific file with the gold annotations (as it appears in the corresponding file of the test set). Each subtask (i.e. A, B and C) is independently evaluated, and then a joint score is computed. For the subtask evaluations, the following metrics are defined:

**Correct matches ( $C_A, C_B, C_C$ ):** When one gold and one given annotation exactly match. Used in all subtasks.

**Partial matches ( $P_A$ ):** When two key phrases have a non-empty intersection. Used only for subtask A.

**Missing matches ( $M_A, M_C$ ):** When an annotation in the gold annotations is not found in the output. Used in subtasks A and C.

**Spurious matches ( $S_A, S_C$ ):** When an annotation in an output file does not appear in the gold annotations. used in subtasks A and C.

**Incorrect matches ( $I_B$ ):** When one assigned label is incorrect. Used only for subtask B.

In order to measure the results on individual tasks as well as overall results, the eHealth-KD challenge proposes three evaluation scenarios.

**Scenario 1.** This scenario consists in performing all subtasks (i.e. A, B and C) sequentially. The input is a first set of 100 plain text sentences. Participants must submit the three corresponding output files (one for each subtask). This scenario is designed to evaluate the overall quality of the participant systems. A combined micro  $F_1$  metric was defined, taking into account results of the three tasks<sup>3</sup>:

$$T_{ABC} = C_A + C_B + C_C$$

$$Rec_{ABC} = \frac{T_{ABC} + \frac{1}{2}P_A}{T_{ABC} + P_A + M_A + M_C + I_B}$$

<sup>3</sup>  $T_{ABC}$  is a subtotal used to simplify the formulas.

$$\begin{aligned}
Prec_{ABC} &= \frac{T_{ABC} + \frac{1}{2}P_A}{T_{ABC} + P_A + S_A + S_C + I_B} \\
F_{1ABC} &= 2 \cdot \frac{Prec_{ABC} \cdot Rec_{ABC}}{Prec_{ABC} + Rec_{ABC}}
\end{aligned}$$

**Scenario 2.** This scenario consists in performing only subtasks B and C sequentially. The input is a second set of 100 plain text sentences, and the corresponding gold annotations for subtask A. Participants must submit the output files corresponding to subtasks B and C. This scenario allows participants to be focused on the key phrases classification, without being affected by errors related to the extraction of key phrases. A combined micro  $F_1$  is defined which takes into account results for Subtask B and C<sup>4</sup>:

$$\begin{aligned}
T_{BC} &= C_B + C_C \\
Rec_{BC} &= \frac{T_{BC}}{T_{BC} + I_B + M_C} \\
Prec_{BC} &= \frac{T_{BC}}{T_{BC} + I_B + S_C} \\
F_{1BC} &= \frac{2 \cdot Prec_{BC} \cdot Rec_{BC}}{Prec_{BC} + Rec_{BC}}
\end{aligned}$$

**Scenario 3.** This scenario consists in performing only subtask C. The input is a third set of 100 plain text sentences, plus the corresponding gold annotations for subtasks A and B. Participants must submit only the output file corresponding to subtask C. This scenario allows participants to focus only on the relation discovery problem, without being affected by errors related to the key phrases extraction or classification. The following metric is defined for evaluation:

$$\begin{aligned}
Rec_C &= \frac{C_C}{C_C + M_C} \\
Prec_C &= \frac{C_C}{C_C + S_C} \\
F_{1C} &= 2 \cdot \frac{Prec_C \cdot Rec_C}{Prec_C + Rec_C}
\end{aligned}$$

### 3 Analysis of eHealth Knowledge Discovery Systems

A total of 31 teams originally were registered for the eHealth-KD challenge, from which six successfully submitted the outputs for the evaluation scenarios. To better compare these participants and highlight the most relevant approaches presented, we define the following tags:

**S:** Shallow supervised models such as CRF, logistic regression, SVM, decision trees, etc.

**D:** Deep learning models, such as LSTM, convolutional networks, etc.

**E:** Word embeddings or other embedding models trained with external corpora.

**K:** External knowledge bases, either explicitly or implicitly (i.e., through third-party tools).

**R:** Rules based on domain expertise.

**N:** Classic NLP techniques or features, i.e., POS-tagging, dependency parsing, etc.

The participant systems, a baseline and an ensemble approach, which has been exclusively built for this study, are briefly described next:

**Team UC3M [SDEN]:** Their technique is based on two embedding models (*Glove* and *Reddit vectors*). Training data is preprocessed to the BIOESV tagging codification. Additionally a BI-LSTM model is trained to generate token-specific codes which encode morphological and syntactic features. The combined features are input to a CRF for label prediction (Zavala, Martínez, & Segura-Bedmar, 2018).

**Team SINAI [KRN]:** Their system performs a morphological analysis in the text for each subtask, identifying all the key phrases in the document. They use their own entity detector system using the UMLS concept dictionary in Spanish. For Subtask B, hand-crafted rules are used to discriminate tokens based on their syntactic features (López-Ubeda, Díaz-Galiano, Martín-Valdivia, & Urena-Lopez, 2018).

**Team UPF-UPC [SKN]:** Their system performs a preprocessing step using *Freeling* (POS-tagging and dependency). Additional semantic features are extracted using *YATE* and some external knowledge bases. With these features a CRF is deployed for jointly learning to extract key phrases (Subtask A) and their labels (Subtask B). For Subtask C, shallow supervised classifiers (*logistic regression*) are used, based on a variety of lexical and semantic features (Palatresi & Hontoria, 2018).

**Team TALP [DEN]:** Their system uses convolutional neural networks to solve simultaneously the classification (Subtask B) and the relation extraction (Subtask C). Vector features are based on pre-trained word embeddings (Word2Vec), and some morphological and syntactic features extracted with *Freeling*. They also apply re-sampling techniques to extend the training set (Medina & Turmo, 2018).

**Team LaBDA [DE]:** Their system consists of a convolutional neural network for the extraction of relations. Additionally, tokens are represented via two embeddings, a classic word embedding and another one for encoding the positional correspondence between related tokens (Suarez-Paniagua, Segura-Bedmar, & Martínez, 2018).

<sup>4</sup> $T_{BC}$  is a subtotal used to simplify the formulas.

**Team UH [RN]:** Their system performs a preprocessing step with standard NLP tools (`spacy`) to extract lexical and syntactic features for each token. Afterwards, they apply a set of hand-crafted heuristics for each task.

**Baseline:** To define a comparison baseline, a basic system was developed and trained on the training corpus. This baseline implementation simply stores all annotations seen in the training corpus. At test time, the output is the set of text spans that exactly match the stored annotations. In addition, an ensemble of as well as for this study we

**Ensemble:** Also for comparison purposes, an ensemble was built with the submissions of all participants. The ensemble is built by selecting the subset of submissions that maximizes the macro  $F_1$  metric across all scenarios.

### 3.1 Comparison of Systems

Table 2 summarizes the competition results, and compares them with the baseline implementation executed in the same conditions. Cells marked with a dash (-) indicate that the corresponding participant did not submit for that task or scenario. The metrics shown for each scenario are the corresponding  $F_1$  measures defined in Section 2.2. Subtasks are each evaluated on the corresponding scenario where they are performed first (i.e., Subtask A in scenario 1, and so on).

To better understand the impact of the characteristics of each system in their results, Table 3 shows the relative importance of each system tag for all tasks. These scores are computed by a linear regression estimate of the results of each task, conditioned on each system’s description. Hence, higher values indicate that systems with such tags tend to perform better in a specific task.

These results show that a variety of approaches are relevant for solving the challenges in the eHealth-KD shared task. The best performing submissions include classic supervised learning, deep learning and knowledge-based techniques. In Subtask A, the best approach (UC3M) is based on a CRF model with pre-trained embeddings as features. This can be considered a pure statistical learning approach, since no domain-specific knowledge is used, besides the knowledge implicitly captured in the embeddings. However, the remaining two approaches (SINAI and UPF-UPC) that perform nearly as well do exploit domain-specific knowledge, classic NLP features and shallow supervised learning. It is interesting that the approach presented by SINAI, which is purely based on knowledge bases and hand-crafted rules, obtains a very competitive result to the other two approaches based on machine learning. These results suggest that perhaps a hybrid approach, in which semantic embeddings are specifically adjusted in health-related documents, could provide an edge over general pur-

pose embeddings. These insights are confirmed by Table 3, which shows that on Subtask A the knowledge-based approach has a considerable higher importance, followed by deep learning and embeddings.

In Subtask B the results are similar. In general this subtask appears to be easier than the rest, which is understandable given that there are only two classes and there is a large correlation between word lemmas and their classes (as shown by the relatively high performance of the baseline). In fact, two of these approaches solve both Subtask A and Subtask B simultaneously, framing it as a problem of entity tagging. In this subtask both learning-based and knowledge-based approaches appear to perform at the same level. However, according to Table 3, the most important characteristic is the use of NLP features.

In Subtask C, the top performing approaches (TALP and LaBDA) are based on convolutional neural networks. An interesting phenomenon is that the best systems in Subtask A are not consistent with the best systems in Subtask C. This might suggest that the optimal approach for either subtask is different. However, the best performer in Subtask A did not submit for Subtask C, and vice-versa. Hence, there is not enough evidence that any of their approaches are inappropriate for the other tasks. In Subtask C, classical approaches based on lexical and semantic features (such as those submitted by UPF-UPC) are not very effective, as confirmed by Table 3.

### 3.2 Analysis of the Results

Table 4 summarizes the annotations of the test set that were correctly identified by zero or more participants. This summary also suggests that Subtask A and B are easier than Subtask C. In Subtask A, around 70% of the annotations in the test set were correctly identified by at least three of the participant systems. Likewise, in Subtask B, 71% of the annotations were correctly classified by at least four systems. On the contrary, 64% of the relations in Subtask C were not recognized by any system. The number of annotations recognized by three or more systems is negligible, since only two participant systems showed competitive results in this scenario. In Subtask C, some relations are apparently easier to recognize. Hence, 50% of annotations of **is-a** relations are recognized by at least one participant whereas less than 15% of the **part-of** instances are recognized by at least one participant. This suggests that some relations might have more consistent textual patterns and are thus easier to extract.

With respect to Subtasks A and B, Table 5 (left part of the table) shows all the key phrases with 6 or more appearances in the test set, sorted by the average number of participants that recognized each instance of each key phrase. No-

	UC3M SDEN	SINAI KRN	UPF-UPC SKN	TALP DEN	LaBDA DE	UH RN	Baseline	Ensemble
<b>Subt. A</b>	0.872	0.798	0.805	-	0.323	0.172	0.597	0.799
<b>Subt. B</b>	0.959	0.921	0.954	0.931	0.594	0.639	0.774	0.946
<b>Subt. C</b>	-	-	0.036	0.448	0.444	0.018	0.107	0.501
<b>Average</b>	0.610	0.573	0.598	0.460	0.454	0.276	0.493	0.749
<b>Scen. 1</b>	0.744	0.710	0.681	-	0.310	0.181	0.566	0.695
<b>Scen. 2</b>	0.648	0.674	0.626	0.722	0.294	0.255	0.577	0.731
<b>Scen. 3</b>	-	-	0.036	0.448	0.444	0.018	0.107	0.501
<b>Average</b>	0.464	0.461	0.448	0.390	0.349	0.151	0.417	0.642

Table 2: Summary of systems and results for the TASS 2018 Task 3 event

Subt.	D	E	K	N	R	S
<b>A</b>	0.38	0.38	0.63	0.36	0.18	0.19
<b>B</b>	0.15	0.15	0.28	0.34	-0.01	0.03
<b>C</b>	0.16	0.16	-0.05	0.00	-0.11	-0.05
<b>All</b>	0.15	0.15	0.30	0.01	0.13	0.15

Table 3: Relative importance of systems’ tags for each task as estimated by linear regression on the task results. Higher numbers indicate that systems by the corresponding tag achieve better results in a specific task

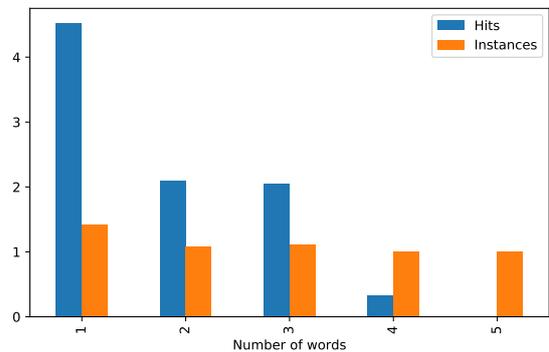
Subtasks A & B						
Hits	0	1	2	3	4	5
Key Phrases	29	37	111	165	251	1
%	4.88	6.22	18.68	27.77	42.25	0.16
Concept	27	28	63	144	608	0
Action	2	9	48	21	236	1
%	2.44	3.11	9.35	13.90	71.10	0.08

Subtask C					
Hits	0	1	2	3	4
is-a	119	64	42	5	5
part-of	82	12	1	1	0
property-of	126	39	11	2	0
same-as	7	1	0	0	0
subject	286	65	41	9	0
target	347	150	76	14	0
Total	967	331	171	31	5
%	64.25	21.99	11.36	02.05	00.33

Table 4: Summary of annotations for each subtask that were correctly identified in the test phase by a given number of participants

tice that these key phrases are a single word. In contrast, the right part of the table shows key phrases with more than 1 word. These are harder to recognize, since fewer instances appear in the training set, and the probability of observing the same sequence of words decreases rapidly with the length of the key phrase.

To support this observation, Figure 2 shows

Figure 2: For all key phrases with the same number of words in the test, **Hits** is the average number of times they are identified by participants, while **Instances** is average number times appear.

the average number of times a key phrase was identified according to the number of words in the phrase, along with the average number of appearances of the key phrase in each text. On average, short key phrases are repeated in the corpus roughly the same number of times than long key phrases. However, long key phrases are harder to identify, presumably because they have less contextual support.

With respect to Subtask C, Table 6 summarizes the most common triplets (left half) and the most often identified (right half). As suggested previously, the **is-a** relation appears to be the easiest to recognize because several consistent textual patterns indicate this relation (e.g., *<Concept> es un <Concept>*). A specific case of **is-a** relation occurs when the target of the relation is a substring of the subject, such as in *is-a (problemas emocionales, problemas)*. In these examples the recall of participants is higher. However, other cases such as *is-a (medicinas, tratamientos)* where there is no direct syntactic pattern to exploit, the recall score is significantly lower. For instance, in the sentence *... los tratamientos incluyen medicinas...* To identify

Key Phrase	Hits	Instances	$\mu$	Key Phrase	Hits	Instances	$\mu$
<i>afecta</i>	20	5	4.0	<i>Estados Unidos</i>	7	2	3.5
<i>cuerpo</i>	20	5	4.0	<i>enfermedades genéticas</i>	3	1	3.0
<i>piel</i>	32	8	4.0	<i>vasos sanguíneos</i>	3	1	3.0
<i>problemas</i>	28	7	4.0	<i>fiebre hemorrágica</i>	3	1	3.0
<i>personas</i>	39	10	3.9	<i>glóbulos rojos</i>	3	1	3.0
<i>tiene</i>	23	6	3.8	<i>síndrome de Marfan</i>	3	1	3.0
<i>proteínas</i>	19	5	3.8	<i>trastorno genético</i>	3	1	3.0
<i>enfermedad</i>	26	7	3.7	<i>terapia intensiva</i>	3	1	3.0
<i>embarazo</i>	18	5	3.6	<i>presión sanguínea</i>	3	1	3.0
<i>vida</i>	18	5	3.6	<i>temperatura corporal</i>	3	1	3.0

Table 5: **Left:** Top key phrases (with 5 or more appearances in the test set) sorted by the average number of hits. **Right:** Top key phrases (with more than 1 word) sorted by the average number of hits

Relation	Hits	Inst.	$\mu$	Relation	Hits	Inst.	$\mu$
<i>is-a (prob. de salud, problemas)</i>	6	5	1.2	<i>is-a (productos químicos, productos)</i>	4	1	4.0
<i>target (tiene, cura)</i>	2	5	0.4	<i>is-a (prob. emocionales, problemas)</i>	4	1	4.0
<i>is-a (contamin. del aire, contamin.)</i>	4	4	1.0	<i>is-a (examen físico, examen)</i>	8	2	4.0
<i>is-a (medicinas, tratamiento)</i>	5	3	1.6	<i>target (tomar, decisiones)</i>	3	1	3.0
<i>part-of (palmas, manos)</i>	2	3	0.6	<i>target (existe, cura)</i>	3	1	3.0
<i>is-a (pruebas genéticas, pruebas)</i>	0	3	0.0	<i>subject (usan, médicos)</i>	6	2	3.0
<i>is-a (medicinas, tratamientos)</i>	2	3	0.6	<i>target (tiene, diabetes)</i>	3	1	3.0
<i>is-a (diabetes gestacional, diabetes)</i>	4	3	1.3	<i>is-a (prof. de la salud, profesional)</i>	3	1	3.0
<i>part-of (aire, contaminación del aire)</i>	0	3	0.0	<i>part-of (tejidos, cuerpo)</i>	3	1	3.0
<i>target (depende, causa)</i>	3	2	1.5	<i>is-a (casos severos, casos)</i>	3	1	3.0

Table 6: **Left:** Top 10 relation triplets sorted by the number of appearances in the test set. **Right:** Top 10 relation triplets sorted by the average number of participants that correctly identified the triplet

such patterns, either external knowledge or some notion of semantic similarity, such as word embeddings is necessary.

Likewise, relation types which are mostly semantic (e.g, **part-of**) obtain a lower recall score in general. An interesting case of *part-of* (*palmas, manos*), which appears in sentences in the form *...las palmas de las manos...*. This textual pattern is similar to many examples of **property-of** relations. Hence, in order to select which is the correct relation, a semantic model is needed to distinguish the concepts **part-of** and **property-of**.

Generally, approaches based on state-of-the-art machine learning seem to dominate individual subtasks. However, by adding domain-specific health related knowledge, less powerful learning techniques can be given a significant boost. Concerning key phrase extraction (Subtask A), most participants use NLP features, either explicitly, or implicitly captured in word embeddings and other representations. The best overall systems do not generalize across the three tasks, while

systems that do generalize do not outperform the baseline in general.

#### 4 Conclusions and Future Work

The following conclusions can be drawn from this study. First, the complexity of all three subtasks is not the same. Subtask B is the easiest and can be considered mostly solved, while Subtask C appears to be the most complex. For Subtask A there is not enough evidence to determine if the top result ( $F_1 = 0.872$ ) is close to human performance, due to the difficulty of arriving at a satisfactory annotation agreement of the corpora. Furthermore, in Subtask C, not all types of semantic relations have equal complexity. The **is-a** relation appears to be simpler to identify, given the relatively straight-forward syntactic patterns in which it occurs. Other relations that have more complex patterns will require a higher degree of semantic understanding of the text for a successful extraction.

The technologies deployed in eHealth-KD challenge indicate that the knowledge discovery

task in health-related documents written in the Spanish language is an attractive future research field. Significant advances in knowledge discovery tasks will require a solid integration of machine learning techniques with knowledge-based approaches, to exploit the strengths of each discipline. The lack of manually tagged Spanish language corpora related to specific domains makes progress more challenging. The eHealth-KD challenge and similar initiatives constitute the first steps towards building friendly competition scenarios, in which researchers from the natural language processing community can evaluate different techniques.

### Acknowledgments

This research has been partially supported by a Carolina Foundation grant in agreement with University of Alicante and University of Havana, sponsoring to Suilan Estevez-Velarde. Moreover, it has also been partially funded by both aforementioned universities and Generalitat Valenciana through the projects PROMETEU/2018/089, PINGVALUE3-18Y and SocialUniv 2.0(ENCARGOINTERNOOMNI-1).

### References

- Doing-Harris, K. M., & Zeng-Treitler, Q. (2011). Computer-assisted update of a consumer health vocabulary through mining of social network data. *Journal of medical Internet research*, 13(2).
- Giunchiglia, F., & Fumagalli, M. (2017). Teleologies: Objects, actions and functions. In *International conference on conceptual modeling* (pp. 520–534).
- Gonzalez-Hernandez, G., Sarker, A., O’Connor, K., & Savova, G. (2017). Capturing the patient’s perspective: a review of advances in natural language processing of health-related text. *Yearbook of medical informatics*, 26(01), 214–227.
- Liu, H., Bielinski, S. J., Sohn, S., Murphy, S., Waghlikar, K. B., Jonnalgadda, S. R., ... Chute, C. G. (2013). An information extraction framework for cohort identification using electronic health records. *AMIA Summits on Translational Science Proceedings, 2013*, 149.
- López-Ubeda, P., Díaz-Galiano, M. C., Martín-Valdivia, M. T., & Urena-Lopez, L. A. (2018). Sinai en tass 2018 task 3. clasificando acciones y conceptos con umls en medline. In *Tass 2018 – taller de análisis semántico en la sepln*.
- Martínez-Cámara, E., Almeida-Cruz, Y., Díaz-Galiano, M. C., Estévez-Velarde, S., García-Cumbreras, M. A., García-Vega, M., ... Julio, V.-R. (2018, September). Overview of TASS 2018: Opinions, health and emotions. In *Proceedings of tass 2018: Workshop on semantic analysis at sepln (tass 2018)* (Vol. 2172). Sevilla, Spain: CEUR-WS.
- Medina, S., & Turmo, J. (2018). Joint classification of key-phrases and relations in electronic health documents. In *Tass 2018 – taller de análisis semántico en la sepln*.
- Moreno, I., Boldrini, E., Moreda, P., & Romá-Ferri, M. T. (2017). Drugsemantics: a corpus for named entity recognition in spanish summaries of product characteristics. *Journal of biomedical informatics*, 72, 8–22.
- Palatresi, J. V., & Hontoria, H. R. (2018). Medical knowledge discovery by combining multiple techniques and resources. In *Tass 2018 – taller de análisis semántico en la sepln*.
- Suarez-Paniagua, V., Segura-Bedmar, I., & Martínez, P. (2018). Labda at tass-2018 task 3: Convolutional neural networks for relation classification in spanish ehealth documents. In *Tass 2018 – taller de análisis semántico en la sepln*.
- Van Landeghem, S., Ginter, F., Van de Peer, Y., & Salakoski, T. (2011). Evex: A pubmed-scale resource for homology-based generalization of text mining predictions. In *Proceedings of bionlp 2011 workshop* (pp. 28–37). Stroudsburg, PA, USA: Association for Computational Linguistics.
- Zavala, R. M. R., Martínez, P., & Segura-Bedmar, I. (2018). A hybrid bi-lstm-crf model for knowledge recognition from ehealth documents. In *Tass 2018 – taller de análisis semántico en la sepln*.

# NEGES 2018: Workshop on Negation in Spanish

## *NEGES 2018: Taller de Negación en Español*

Salud María Jiménez-Zafra<sup>1</sup>, Noa P. Cruz Díaz<sup>2</sup>,  
Roser Morante<sup>3</sup>, María Teresa Martín-Valdivia<sup>1</sup>

<sup>1</sup>SINAI, Centro de Estudios Avanzados en TIC (CEATIC), Universidad de Jaén  
{sjzafra, maite}@ujaen.es

<sup>2</sup>Savana Médica, Madrid  
contact@noacruz.com

<sup>3</sup>CLTL Lab, Computational Linguistics, VU University Amsterdam  
r.morantevallejo@vu.nl

**Abstract:** This paper presents the 2018 edition of NEGES, Workshop on Negation in Spanish, that took place on September 18 as part of the 34th International Conference of the Spanish Society for Natural Language Processing. In this edition, three tasks were proposed: Task 1: “Annotation guidelines”, Task 2: “Negation cues detection”, and Task 3: “Sentiment analysis”. The dataset used for Task 2 and Task 3 was the SFU ReviewSP-NEG corpus. About 10 teams showed interest in the tasks and 4 teams finally submitted results.

**Keywords:** NEGES 2018, negation, annotation guidelines, negation processing, cue detection, sentiment analysis

**Resumen:** Este artículo presenta la edición 2018 de NEGES, Taller de NEGación en ESpañol, que tuvo lugar el 18 de septiembre como parte de la 34 edición de la Conferencia Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural. En esta edición se propusieron 3 tareas: Tarea 1: “Guías de anotación”, Tarea 2: “Detección de claves de negación”, y Tarea 3: “Análisis de sentimientos”. El conjunto de datos utilizado para la Tarea 2 y para la Tarea 3 fue el corpus SFU ReviewSP-NEG. Unos 10 equipos mostraron interés por las tareas y 4 equipos presentaron resultados.

**Palabras clave:** NEGES 2018, negación, guías de anotación, procesamiento de la negación, detección de claves, análisis de sentimientos

## 1 Introduction

Negation is a complex linguistic phenomenon of growing interest in computational linguistics. Detection and treatment of negation is relevant in a wide range of applications such as sentiment analysis or information retrieval, where it is crucial to know when a part of the text should have a different meaning due to the presence of negation. If we want to develop systems that approach human understanding, it is necessary to incorporate the treatment of negation, a linguistic phenomenon that we use constantly. In recent years, several challenges and shared tasks have focused on processing negation (Morante and Sporleder, 2010; Farkas et al., 2010; Morante and Blanco, 2012). However, most of the research on negation has been done for English. Therefore, the 2018 edi-

tion of the NEGES Workshop<sup>1</sup> aimed to advance the study of this phenomenon in Spanish, the second most widely spoken language in the world and the third most widely used on the Internet. The main objective was to bring together the scientific community that is working on negation to discuss how it is being addressed, what are the main problems encountered, as well as sharing resources and tools aimed at processing negation in Spanish.

The rest of this paper is organized as follows. The proposed tasks are described in Section 2, and the data used in Section 3. Evaluation measures are introduced in Section 4. Participating systems and their results are summarized in Section 5. Finally, Section 6 concludes the paper.

<sup>1</sup><http://www.sepln.org/workshops/neges/index.php?lang=en>

Corpus	Domain	Annotation guidelines
UAM Spanish TreeBank	News	pp. 51-55 (Sandoval and Salazar, 2013)
IxaMed-GS	Clinical reports	pp. 322 (Oronoz et al., 2015)
SFU ReviewSP-NEG	Product reviews	pp. 538-559 (Jiménez-Zafra et al., 2018)
UHU-HUVR	Clinical reports	pp. 54-57 (Cruz Díaz et al., 2017)
IULA Spanish Clinical Record	Clinical reports	pp. 45-49 (Marimon, Vivaldi, and Bel, 2017)

Table 1: Annotation guidelines provided for Task 1

## 2 Tasks description

In the 2018 edition of the Workshop on Negation in Spanish, three tasks were proposed:

- **Task 1:** “Annotation guidelines”
- **Task 2:** “Negation cues detection”
- **Task 3:** “Sentiment analysis”

The following is a description of each task.

### 2.1 Task 1

Task 1 of NEGES 2018, “Annotation guidelines”, had as goal to reach an agreement on the guidelines to follow for the annotation of negation in Spanish texts. Although there have already been several annotation efforts, the community lacks a standard for the annotation of negation, contrary to what happens with other phenomena, such as semantic roles.

The corpora annotated so far in Spanish belong to 3 domains (news, clinical reports and product reviews) and are based on different guidelines. In this task, the guidelines used for the annotation of the corpora were made available to the participants so that they could analyze them (Table 1). A period of analysis was provided and once it was over, participants sent a document indicating which aspects of the guidelines they agreed with and which they did not, all duly justified. The documents describing the perspective of each team were sent to the rest of participants prior to the workshop in order to enhance a discussion about the main aspects of interest and try to reach a consensus.

### 2.2 Task 2

Task 2 of NEGES 2018, “Negation cues detection”, had the aim to promote the development and evaluation of systems for identifying negation cues in Spanish. For example, in sentence [1] the systems had to identify three negation cues: i) *En mi vida*, ii) *no* and iii) *sin*.

- [1] **En mi vida** he hecho una reserva con tanta antelación, **no** quería quedarme **sin** sitio.

Participants received a set of training and development data to build their systems during the development phase. The manual annotation of the negation cues was performed by domain experts, following well-defined annotation guidelines (Jiménez-Zafra et al., 2018; Martí et al., 2016). At a later stage, a set of tests were made available for evaluation. The participant’s submissions were evaluated against the gold standard annotations. Negation cues could be single words (e.g., “no” [*no/not*]), multiwords (e.g., “ni siquiera” [*not even*]) or discontinuous words (e.g., “no...apenas” [*not...hardly*]).

### 2.3 Task 3

Task 3 of NEGES 2018, “Sentiment analysis”, was proposed to evaluate the role of negation in sentiment analysis. In this task, participants had to develop a system that used the negation information contained in a corpus of reviews of movies, books and products (Jiménez-Zafra et al., 2018) to improve the task of polarity classification. They had to classify each review as *positive* or *negative* using an heuristic that incorporated negation processing.

## 3 Data

The SFU ReviewSP-NEG corpus<sup>2</sup> (Jiménez-Zafra et al., 2018) was the collection of documents used to train and test the systems presented in *Task 2* and *Task 3*<sup>3</sup>. This corpus is an extension of the Spanish part of the SFU Review corpus (Taboada, Anthony, and Voll, 2006) and it could be considered the

<sup>2</sup><http://sinai.ujaen.es/sfu-review-sp-neg-2/>

<sup>3</sup>To download the data in the format provided for *Task 2* and *Task 3* go to <http://www.sepln.org/workshops/neges/index.php?lang=en> or send an email to the organizers

counterpart of the SFU Review Corpus with negation and speculation annotations (Konstantinova et al., 2012).

The Spanish SFU Review corpus (Taboada, Anthony, and Voll, 2006) consists of 400 reviews extracted from the website Ciao.es that belong to 8 different domains: cars, hotels, washing machines, books, cell phones, music, computers, and movies. For each domain there are 50 positive and 50 negative reviews, defined as positive or negative based on the number of stars given by the reviewer (1-2=negative; 4-5=positive; 3-star review were not included). Later, it was extended to the SFU ReviewSP-NEG corpus (Jiménez-Zafra et al., 2018) in which each review was automatically annotated at the token level with PoS-tags and lemmas using Freeling (Padró and Stanilovsky, 2012), and manually annotated at the sentence level with negation cues and their corresponding scopes and events. Moreover, it is the first corpus in which it was annotated how negation affects the words within its scope, that is, whether there is a change in the polarity or an increase or decrease of its value. Finally, it is important to note that the corpus is in XML format and it is freely available for research purposes.

### 3.1 Datasets Task 2

The SFU ReviewSP-NEG corpus was randomly splitted into development, training and test sets with 33 reviews per domain in training, 7 reviews per domain in development and 10 reviews per domain in test. The data was converted to CoNLL format (Buchholz and Marsi, 2006) where each line corresponds to a token, each annotation is provided in a column and empty lines indicate the end of the sentence. The content of the given columns is:

- Column 1: domain\_filename
- Column 2: sentence number within domain\_filename
- Column 3: token number within sentence
- Column 4: word
- Column 5: lemma
- Column 6: part-of-speech
- Column 7: part-of-speech type

- Columns 8 to last: if the sentence has no negations, column 8 has a “\*\*\*” value and there are no more columns. Else, if the sentence has negations, the annotation for each negation is provided in three columns. The first column contains the word that belongs to the negation cue. The second and third columns contain “-”.

The distribution of reviews and negation cues in the datasets is provided in Table 2. Moreover, in Figure 1 and Figure 2, we show 2 examples of the format of the files with different types of sentences. In the first example (Figure 1) there is no negation so the 8th column is “\*\*\*” for all tokens, whereas the second example (Figure 2) is a sentence with two negation cues in which information for the first negation is provided in columns 8-10, and for the second in columns 11-13.

	Reviews	Negation cues
Training	264	2,511
Development	56	594
Test	80	836

Table 2: Distribution of reviews and negation cues in the datasets of Task 2

### 3.2 Datasets Task 3

For this task, we provided the SFU ReviewSP-NEG corpus with the original format (XML). The meaning of the labels used are the following:

- <review\_polarity=“positive/negative”>. It describes the polarity of the review, which can be “*positive*” or “*negative*”.
- <sentence\_complex=“yes/no”>. This label corresponds to a complete phrase or fragment thereof in which a negation structure can appear. It has associated the *complex* attribute that can take one of the following values:
  - “yes”, if the sentence contains more than one negation structure (<neg\_structure>).
  - “no”, if the sentence only has a negation structure.
- <neg\_structure>. This label corresponds to a syntactic structure in which a negation cue appears. It has 4 possible

coches_no_1_21	5	1	Solo	solo	rg	-	***		
coches_no_1_21	5	2	tenia	tener	vmi3s0	main	***		
coches_no_1_21	5	3	57000	57000	z	-	***		
coches_no_1_21	5	4	km	kilómetro	ncmn000	common	***		
coches_no_1_21	5	5	y	y	cc	coordinating	***		
coches_no_1_21	5	6	algo	algo	pi0cs000	indefinite	***	***	
coches_no_1_21	5	7	menos	menos	rg	-	***		
coches_no_1_21	5	8	de	de	sps00	preposition	***		
coches_no_1_21	5	9	tres	3	z	-	***		
coches_no_1_21	5	10	años	año	ncmp000	common	***		
coches_no_1_21	5	11	.	.	fp	-	***		

Figure 1: Sentence without negation in CoNLL format

hoteles_no_2_6	9	1	Aun	aun	np00000	proper	-	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	2	estoy	estar	vaip1s0	auxiliary	-	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	3	esperando	esperar	vmg0000	main	-	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	4	que	que	cs	subordinating	-	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	5	me	me	pp1cs000	personal	-	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	6	carguen	cargar	vmsp3p0	main	-	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	7	los	el	da0mp0	article	-	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	8	puntos	punto	ncmp000	common	-	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	9	en	en	sps00	preposition	-	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	10	mi	mi	dp1cs	possessive	-	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	11	tarjeta	tarjeta	ncfs000	common	-	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	12	más	más	rg	-	-	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	13	,	,	fc	-	-	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	14	no	no	rn	negative	no	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	15	sé	saber	vmip1s0	main	-	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	16	dónde	dónde	pt000000	interrogative	-	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	17	tienen	tener	vmip3p0	main	-	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	18	la	el	da0fs0	article	-	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	19	cabeza	cabeza	ncfs000	common	-	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	20	pero	pero	cc	coordinating	-	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	21	no	no	rn	negative	-	-	-	-	-	-	-	no	-	-	-	-	-
hoteles_no_2_6	9	22	la	lo	pp3fsa00	personal	-	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	23	tienen	tener	vmip3p0	main	-	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	24	donde	donde	pr000000	relative	-	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	25	deberían	deber	vmic3p0	main	-	-	-	-	-	-	-	-	-	-	-	-	-
hoteles_no_2_6	9	26	.	.	fp	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Figure 2: Sentence with two negations in CoNLL format

attributes, two of which (*change* and *polarity\_modifier*) are mutually exclusive.

- *polarity*: it presents the semantic orientation of the negation structure (“*positive*”, “*negative*” or “*neutral*”).
- *change*: it indicates whether the polarity or meaning of the negation structure has been completely changed because of the negation (*change*=“*yes*”) or not (*change*=“*no*”).
- *polarity modifier*: it states whether the negation structure contains an element that nuances its polarity. It can take the value “*increment*” if there is an increment in the intensity of the polarity or, on the contrary, it can take the value “*reduction*” if there is a reduction.
- *value*: it reflects the type of the negation structure, that is, “*neg*” if it expresses negation, “*contrast*” if it indicates contrast or opposition between terms, “*comp*” if it expresses a comparison or inequality

ity between terms or “*noneg*” if it does not negate despite containing a negation cue.

- *<scope>*. This label delimits the part of the negation structure that is within the scope of negation. It includes both, the negation cue (*<negexp>*) and the event (*<event>*).
- *<negexp>*. It contains the word(s) that constitute(s) the negation cue. It can have associated the attribute *discid* if negation is represented by discontinuous words.
- *<event>*. It contains the words that are directly affected by negation (usually verbs, nouns or adjectives).

The distribution of reviews in the training, development and test sets is provided in Table 3, as well as the distribution of the different negation structures per dataset. The total of positive and negative reviews can be seen in the rows named as *+ Reviews* and *- Reviews*, respectively.

Total	Training	Devel.	Test
Reviews	264	56	80
+ Reviews	134	22	44
- Reviews	130	34	36
neg	2.511	594	836
noneg	104	22	55
contrast	100	23	52
comp	18	6	6

Table 3: Distribution of reviews and negation cues in the datasets of Task 3

#### 4 Evaluation measures

The evaluation script used to evaluate **Task 2** was the same used to evaluate the \*SEM 2012 Shared Task: “Resolving the Scope and Focus of Negation” (Morante and Blanco, 2012). It is based on the following criteria:

- Punctuation tokens are ignored.
- A True Positive (TP) requires all tokens of the negation element have to be correctly identified.
- To evaluate cues, partial matches are not counted as False Positive (FP), only as False Negative (FN). This is to avoid penalizing partial matches more than missed matches.

The measures used to evaluate the systems were Precision (P), Recall (R) and F-score (F1). In the proposed evaluation, FN are counted either by the system not identifying negation elements present in the gold annotations, or by identifying them partially, i.e., not all tokens have been correctly identified or the word forms are incorrect. FP are counted when the system produces a negation element not present in the gold annotations and TP are counted when the system produces negation elements exactly as they are in the gold annotations.

For evaluating **Task 3**, the traditional measures used in text classification were applied: Precision (P), Recall (R), F-score (F1) and Accuracy (Acc). P, R and F-score were measured per class and averaged using macro-average method.

$$P = \frac{TP}{TP + FP} \quad (1)$$

$$R = \frac{TP}{TP + FN} \quad (2)$$

$$F1 = \frac{2PR}{P + R} \quad (3)$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

#### 5 Participants

10 teams showed interested and 4 teams submitted results.

**Task 1** had two participants: the CLiC team composed of M. Antonia Martí and Mariona Taulé from the University of Barcelona, and Lucia Donatelli from the Georgetown University.

Martí and Taulé (2018) carry out an analysis of 5 fundamental aspects of the corpora analyzed: i) the negation cue, ii) the scope and the inclusion of the subject in the scope, iii) the coordinated structures, iv) the negative locutions and v) the lexical and morphological negation. Taking into account the differences observed in the annotation of the corpora, they proposed the following guidelines:

- Annotate the negation cue whenever possible, as it will allow to use it whenever necessary or to ignore it otherwise. Moreover, they consider that it should be distinguished between simple markers (e.g. “no” [no/not], “sin” [without]) and complex markers (e.g. “no...nadie” [no...nobody]), where one implies the presence of the other. They propose to make use of the typology defined for the annotation of the SFU ReviewSP-NEG corpus.
- Annotate the scope including the subject within it. They mention that in many cases the focus of negation corresponds to the subject and this would facilitate future annotations of the corpus.
- Perform coordinated negation treatment. They propose to distinguish between coordinated structures affected by the same predicate and negation marker [2] and coordinated structures with independent negation cues and predicates [3], so that in the first case a single negation marker is considered and the rest of the negation structure as scope and, in the second case, a separate scope is annotated for each coordinated negation marker.

- Annotate negative locutions (e.g. “en absoluto” [*not at all*]), even if they do not contain explicit negation markers.
- Annotate lexical and morphological negation, which have only been addressed restrictively in the UHU-HUVR and IULA Spanish Clinical Record corpora.
- Annotate the focus of negation, which is not deal with in any of the guidelines analyzed.

[2] No [es ni muy pesado ni muy ligero] (SFU Review<sub>SP-NEG</sub>)

[3] No [soy muy alta] tampoco [un pitufo] (SFU Review<sub>SP-NEG</sub>)

Donatelli (2018) describes each corpus individually and indicates which elements are missing in the annotation of each of them and those aspects that should have been taken into account. She considers that some components of the different guidelines can be combined in order to set linguistically precise guidelines and neutral guidelines with regard to the domain. She indicates that in order to represent the semantic of negation, the following elements must be annotated:

- The negation cue: lexical item that expresses negation.
- The scope: part of the text that is negated.
- The focus: part of the scope that is prominently or explicitly negated.
- The reinforcement (if exists): auxiliary negation or element of negative polarity, known as NPI (Negative Polarity Item) (Altuna, Minard, and Speranza, 2017).

Below we can see, in an example provided by the author [4], the different elements explained above. The negation cue appears in bold, the scope in brackets, the focus in italics, and the reinforcement underlined.

[4] John **no** [come *carne* sino verduras].

Donatelli considers that the scheme proposed by Jiménez-Zafra et al. (2018) for the annotation of the SFU Review<sub>SP-NEG</sub> corpus is suitable for capturing the layers of

negation complexity and proposes to combine it with the use of the label *NegPolItem* used by Marimon, Vivaldi, and Bel (2017) in the annotation of the IULA Spanish Clinical Record corpus to annotate items of negative polarity (NPI) or auxiliary negations.

**Task 2** also had two participants: the UPC team composed of Henry Loharja, Lluís Padró and Jordi Turmo from the Universitat Politècnica de Catalunya (Loharja, Padró, and Turmo, 2018), and Hermenegildo Fabregat, Juan Martínez-Romo and Lourdes Araujo from the National Distance Education University of Spain (UNED) (Fabregat, Martínez-Romo, and Araujo, 2018). The official results are shown in Tables 4, 5 and 6, evaluated in terms of P, R and F1.

Domain	P	R	F1
Cars	94.23	72.06	81.67
Hotels	97.67	71.19	82.35
Washing machines	92.00	66.67	77.31
Books	79.52	66.27	72.29
Cell phones	93.33	73.68	82.35
Music	92.59	57.47	70.92
Computers	-	-	-
Movies	86.26	69.33	76.87
Total	79.45	59.58	67.97

Table 4: Official results by domain for the UNED team

Domain	P	R	F1
Cars	95.08	85.29	89.92
Hotels	94.00	79.66	86.24
Washing machines	94.74	78.26	85.72
Books	84.19	84.52	84.35
Cell phones	89.80	77.19	83.02
Music	92.96	75.86	83.54
Computers	91.36	91.36	91.36
Movies	89.68	85.28	87.42
Total	91.48	82.18	86.45

Table 5: Official results by domain for the UPC team

The results by domain, described in Tables 4 and 5, show that there are sub-collections such as books and music in which both systems obtain worse results compared to the rest of the sub-collections. The system developed by the UNED team obtains the highest performance in cell phones and hotels sub-collections, while the UPC system

shows a better detection of negation cues in the computers sub-collection, in particular, it obtains an F1 of 91.36%.

The overall results presented in Table 6 correspond to the performances without considering the computers subset, since the UNED team could not submit the results for computers due to technical problems. In terms of overall performance, both systems obtain similar precision. However, the recall achieved by the UNED system is lower. Therefore, the best result is obtained by the UPC team with an F1 of 85.74%.

Team	P	R	F1
UNED	90.80	68.10	77.68
UPC	91.49	80.87	85.74

Table 6: Overall official results for Task 2

In terms of the approaches applied, both proposals use the standard labelling scheme *BIO* where the first word of a negation structure denotes by *B* and the remaining words by *I*. The label *O* indicates that the word does not correspond with a negation cue.

The UNED team applies a model of deep learning inspired by named entity recognition architectures and negation detection models. Specifically, this system is focused on the use of several neural networks together with a bidirectional LSTM (Long Short-Term Memory). This supervised approach is based on pretrained word embeddings for Spanish. For its part, the UPC team uses Conditional Random Fields with a set of features such as the part-of-speech of the word and information about how the words are written.

Finally, the resources used by the participants are diverse. The UNED team uses Keras (Chollet and others, 2015) and TensorFlow (Abadi et al., 2016) libraries, as well as pretrained word embeddings for Spanish (Cardellino, 2016), and the UPC team uses NLTK (Loper and Bird, 2002).

**Task 3** had no participants. Some of the teams registered for the workshop showed interest in the task, but expressed that they did not participate due to lack of time.

## 6 Conclusions

This paper presents the description of the 2018 edition of NEGES, which consisted of three different tasks related to different aspects of negation: Task 1 on reaching an

agreement on the guidelines to follow for the annotation of negation in Spanish, Task 2 on identifying negation cues, and Task 3 on evaluating the role of negation in sentiment analysis. The SFU ReviewSP-NEG corpus was the collection of documents used to train and test the systems presented in Task 2 and Task 3. As far as we know, this is the first task that focuses on the development and evaluation of systems for identifying negation cues in Spanish in the area of sentiment analysis.

A total of 4 teams participated in the workshop, 2 for developing annotation guidelines and 2 for cues detection. Task 3 had no participants. For a future edition of the workshop we would like to continue working on the unification of the annotation schemes and propose different tasks to detect negation in other domains such as biomedical.

## 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. RM is supported by the Netherlands Organization for Scientific Research (NWO) via the Spinoza-prize awarded to Piek Vossen (SPI 30-673, 2014-2019).

## References

- Abadi, M., P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. 2016. Tensorflow: a system for large-scale machine learning. In *Proceedings of OSDI 2016*, volume 16, pages 265–283.
- Altuna, B., A.-L. Minard, and M. Speranza. 2017. The Scope and Focus of Negation: A Complete Annotation Framework for Italian. In *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*, pages 34–42.
- Buchholz, S. and E. Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the tenth conference on computational natural language learning*, pages 149–164.
- Cardellino, C. 2016. Spanish Billion Words Corpus and Embeddings. <https://crscardellino.github.io/SBWCE/>. [Accessed 29-January-2019].

- Chollet, F. et al. 2015. Keras. <https://keras.io>. [Accessed 29-January-2019].
- Cruz Díaz, N., R. Morante Vallejo, M. J. Maña López, J. Mata Vázquez, and C. L. Parra Calderón. 2017. Annotating Negation in Spanish Clinical Texts. In *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*, pages 53–58.
- Donatelli, L. 2018. Cues, Scope, and Focus: Annotating Negation in Spanish Corpora. In *Proceedings of NEGES 2018: Workshop on Negation in Spanish, CEUR Workshop Proceedings*, volume 2174, pages 29–34.
- Fabregat, H., J. Martínez-Romo, and L. Araujo. 2018. Deep Learning Approach for Negation Cues Detection in Spanish at NEGES 2018. In *Proceedings of NEGES 2018: Workshop on Negation in Spanish, CEUR Workshop Proceedings*, volume 2174, pages 43–48.
- Farkas, R., V. Vincze, G. Móra, J. Csirik, and G. Szarvas. 2010. The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*, pages 1–12.
- Jiménez-Zafra, S. M., M. Taulé, M. T. Martín-Valdivia, L. A. Ureña-López, and M. A. Martí. 2018. SFU Review SP-NEG: a Spanish corpus annotated with negation for sentiment analysis. a typology of negation patterns. *Language Resources and Evaluation*, 52(2):533–569.
- Konstantinova, N., S. C. De Sousa, N. P. C. Díaz, M. J. M. López, M. Taboada, and R. Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In *Proceedings of LREC 2012*, pages 3190–3195.
- Loharja, H., L. Padró, and J. Turmo. 2018. Negation Cues Detection Using CRF on Spanish Product Review Text at NEGES 2018. In *Proceedings of NEGES 2018: Workshop on Negation in Spanish, CEUR Workshop Proceedings*, volume 2174, pages 49–54.
- Loper, E. and S. Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, pages 63–70.
- Marimon, M., J. Vivaldi, and N. Bel. 2017. Annotation of negation in the iula spanish clinical record corpus. In *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*, pages 43–52.
- Martí, M. A. and M. Taulé. 2018. Análisis Comparativo de los Sistemas de Anotación de la Negación en Español. In *Proceedings of NEGES 2018: Workshop on Negation in Spanish, CEUR Workshop Proceedings*, volume 2174, pages 23–28.
- Martí, M. A., M. Taulé, M. Nofre, L. Marsó, M. T. Martín-Valdivia, and S. M. Jiménez-Zafra. 2016. La negación en español: análisis y tipología de patrones de negación. *Procesamiento del Lenguaje Natural*, (57):41–48.
- Morante, R. and E. Blanco. 2012. \* SEM 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 265–274.
- Morante, R. and C. Sporleder. 2010. Proceedings of the workshop on negation and speculation in natural language processing. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 1–109.
- Oronoz, M., K. Gojenola, A. Pérez, A. D. de Ilarraza, and A. Casillas. 2015. On the creation of a clinical gold standard corpus in spanish: Mining adverse drug reactions. *Journal of biomedical informatics*, 56:318–332.
- Padró, L. and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of LREC 2012*, Istanbul, Turkey, May.
- Sandoval, A. M. and M. G. Salazar. 2013. La anotación de la negación en un corpus escrito etiquetado sintácticamente. *Revista Iberoamericana de Lingüística: RIL*, (8):45–60.
- Taboada, M., C. Anthony, and K. D. Voll. 2006. Methods for creating semantic orientation dictionaries. In *Proceedings of LREC 2016*, pages 427–432.

# Creación de un corpus de noticias de gran tamaño en español para el análisis diacrónico y diatópico del uso del lenguaje

## *Creation of a large news corpus in Spanish for the diachronic and diatopic analysis of the use of language*

Pavel Razgovorov<sup>1</sup>, David Tomás<sup>2</sup>

<sup>1</sup>Ingeniería Informática Empresarial (i2e), C/ Auso y Monzó 16, 03006, Alicante (España)

<sup>2</sup>Universidad de Alicante, C/ San Vicente del Raspeig s/n, 03690, Alicante (España)  
pavel.razgovorov@i2e.es, dtomas@dlsi.ua.es

**Resumen:** Este artículo describe el proceso llevado a cabo para desarrollar un corpus de noticias periodísticas de gran tamaño en español. Todos los textos recopilados están ubicados tanto temporal como geográficamente. Esto lo convierte en un recurso de gran utilidad para trabajos en el ámbito de la lingüística, la sociología y el periodismo de datos, permitiendo tanto el estudio diacrónico y diatópico del uso del lenguaje como el seguimiento de la evolución de determinados eventos. El corpus se puede descargar libremente empleando el software que se ha desarrollado como parte de este trabajo. El artículo se completa con un análisis estadístico del corpus y con la presentación de dos casos de estudio que muestran su potencial a la hora de analizar sucesos.

**Palabras clave:** Corpus, minería de texto, análisis diacrónico, análisis diatópico

**Abstract:** This article describes the process carried out to develop a large corpus of news stories in Spanish. The collected texts are located both temporally and geographically. This makes it a very useful resource to work with in the field of linguistics, sociology and data journalism, allowing the diachronic and diatopic study of the use of language and tracking the evolution of specific events. The corpus can be freely downloaded using the software developed as part of this work. The article includes a statistical analysis of the corpus and two case studies that show its potential for event analysis.

**Keywords:** Corpus, text mining, diachronic analysis, diatopic analysis

## 1 Introducción

Las noticias periodísticas cada vez se producen y consumen de manera más frecuente a través de Internet, favoreciendo la existencia de grandes volúmenes de este tipo de textos en formato digital. El análisis computacional de estos corpus de noticias puede llevar al descubrimiento de interesantes hallazgos desde un punto de vista tanto sociológico como lingüístico (Leetaru, 2011).

Este artículo presenta el desarrollo de un corpus de noticias en español de gran tamaño ubicadas tanto geográfica (lugar en el que se produjeron) como temporalmente (momento en el que tuvieron lugar). El objetivo de este corpus es el de servir de fuente para estudios en áreas como la lingüística, la sociología y el

periodismo de datos, permitiendo un análisis diacrónico (evolución en el tiempo) y diatópico (evolución en el espacio) del uso del lenguaje. Además de describir las características del corpus y el proceso llevado a cabo para su obtención, este trabajo presenta ejemplos de casos de estudio centrados en su explotación.

La fuente de información empleada para la construcción de este corpus ha sido el periódico gratuito de información general *20 minutos*,<sup>1</sup> el cual ofrece en formato digital todas sus noticias desde enero de 2005 hasta la actualidad. Este periódico contiene secciones locales para todas y cada una de las provincias de España, lo que permite tener localizadas geográficamente cada una de las noticias. Pa-

<sup>1</sup><https://www.20minutos.es>

ra el desarrollo de este corpus se han extraído todas las noticias publicadas en la sección local de cada una de las cincuenta provincias y las dos ciudades autónomas de Ceuta y Melilla, dando lugar a un corpus de casi dos millones de noticias publicadas a lo largo de los últimos trece años.

Además del contenido original de las noticias, se ha realizado un análisis lingüístico del texto incorporando información sobre lemas, entidades y palabras con contenido semántico, codificando toda esta información en formato JSON para facilitar su posterior procesamiento. Aunque el corpus no puede distribuirse libremente por los derechos de uso de *20 minutos*, se proporciona el software necesario para que cualquier investigador pueda descargar y replicar este corpus en su ordenador de manera sencilla.<sup>2</sup>

La gran ventaja que ofrece este corpus frente a otros existentes es la localización geográfica de los artículos, lo que permite realizar análisis diatópicos del lenguaje y estudiar los rasgos dialectales de distintas regiones. Hasta donde tenemos conocimiento, no existe en la actualidad un corpus de noticias con estas características libremente disponible.

El resto del artículo se estructura como sigue: la Sección 2 presenta diversos trabajos en el ámbito del desarrollo de corpus periodísticos; la Sección 3 describe el proceso llevado a cabo para la generación del corpus; en la Sección 4 se presentan distintas estadísticas extraídas del corpus y detalles sobre el uso del lenguaje a nivel geográfico y temporal; en la Sección 5 se ofrece un análisis cuantitativo de dos sucesos explotando la información contenida en el corpus; finalmente, la Sección 6 muestra las conclusiones y posibles trabajos futuros.

## 2 Trabajo relacionado

Existen diferentes estudios y proyectos que tratan sobre la construcción de corpus de textos periodísticos para su posterior análisis lingüístico. En esta sección nos vamos a centrar en revisar los trabajos realizados en idioma español y, por tanto, más afines con el corpus descrito en este artículo.

El primero de estos corpus es el recopilado en el proyecto Aracne,<sup>3</sup> donde se pre-

senta un estudio sobre la variación de la riqueza lingüística en la prensa española desde 1914 hasta 2014. Sobre este corpus de noticias se realizó un procesamiento lingüístico de los textos midiendo rasgos de variación léxica, densidad y complejidad de los textos. El resultado de este proyecto fue un corpus de 5.167 artículos y 1.921.566 de palabras. El corpus no está disponible para su estudio.

Otro trabajo en esta línea es el corpus *Spanish News Text* (Graff y Gallegos, 1995), compuesto de textos periodísticos extraídos de diferentes periódicos y agencias de noticias de hispanoamérica entre el año 1995 y 1996. El corpus cuenta con 170 millones de palabras y está disponible para los miembros del *Linguistic Data Consortium*<sup>4</sup> (LDC). Existe una versión posterior del corpus, el *Spanish Newswire Text, Volume 2* (Graff y Gallegos, 1999), que recoge noticias entre los años 1996 y 1998. Esta versión está disponible para el público en general previo pago.

El corpus *Timestamped JSI web*<sup>5</sup> está formado por artículos de noticias obtenidos de distintos servicios RSS a nivel mundial. Contiene textos desde el año 2014 hasta la actualidad en diferentes idiomas, entre ellos el español. El corpus está accesible para usuarios suscritos a la plataforma *Sketchengine* y sólo puede consultarse dentro de ésta, ofreciendo funcionalidades para análisis del lenguaje como la búsqueda de sinónimos, ejemplos de uso, frecuencia de palabras o identificación de neologismos.

Otro corpus relevante es el *Corpus del Español NOW (News on the Web)*<sup>6</sup> que contiene cerca de 5.700 millones de palabras obtenidos de periódicos digitales y revistas de 21 países de habla hispana, desde el año 2012 hasta la actualidad. Al igual que el corpus anterior, su acceso está limitado a la plataforma en la que se oferta, en la que se pueden realizar diferentes consultas a través de una interfaz.

Finalmente, *Molino Labs*<sup>7</sup> presenta un corpus formado a partir de artículos de prensa de España, Argentina y México, producidas entre los años 1997 y 2009. Cuenta con más de 1.700 millones de artículos y 660 millones de palabras. El corpus se puede consultar a través de una interfaz web, pero no

<sup>4</sup><https://www ldc.upenn.edu>

<sup>5</sup><https://www.sketchengine.eu/>

jozef-stefan-institute-newsfeed-corpus

<sup>6</sup><https://www.corpusdelespanol.org>

<sup>7</sup><http://www.molinolabs.com/corpus.html>

<sup>2</sup><https://github.com/analisis-20minutos/herramientas-analisis>

<sup>3</sup><https://www.funfeu.es/aracne>

está disponible para descarga. Esta interfaz ofrece, entre otras, la posibilidad de buscar palabras de una longitud fija que empiezan por determinados caracteres o localizar palabras que aparecen en compañía de otras.

Si bien algunos de estos trabajos permiten hacer un estudio diacrónico del corpus, a diferencia de nuestra propuesta ninguno de ellos ofrece la posibilidad de hacer un estudio de los textos a nivel diatópico. Adicionalmente, existe el problema del acceso a los documentos. La mayoría de las propuestas estudiadas sólo permiten analizar el corpus a través de una interfaz web con funciones limitadas, mientras que otras ofrecen su descarga solo a suscriptores o previo pago. En el caso de nuestro corpus, aunque por limitaciones de derechos de uso no se pueda distribuir libremente, se ha publicado el software desarrollado para que cualquiera pueda descargarlo y replicarlo en su ordenador. Finalmente, otro de los puntos fuertes de nuestro corpus frente a otras propuestas es el volumen de noticias que presenta, la información de análisis lingüístico incluida y el contar con noticias actuales (ver Sección 4).

### 3 Creación del corpus

En esta sección se describe todo el proceso de obtención del corpus. En primer lugar se describirá el proceso de descarga y limpieza de las noticias. A continuación se expone el proceso llevado a cabo para la eliminación de duplicados y casi duplicados. Finalmente, se describe el análisis lingüístico llevado a cabo sobre los documentos.

#### 3.1 Obtención de noticias

La primera tarea llevada a cabo para la creación del corpus fue la obtención de las noticias en formato digital de la hemeroteca del diario *20 minutos*. Las noticias están accesibles a través de la sección *Archivo*<sup>8</sup> del diario donde se encuentran agrupadas por días. Se puede acceder a la página web de cada una de ellas pinchando en el correspondiente enlace, pero no existe una forma directa de descargar el contenido de la noticia.

Para poder obtener el texto limpio de las noticias se tuvo que desarrollar un programa para la extracción de datos de las páginas web (*web scraping*). El objetivo era obtener el título, resumen y cuerpo de las noticias, eliminando todas las etiquetas HTML y conte-

nidos adicionales que se muestran en la página del diario (ej. menús, anuncios, noticias relacionadas e imágenes). Para dicha tarea se utilizó la herramienta *Scrapy*,<sup>9</sup> que permite extraer elementos de las páginas web mediante la definición de selectores CSS.

La principal dificultad de esta tarea fue la falta de homogeneidad en la estructura HTML de las páginas. Ésta variaba en función de los años y de algunas localizaciones, por lo que se tuvieron que realizar ajustes en la herramienta para contemplar estas variaciones. Se revisaron manualmente y de manera sistemática subconjuntos de noticias en las distintas localizaciones y años para comprobar que la obtención del contenido de todas ellas fuera correcto.

Como resultado de este proceso se obtuvo un volcado completo de las noticias del portal desde el 17 de enero de 2005, primer día del que se tiene registro, hasta el 5 de julio de 2018, momento en el que se finalizó la recolección de datos. Para cada noticia se almacenó en formato JSON, tal y como se puede ver en la Figura 1, su localidad (*province*), fecha en formato ISO 8601 (*date*), URL de la página de donde fue extraída (*url*), título (*title*), resumen (*lead*) y cuerpo (*body*). Para estos tres últimos campos, además del texto original (*raw\_text*), se incluyó una serie de información adicional resultado del análisis lingüístico realizado, tal y como se describe en la Sección 3.3.

En total se obtuvieron 2.215.078 artículos de 52 localizaciones diferentes: las cincuenta provincias de España y las dos ciudades autónomas de Ceuta y Melilla.

#### 3.2 Eliminación de duplicados

Dado que *20 minutos* cuenta con numerosas secciones locales, era de esperar que algunas noticias pudieran estar duplicadas o casi duplicadas entre distintas provincias. Por ejemplo, en ocasiones se usan noticias “plantilla” en las que sólo cambian los datos específicos correspondientes a cada localidad. Es el caso de eventos como “La fiesta del cine”, donde lo único que varía de una provincia a otra es el número de asistentes. Otros ejemplos de noticias casi duplicadas son aquellas que presentan actualizaciones sobre una noticia anterior. Dentro de este grupo entran casos como el de las crecidas del Ebro, donde cada día se publican datos actualizados

<sup>8</sup><https://www.20minutos.es/archivo>

<sup>9</sup><https://scrapy.org>

```

{
  "province": "SEVILLA",
  "date": "2005-01-28T00:00:00",
  "url": "https://www.20minutos.es/noticia/1994/0/sevilla/bajo/cero/",
  "title": {
    "raw_text": "Y en Sevilla, bajo cero",
    "lemmatized_text": "y en sevilla bajo 0",
    "lemmatized_text_reduced": "sevilla",
    "persons": [], "locations": ["Sevilla"], "organizations": [], "dates": [],
    "numbers": ["0"], "others": []
  },
  "lead": {
    ...
  },
  "body": {
    ...
  }
}

```

Figura 1: Ejemplo de noticia del corpus en formato JSON. Los campos contenidos en `title` se encuentran también en `lead` y `body`. Se omiten aquí por motivos de espacio

utilizando el mismo cuerpo de noticia. Finalmente, también hay noticias de interés global que simplemente se duplican literalmente entre distintas localidades.

Para evitar textos repetidos que puedan afectar al análisis estadístico del corpus, se desarrolló un programa para la eliminación de duplicados y casi duplicados. El problema inicial de este proceso era la imposibilidad de cotejar cada noticia del corpus con todas las demás, ya que implicaba más de cuatro billones de comparaciones. Una de las posibilidades que se planteó para reducir este número fue la de comparar noticias sólo entre provincias cercanas. Sin embargo, tras un análisis realizado sobre un subconjunto de ellas, se comprobó que existían noticias duplicadas en localidades muy distantes. Lo que sí permitió este estudio preliminar fue identificar un patrón temporal en las noticias duplicadas: todas ellas se daban en el mismo día o en días muy próximos. Para dar margen suficiente, se decidió comparar las noticias en bloques de sesenta días estableciendo un solapamiento de seis días entre bloques consecutivos.

Para acelerar aún más el proceso de comparación entre pares de noticias, se utilizó una estructura de tipo *MinHash* (Broder, 1997) que permitía calcular la similitud de Jaccard (Leskovec, Rajaraman, y Ullman, 2014) entre dos textos en un tiempo lineal con respecto al tamaño del conjunto de documentos a comparar. En la implementación se empleó *Locality-Sensitive Hashing* (LSH) (Indyk

y Motwani, 1998) para optimizar el proceso de comparación entre documentos. De esta manera se consiguió un algoritmo de complejidad  $O(n \cdot \sqrt{n})$  para la búsqueda de duplicados, reduciendo la complejidad de tipo  $O(n^3)$  que hubiera tenido de no haber aplicado estas optimizaciones.

Una vez establecido el procedimiento de comparación, era necesario determinar cuándo dos noticias se iban a considerar como duplicadas. Se probaron distintos porcentajes de solapamiento a nivel de palabra, comprobando el número de noticias que eran consideradas como duplicadas. Finalmente este umbral de solapamiento se estableció en un 70%, ya que con este valor se alcanzaba estabilidad en el número de noticias eliminadas, y para valores menores se consideraba que podía llevar a la obtención de falsos duplicados. El número total de noticias que quedó como resultado de este proceso fue de 1.826.985, eliminando en total casi cuatrocientas mil noticias.

### 3.3 Procesamiento lingüístico

Con el objetivo de enriquecer el corpus con información lingüística que permitiera un análisis más profundo de los textos, se procedió a la extracción de distintas características de éste para incorporarlas a la estructura JSON desarrollada. Concretamente, se extrajeron los siguientes elementos de cada noticia para el título, cuerpo y resumen (ver ejemplo en la Figura 1):

- Texto lematizado, sin signos de puntuación

ción (`lemmatized_text`).

- Texto lematizado sólo de adjetivos, nombres, verbos y adverbios acabados en “mente” (`lemmatized_text_reduced`). El objetivo es mantener aquí sólo aquellos términos que tienen valor semántico y aportan significado al texto.
- Lista de entidades nombradas: organizaciones (`organizations`), personas (`persons`), lugares (`locations`), fechas (`dates`), números (`numbers`) y otras (`others`).

Para obtener esta información se utilizó la herramienta FreeLing (Padró y Stanilovsky, 2012). Después de realizar unas pruebas de rendimiento, se estimó que el procesado del corpus llevaría aproximadamente 108 horas de ejecución. Para reducir este tiempo se paralelizó el análisis de las noticias mediante la librería *OpenMP*,<sup>10</sup> consiguiendo una reducción del 72 % en el tiempo de procesamiento.

#### 4 Análisis del corpus

Sobre el corpus creado se han realizado una serie de análisis estadísticos para determinar la riqueza léxica de los textos, estudiar la evolución temporal del lenguaje y también su uso en distintas zonas geográficas. Los siguientes apartados de esta sección profundizan en cada uno de estos aspectos.

##### 4.1 Estadísticas generales

Como se ha comentado anteriormente, el corpus final cuenta con un total de más de 1.8 millones de noticias (tras la eliminación de duplicados) y un tamaño cercano a los 20 GB después de incorporarle la información resultante del procesamiento con FreeLing. La Figura 2 muestra un mapa coroplético con el número de noticias publicadas en cada provincia para el total del corpus. Se ha creado una infografía interactiva completa, accesible en Internet, donde se puede ver el número exacto de noticias por provincia y su evolución a lo largo de los años.<sup>11</sup>

Las provincias con mayor número de noticias son Sevilla (176.903), seguida de Valencia (100.455) y Cantabria (92.268), mientras que la que menos tiene es Ceuta (2.789). El lugar con mayor número de noticias en un

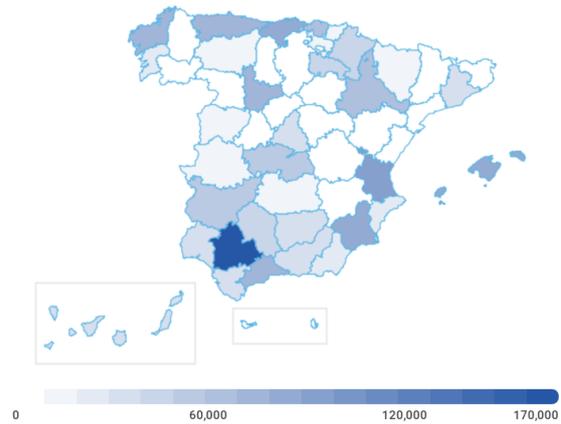


Figura 2: Número de noticias por provincia para el total de años analizado

único día es Murcia, el 12 de mayo de 2011, con 164 artículos (este punto se describe con más detalle en la Sección 5.1). La media de noticias por localización es de 35.134,33, mientras que su desviación estándar es de  $\pm 33.989,52$ , reflejando éste último dato una diferencia notable en el número de publicaciones entre distintas ediciones locales.

En cuanto a la distribución de noticias por años, los periodos de mayor actividad fueron 2011 (200.599), 2014 (199.721) y 2013 (199.334), mientras que los que menos publicaciones tuvieron fueron 2005 (26.845), 2009 (28.203) y 2006 (42.809). La media por años se sitúa en 130.498,93 noticias con una desviación estándar de  $\pm 73.869,20$ , situación análoga al análisis anterior por provincia.

Por lo que respecta al número de palabras del corpus, éste se compone de un total de 711.840.945 términos. Para medir la riqueza léxica del corpus analizamos el *Type-Token Ratio* (TTR) (Holmes, 1985), que representa el cociente entre el número de palabras diferentes que contiene un texto y el número total de palabras de ese texto. Concretamente, calculamos el TTR de cada noticia de manera individual, obteniendo posteriormente para cada año y localización el TTR promedio de todas sus noticias. En la Figura 3 se puede ver la evolución del TTR a lo largo de los años, teniendo en cuenta tanto el vocabulario total (etiqueta *TTR* en el gráfico) como el que incluye únicamente palabras con valor semántico (gráfico *TTR reducido*). En la gráfica se observa una pérdida de riqueza léxica con el paso de los años, desde 2005 hasta 2011, estabilizándose posteriormente con un

<sup>10</sup><https://www.openmp.org>

<sup>11</sup><http://cort.as/-C56M>, accedida en noviembre de 2018.

repunte de la riqueza en los últimos tres años estudiados.

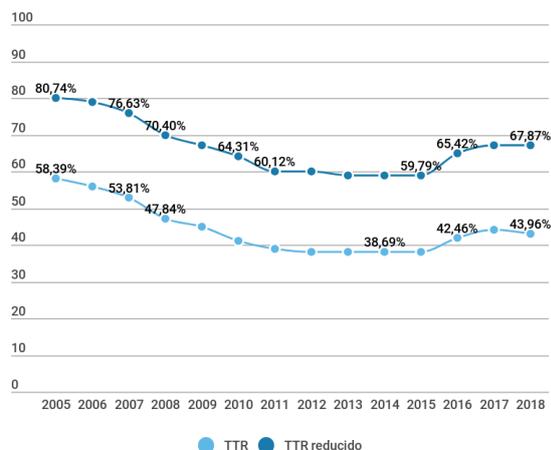


Figura 3: Evolución del TTR a lo largo de los años, tanto para el léxico completo (*TTR*) como para el subconjunto de los términos con valor semántico (*TTR reducido*)

Este cálculo se llevó a cabo también a nivel de localización geográfica.<sup>12</sup> El estudio revela que las provincias con mayor TTR son Alicante (69,08%), Pontevedra (68,46%) y Madrid (67,91%). En la otra cara de la moneda, Toledo (60,72%), Castellón (60,81%) y Sevilla (61,45%) presentan los índices más bajos de riqueza léxica.

## 4.2 Análisis diacrónico

El objetivo del estudio llevado a cabo en este apartado es averiguar, para cada año, cuáles han sido los temas más populares entre las publicaciones del periódico a partir del léxico empleado. Para la obtención de los términos más populares se tuvieron en cuenta solo las palabras con contenido semántico, tal y como fueron descritas en la Sección 3.3. Se ha realizado una infografía donde se pueden apreciar en forma de nube de palabras los cincuenta términos (lemas) más importantes y su frecuencia de aparición para cada anualidad.<sup>13</sup>

La Figura 4 muestra como ejemplo la nube de palabras correspondiente al año 2017. Hay términos esperables en esta lista de palabras más populares como “gobierno”, “ayuntamiento”, “PP” (muy por encima de otros partidos políticos) o “presidente”. Llama la

<sup>12</sup>Todos los detalles están disponibles en la siguiente infografía: <http://cort.as/-C56R>, accedida en noviembre de 2018.

<sup>13</sup><http://cort.as/-C56W>, accedida en noviembre de 2018.

atención la frecuencia de uso de la palabra “persona”, hecho que puede deberse a la neutralidad que presenta a nivel de género, siendo por ello muy utilizada por los periodistas para producir textos que contengan un lenguaje inclusivo y no sexista.



Figura 4: Nube de palabras con los cincuenta términos (lemas) más frecuentes en el año 2017

## 4.3 Análisis diatópico

De manera análoga al apartado anterior, se procedió a realizar un estudio de los términos más usados en las distintas localizaciones estudiadas, obteniendo los cincuenta términos más frecuentes en cada una de ellas. También se evaluaron las entidades más populares que habían sido extraídas por FreeLing, tal y como se describe en la Sección 3.3.<sup>14</sup> Toda esta información está incluida en la infografía mencionada en el apartado anterior.

Como era de esperar, entre las palabras más repetidas para cada lugar está el propio nombre de la provincia y su gentilicio, junto con la comunidad autónoma a la que pertenece. Es destacable también la aparición de partidos políticos locales (como “BNG” en A Coruña, “PNV” en País Vasco y “Foro [Asturias]” en Asturias) y la preponderancia que tienen algunos grupos en determinadas regiones (en Cádiz tanto “PP” como “PSOE” son términos muy frecuentes) frente a otras en las que desaparecen (por ejemplo, en Barcelona no aparecen ninguno de estos dos entre los cincuenta términos más frecuentes).

En aquellas regiones en las que existen lenguas propias, se pueden encontrar entre los

<sup>14</sup>A modo de referencia, el rendimiento de FreeLing clasificando entidades se evaluó revisando manualmente 10 artículos al azar, con un total de 4.960 términos y 378 entidades, obteniendo un 84,92% de precisión en esta tarea.

más frecuentes términos muy representativos de dichos lugares, como “generalitat” en Valencia, “xunta” en A Coruña, “euskadi” en Vizcaya o “mossos” en Barcelona.

## 5 Casos de estudio

En esta sección se muestran dos casos de ejemplo del tipo de información que se puede obtener del corpus desarrollado mediante el análisis de texto. Una de las aplicaciones interesantes de este corpus es su uso para el periodismo de datos (Gray, Chambers, y Bou-negru, 2012), una especialidad del periodismo que refleja el creciente valor de los datos en la producción y distribución de información, cuyo objetivo es recabar gran cantidad de datos y hacer la información comprensible a la audiencia ayudándose de herramientas como las infografías, representaciones gráficas o aplicaciones interactivas.

En el primero de los casos de estudio se analiza el terremoto que tuvo lugar en Lorca el 12 de mayo de 2011, mientras que en el segundo estudio se aborda el tema de la independencia de Cataluña a lo largo de los dos últimos años.

### 5.1 Terremoto de Lorca

Tal y como se comentó en la Sección 4.1, la provincia que más noticias tuvo en un único día fue Murcia, el 12 de mayo de 2011, con 164 artículos. El suceso que provocó este aluvión de información fue el seísmo ocurrido el día anterior en la localidad de Lorca con una magnitud de 5,1 en la escala Richter, causando numerosos daños materiales, más de 300 heridos y 9 muertos. Para recuperar las noticias relacionadas con este evento, se localizaron todas aquellas que contenían el término “Lorca” junto con “terremoto” o “seísmo”.

Para analizar el suceso partiendo del corpus desarrollado, se ha creado una nueva infografía<sup>15</sup> para identificar los términos más relevantes asociados con esta noticia, los lugares en los que más se habló del tema (desde el punto de vista del número de noticias publicadas) y un estudio de la repercusión de la noticia a lo largo del tiempo, identificando el momento en el que los medios redujeron su interés en este suceso. Este último aspecto se puede apreciar en la Figura 5.

Cabe destacar las posibilidades que ofrece el corpus como herramienta para determinar

<sup>15</sup><http://cort.as/-C56d>, accedida en noviembre de 2018.

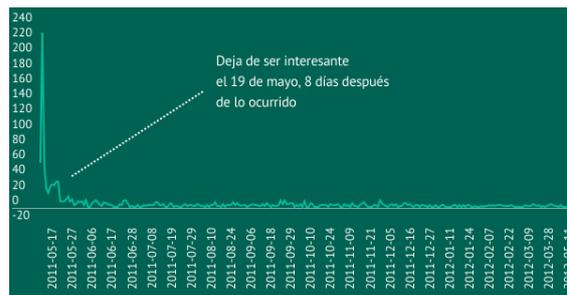


Figura 5: Seguimiento periodístico del terremoto de Lorca durante un año, mostrando el número de noticias publicadas cada día

el interés en el tiempo que produce una noticia. En el caso de ésta, se aprecia como una semana después del seísmo su interés decreció notablemente en cuanto al número de publicaciones relacionadas, mostrando la limitada vida que puede tener en los medios de comunicación un evento de esta magnitud.

### 5.2 Independencia de Cataluña

En este apartado se presenta un segundo estudio siguiendo las pautas del anterior, con el foco puesto en un tema de relevancia política nacional como es la independencia de Cataluña. Los términos que se emplearon para localizar noticias relacionadas con esta temática fueron: “independencia”, “independència”, “independentismo”, “independentisme” e “independentista”. El periodo de estudio se fijó desde el 1 de enero de 2017 hasta el 5 de julio de 2018 (último día en el que se incorporaron noticias al corpus).

Al igual que para el caso del terremoto de Lorca, se ha diseñado una infografía que hace un análisis temporal y espacial de la cobertura del suceso en el corpus.<sup>16</sup> En este análisis se pueden observar datos destacables, como que la cobertura de este tema en el periódico *20 minutos* se ha llevado a cabo de manera prácticamente ininterrumpida en el año y medio a estudio: solo en 6 días de los 445 analizados no se produjo ninguna noticia relacionada con este asunto. También llama la atención detalles como que en Ceuta y Melilla no se hace ninguna mención al proceso en todo el periodo analizado, que Barcelona (1.289 noticias), Valencia (336 noticias) y A Coruña (335 noticias) son las provincias con mayor cobertura del tema, y que en Madrid apenas se encuentran 28 noticias sobre este

<sup>16</sup><http://cort.as/-C56g>, accedida en noviembre de 2018.

asunto, un número llamativamente bajo teniendo en cuenta el volumen de publicaciones de su sección local. En la Figura 6 se puede ver una captura de la infografía centrada en el análisis temporal.



Figura 6: Análisis temporal de las noticias relacionadas con la independencia de Cataluña

## 6 Conclusiones y trabajo futuro

Este artículo presenta el desarrollo de un corpus de noticias periodísticas de gran tamaño que contempla tanto la dimensión temporal como la geográfica de las noticias. Hasta donde tenemos conocimiento, no existe en la actualidad un corpus de noticias con estas características libremente disponible. Este corpus puede servir de fuente para realizar distintos estudios del uso y evolución del lenguaje, además de ser un recurso de utilidad para el periodismo de datos, tal y como se mostró en la Sección 5 mediante dos casos de estudio. Si bien el corpus, por cuestiones de derechos de uso no está disponible para su libre distribución, una de las aportaciones de este trabajo es proporcionar las herramientas software necesarias para que cualquier persona que quiera trabajar con él pueda descargarlo y replicarlo en su ordenador.

En este trabajo se ha descrito todo el proceso llevado a cabo para la extracción y obtención de las noticias, dificultades afrontadas para la limpieza y eliminación de duplicados, así como el posterior análisis lingüístico y enriquecimiento del mismo. Se han proporcionado también datos estadísticos del corpus resultante desde las dimensiones temporal y geográfica, desarrollando diversas infografías interactivas que permiten un análisis más profundo y detallado del mismo.

Como trabajo futuro se plantea el desarrollo de nuevos estudios de carácter lingüístico y sociológico que puedan resultar de interés a

partir de esta fuente de información. Se proyecta también extender este trabajo para su aplicación a otros corpus de textos en la Web que puedan ser recopilados de manera similar.

## Agradecimientos

Queremos agradecer al Dr. Borja Navarro-Colorado sus valiosos comentarios y sugerencias previos al envío de este artículo.

## Bibliografía

- Broder, A. Z. 1997. On the resemblance and containment of documents. En *Proceedings of the Compression and Complexity of Sequences 1997*, SEQUENCES '97, páginas 21–29, Washington, DC, USA. IEEE Computer Society.
- Graff, D. y G. Gallegos. 1995. Spanish news text. Download at Linguistic Data Consortium: <https://catalog.ldc.upenn.edu/LDC95T9>.
- Graff, D. y G. Gallegos. 1999. Spanish newswire text, volume 2. Download at Linguistic Data Consortium: <https://catalog.ldc.upenn.edu/LDC99T41>.
- Gray, J., L. Chambers, y L. Bounegru. 2012. *The data journalism handbook: How journalists can use data to improve the news*. O'Reilly Media.
- Holmes, D. I. 1985. The analysis of literary style. *Journal of the Royal Statistical Society. Series A (General)*, 148(4):328–341.
- Indyk, P. y R. Motwani. 1998. Approximate nearest neighbors: Towards removing the curse of dimensionality. En *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, páginas 604–613, New York, NY, USA. ACM.
- Leetaru, K. 2011. Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, 16(9).
- Leskovec, J., A. Rajaraman, y J. D. Ullman. 2014. *Mining of Massive Datasets*. Cambridge University Press, New York, NY, USA, 2nd edición.
- Padró, L. y E. Stanilovsky. 2012. Free-ling 3.0: Towards wider multilinguality. En *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey. ELRA.

# Deep learning approach for negation trigger and scope recognition

## *Experimentación basada en deep learning para el reconocimiento del alcance y disparadores de la negación*

Hermenegildo Fabregat<sup>1</sup>, Lourdes Araujo<sup>1,2</sup>, Juan Martinez-Romo<sup>1,2</sup>

<sup>1</sup>NLP & IR Group, Dpto. Lenguajes y Sistemas Informáticos,  
Universidad Nacional de Educación a Distancia (UNED),  
Juan del Rosal 16, Madrid 28040, Spain

<sup>2</sup>IMIENS: Instituto Mixto de Investigación, Escuela Nacional de Sanidad,  
Monforte de Lemos 5, Madrid 28019, Spain

gildo.fabregat@lsi.uned.es, lurdes@lsi.uned.es, juaner@lsi.uned.es

**Abstract:** The automatic detection of negation elements is an active area of study due to its high impact on several natural language processing tasks. This article presents a system based on deep learning and a non-language dependent architecture for the automatic detection of both, triggers and scopes of negation for English and Spanish. The presented system obtains for English comparable results with those obtained in recent works by more complex systems. For Spanish, the results obtained in the detection of negation triggers are remarkable. The results for the scope recognition are similar to those obtained for English.

**Keywords:** Negation scope, negation triggers detection, deep learning

**Resumen:** La detección automática de los distintos elementos de la negación es un frecuente tema de estudio debido a su alto impacto en diversas tareas de procesamiento de lenguaje natural. Este artículo presenta un sistema basado en *deep learning* y de arquitectura no dependiente del idioma para la detección automática tanto de disparadores como del alcance de la negación para inglés y español. El sistema presentado obtiene para inglés resultados comparables a los obtenidos en recientes trabajos por sistemas más complejos. Para español destacan los resultados obtenidos en la detección de claves de negación. Por último, los resultados para el reconocimiento del alcance de la negación, son similares a los obtenidos en inglés.

**Palabras clave:** Detección de negación, disparadores de la negación, *deep learning*

## 1 Introduction

The study of negation is an active research topic due to its effects and importance within the different challenges of the natural language processing (NLP) research area. Although many tasks and domains are affected by negation, its study in the biomedical domain is of particular relevance. Chapman et al. (2001a) shows the importance of consider possible negated phrases during the analysis of electronic health records (EHR), documents in which much of the information contained is expressed in a negated way. The contributions of negation in tasks such as sentiment analysis and relationship extraction stand out due to the performance improvements obtained after considering it. Coun-

cill, McDonald, and Velikovich (2010) examine the achieved improvements after including the study of negation in the task of sentiment analysis in online product reviews. The authors take into account the negation during the evaluation of the score of each term in a sentence, modifying the score sign if the term is part of a negation. On the other hand, Chowdhury and Lavelli (2013) highlight the significant performance improvements obtained in the detection of drug-drug relationships after considering negation.

There are many possible elements of study concerning negation. This paper deals with both, identification of negation triggers and the delimitation of negation scope. The detection of negation triggers, can be considered a basic task in the study of negation.

It refers to the identification of expressions that work as markers of negation. The identification of negation scope refers to finding segments of a sentence that are part of one or more negations. This study presents the experimentation carried out using a deep learning approach for the study of the negation scope for both English and Spanish. The following sections present a study of the state of the art for the proposed work (Section 2), a summarization of the proposed approach including a brief description of the explored datasets (Section 3) and finally, a review of the obtained results (Section 4) and the conclusions reached after the experimentation (Section 5).

## 2 Background

Chapman et al. (2001b) presented an algorithm called NegEx based on the use of regular expressions for the detection of negation in clinical documents. Tools such as cTAKES (Savova et al., 2010), designed for processing medical documents in free text format, use NegEx for the treatment of negation. Nowadays, NegEx is considered a baseline in many of the works dealing with the automatic study of negation. Although this algorithm shows a high *performance*, an important issue to take into account, is the low precision obtained evaluating sentences where the term “no” appears. Goldin and Chapman (2003) extend the study of this case developing a set of experiments in order to compare the results obtained by NegEx with those obtained by a set of different machine learning algorithms. Among them Naive Bayes (NB) and Decision Tree (DT), achieved better results than NegEx. Although NegEx has been designed for English, recent works such as Chapman et al. (2013) and Skeppstedt (2011) have studied its use for other languages such as French, German, Swedish and Spanish. Cotik et al. (2016) show the results obtained by an adaptation of NegEx to Spanish. Their results are better than the use of dictionaries and comparable with those obtained by a system of rules based on patterns of PoS Tagging.

There are many systems evaluated using the Bioscope corpus (Vincze et al., 2008) which is a linguistic resource containing annotations about negation and speculation in the biomedical domain. Fancellu, Lopez, and Webber (2016) shows the good performance

of Bidirectional Long Short-Term Memory (bi-LSTM) based models for the identification of multi-term expressions such as “by no means of” and “no longer”. Fancellu et al. (2017) extend the study to other domains and languages (Chinese), presenting, among others, results for the Bioscope corpus and for the SFU corpus (Konstantinova et al., 2012). The study shows a comparison of the results obtained by a bidirectional long short-term memory (bi-LSTM) based model with some state of the art systems. For different domains, Li and Lu (2018) deal with both, the detection of negation triggers and the recognition of the scope of negation, using different kinds of conditional random fields (CRF), *linear CRF* (Lafferty, McCallum, and Pereira, 2001), *semi-CRF* (Sarawagi and Cohen, 2005) and *latent variable CRF*. Taking into account the obtained results, one of the conclusions the authors reached was the good performance of this kind of algorithms in sequence labeling tasks, having obtained remarkable results even after extending the evaluation to languages such as Chinese.

## 3 Materials and methods

The aim of this work has been to look for a simple deep learning architecture for negation detection valid for different languages. We consider both, the detection of negation triggers and the recognition of the scopes. We have built a simple architecture in which input data have proven to be useful. In the following sections, we describe the proposed model and details of the corpora we use for the evaluation. We also explain the details of both, text pre-processing and the system output post-processing processes.

### 3.1 Features

The proposed model uses the following features:

**Words.** For Spanish, we have used the embedding vectors presented by Cardellino (2016). They are vectors of 300 dimensions and collect a total of 1000653 unique tokens. They were generated using the word2vec algorithm by means of multiple repositories of information in Spanish for training. For English, the word embedding presented by Pyysalo et al. (2013) were used. They are 200-dimensional vectors and collect about twenty-four million unique tokens. This

resource was generated using word2vec and taking as a source of information several wikipedia dumps and some biomedical repositories such as PubMed and PMC.

**PoS-Tagging.** We have used FreeLing PoS-tagger (Padró and Stanilovsky, 2012) for Spanish and the maximum entropy PoS tagger implemented in NLTK (Bird and Loper, 2004) for English.

**Casing.** Another feature used is a matrix for the representation of word casing information. Each token has been represented with the corresponding index of the matrix embedding. The casing embedding matrix is a hot-one encoding matrix of size 14. This feature provides additional support to the model by representing each token in a summary category.

**Chars.** We use character embeddings in order to collect expressions not included in the pre-trained word embeddings vocabulary, taking into account that the vocabulary of health records is not standard and that the vocabulary of product reviews may contain spelling errors. This vectorial representation allows to represent the information contained in both prefixes and suffixes.

Both PoS-tagging, casing and character embedding models have been implemented using three Keras Embedding Layers initialized using a random uniform distribution. In Section 4 the performance improvement obtained after considering each of the features is shown.

### 3.2 Proposed model

Figure 1 shows the architecture of the proposed model. This architecture consists of a character-level processing module (Santos and Zadrozny, 2014) and a word-level processing module (Fabregat, Araujo, and Martinez-Romo, 2018). Character-level processing is essentially a transformation of the characters of each word into character embeddings and on the concatenation of the most important features obtained by the application of a convolutional layer. The result of this process is concatenated to the input of the other part of the model.

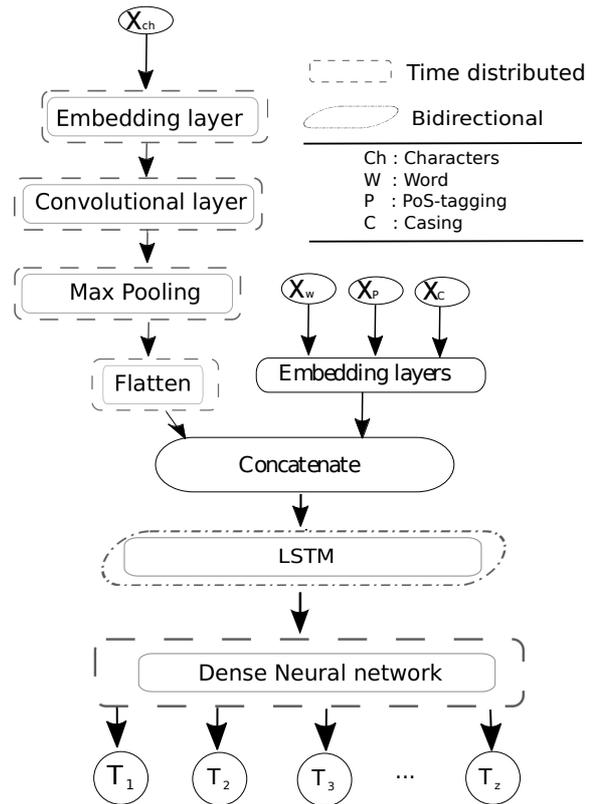


Figure 1: Architecture of the proposed model, where  $X_{Ch}$  and  $X_W$  (Ch: Characters, W: Raw word) are the encoded word inputs and  $X_P$  and  $X_C$  are the encoded inputs representing the PoS-tagging and casing information. Bi-LSTM inputs ( $Y_x$ ) are the concatenated embedded features for each word. In the output layer,  $T_x$  represents the assigned tag.

The second part of the model consists of a Bidirectional Long Short-Term Memory (Bi-LSTM) network connected to a neural network. On the one hand, the Bi-LSTM is responsible of processing each part of the concatenated features in order to obtain positional/semantic relationships between terms of a sentence. On the other hand, the neural network is responsible of generating the correct classification sequence. This neural network has softmax as an activation function and processes the output of each part of the sequence returned by the previous LSTM.

Taking into account the obtained results in Fabregat, Araujo, and Martinez-Romo (2018), which deals with a problem of recognition of negation triggers by means of an approach based on deep learning, the adjustment of parameters has been made according to the configuration of the model presented in

that paper. For the configuration of the additional layers that form the character-level processing, we have used of a configuration focused on extracting syntactic/semantic features from immediately adjacent characters. The final configuration is as follows:

- Convolutional layer (kernel\_size / filter): 3 / 30
- Embeddings dimension (Casing / PoS-tagging / Char): 12 / 50 / 50
- LSTM output dimension: 250
- Dropout: 0.5
- Batch size / Model optimizer: 32 / Adam
- Hidden Dense units (output dimension / activation function): 17 / softmax

The number of neurons in the output layer of the neural network corresponds to the total of classes to be considered in the annotation.

### 3.3 Pre-processing

During the pre-processing phase, the different datasets are transformed into the BILOU labelling scheme (Ratinov and Roth, 2009). In this annotation scheme the information is represented applying the following map: {I:In - For tokens part of the annotation. O: Out - For tokens outside the annotation. B:Begin - For the first token of each annotation. L:Last - For the last token of each annotation. U:Unique - Those annotations that have a single token.}. This annotation scheme, used in entity recognition tasks, allows the partial overlapping and nesting of one entity within another, a characteristic necessary to represent cases such as two or more negations starting in the same term or a negation included within another. We have carried out this encoding in order to be able to deal with this problem from the perspective of a classification problem. We have used the labelling code to represent both, the scope and the negation triggers separately. These two codifications are combined into a single one by concatenating the labels. For example, if an expression is both, the beginning of a negation and the beginning of a negation triggers, this will be re-labeled with the label “BB”. Table 1 shows an example of BILOU annotation format. The first column contains the word and the second column contains the label after joining the scope label and the trigger associated label. The

example shows the annotation of both scopes. While the first one spans from the first term “no” up to the term “dinero”, the other is nested and spans from the second term “no” up to the term “gusta”.

Word	Label	Word	Label
no	BU	no	BU
tendré	IO	me	IO
jamás	IU	gusta	LO
que	IO	por	IO
aceptar	IO	el	IO
un	IO	dinero	LO
trabajo	IO	.	OO
que	IO		

Table 1: SFU Review SP-NEG fragment with tag assignment.

Considering that a negation must have associated a scope, i.e. there are combinations of labels that cannot occur, a total of 17 labels are generated.

### 3.4 Post-processing

The post-processing phase aims to ensure that the format generated by the model is correct. This format must satisfy the following requirements: Each scope must have at least one associated negation trigger and each annotation must have both, a start and an end label, except in the case of a single token annotation. This phase applies the following rules:

- If a scope does not have at least one negation trigger associated to it, it is not a scope.
  - **Sentence** Don’t you think it’s late?
  - **Proposed labels** BO BO BO BO BO
  - **Processed labels** OO OO OO OO OO
- If a negation trigger does not have one negation scope associated to it, it is not a negation trigger.
  - **Sentence** Don’t you think it’s late?
  - **Proposed labels** IS OO OO OO OO
  - **Processed labels** OO OO OO OO OO
- If an annotation starts but does not closes, then it finishes in the last term considered by the system as part of the annotation.
  - **Sentence** don’t buy it.
  - **Proposed labels** BS IO IO

- **Processed labels** BS IO LO

- If an annotation closes a scope but it is not open, it starts with the first trigger of the phrase detected by the system.

- **Sentence** don't buy it.

- **Proposed labels** IS IO LO

- **Processed labels** BS IO LO

### 3.5 Corpora

In order to extract conclusions for both languages, English and Spanish, we have selected two corpora with similar annotation guidelines. They are Bioscope corpus and SFU Review SP-NEG corpus (Jiménez-Zafra et al., 2018). Bioscope corpus consists of three parts, electronic health records (EHRs) presented in free text format, full biological articles and abstracts of both scientific and biological articles. The domains included in this corpus present a complex structure, being the most different the domain of EHRs for the use of a free writing style. The subset of abstracts stands out because it contains more negations than the rest and it is the largest subset. On the other hand, SFU Review SP-NEG corpus consists of a collection of 400 comments on cars, hotels, washing machines, books, mobile phones, music, computers and films from the *Ciao.es* website. This corpus presents a mixture of free-writing and formal writing styles.

In both collections, each document has been annotated at token and sentence level

with labels related to negation triggers and their linguistic scope. In addition, both collections have used an annotation style without gaps. In summary, using both datasets and taking into account the different writing scenarios of their documents, this work studies the performance of the proposed architecture for free text in Spanish and for both free text and well-structured text in English.

## 4 Evaluation

This model has been evaluated using 10 fold cross-validation and we have made a study of the improvement obtained considering each attribute of the model. We have used two separate workflows for the model evaluation: one for the evaluation of negation scope detection and a second for the evaluation of negation triggers recognition. For the evaluation of negation scope detection task, we used the percentage of correctly identified scopes (PCS), F1 measure at scope level (F1s) and F1 measure at token level (F1t). F1s measure only considers as false positives those scopes that have been identified but were not found in the gold standard. For the evaluation of triggers, we used precision, recall and F1-measure. In both cases, the evaluation metrics have been used in previous works. The results obtained during this preliminary evaluation suggest that the proposed model is appropriate to deal with the different domains proposed. The experiments carried out show an improvement in performance af-

Training Features	Bioscope									SFU SP-NEG		
	Abstracts			Clinical records			Full papers			All categories		
	PCS	F1s	F1t	PCS	F1s	F1t	PCS	F1s	F1t	PCS	F1s	F1t
W	75.6	84.8	75.52	89.11	94.07	91.28	46.77	59.55	46.17	69.76	82.04	67.32
W+P	81.83	89.9	80.3	94.57	97.12	94.78	49.05	64.79	50.41	72.40	84.05	71.34
W+P+C+Ch	80.52	88.54	80.05	90.03	94.63	91.85	58.99	70.67	58.50	74.29	85.25	72.00

Table 2: Evaluation of each elements considered in the proposed model for the identification of the scope, {W:Words - P:PoS - C: Casing - Ch: Chars}. PCS (percentage of correctly identified scopes), F1 at scope level (F1s) and F1 at token level (F1t) are the metrics analyzed

Training Features	Bioscope									SFU SP-NEG		
	Abstracts			Clinical records			Full papers			All categories		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
W	96.03	94.19	95.0	99.11	96.14	97.59	90.99	80.6	84.81	99.04	91.18	94.95
W+P	99.59	95.67	97.59	99.76	97.42	98.57	99.42	82.58	90.02	97.10	92.43	94.70
W+P+C+Ch	97.4	94.28	95.75	98.82	96.24	97.46	93.25	83.21	87.45	99.69	91.85	95.60

Table 3: Evaluation of each elements considered in the proposed model for the exact identification of negation triggers, {W:Words - P:PoS - C: Casing - Ch: Chars} Prec: Precision, Rec: Recall and F1 measure are the metrics analyzed

ter the addition of each of the features studied. The decrease of performance for the corpus Bioscope in the categories of abstracts and clinical records after adding casing and chars features is remarkable. This may be because the model is over adjusted to certain patterns discovered with these features. According to the difference in negation occurrences between the different subsets in the Bioscope corpus, evaluating the subset of articles we obtain the highest standard deviation ( $\pm 11\%$ ) and the worst performance. In this study, some errors were detected in the treatment of double negation and in the handling of multi-term expressions. The Spanish results show improvements as new features are added to the study. Many of the errors detected in the identification of the scope correspond to sentences with large negations. In order to study the robustness of the studied model, Table 4 shows the results obtained by evaluating the behaviour of the model detecting the scope, training it with data from a subset and validating it with data from the other sets. We have only been able to carry out this experiment with the corpus in En-

glish because it is the only one that shows a structure divided into categories with strong differences in the writing style. As can be observed, the best inter-domain performance is obtained when training with the abstracts subset. It is because the set of abstracts is the group that contains more negations and it uses a language very similar to the rest of subsets. The main problem detected is the mean length difference of the negations contained in the different subsets. The system trained with the set of abstracts tends to lose performance evaluating long sentences contained in the set of articles.

Finally, results in both datasets have been compared with the results of the state of the art systems results (Table 5 and Table 6). Competitive results have been obtained evaluating with Bioscope, although with a remarkable lower performance in terms of precision. This is mainly due to the fact that our system tends to generate a shorter length range than that collected in the gold standard. Some detected errors are those in which the first negation trigger appears far from the beginning of the scope. In these

Training with	Testing with					
	Abstracts		Clinical records		Full papers	
	PCS	F1	PCS	F1	PCS	F1
Abstracts	-.	-.	<b>84.90</b>	<b>91.83</b>	<b>55.85</b>	<b>71.67</b>
Clinical records	40.22	53.68	-.	-.	40.05	53.94
Full papers	<b>76.81</b>	<b>64.00</b>	81.68	89.91	-.	-.

Table 4: Bioscope corpus (English) - Evaluation of interdomain scope recognition. Results obtained by training with one of the Bioscope subsets (abstracts, clinical records and full papers) and testing with other

System	Abstracts		Clinical records		Full papers	
	PCS	F1	PCS	F1	PCS	F1
Proposed model	80.52	88.54	90.03	94.63	58.99	70.67
Li and Lu (2018)	<b>84.1</b>	91.3	94.4	95.59	<b>60.1</b>	69.23
Fancellu et al. (2017)	81.38	<b>92.11</b>	94.21	<b>97.94</b>	54.54	77.73
Fancellu, Lopez, and Webber (2016)	73.72	91.35	<b>95.78</b>	97.66	51.24	<b>77.85</b>

Table 5: Bioscope corpus (English) - Evaluation of negation scope recognition: Comparison with other state-of-the-art approaches

System	P	Triggers			Scope		
		R	F1	PCS	F1s	F1t	
Proposed Model	<b>99.69</b>	<b>91.85</b>	<b>95.60</b>	<b>74.29</b>	<b>85.25</b>	<b>72.00</b>	
Fabregat, Araujo, and Martínez-Romo (2018)	79.45	59.58	67.97	-	-	-	
Loharja, Padró, and Turmo Borrás (2018)	91.48	82.18	86.45	-	-	-	

Table 6: SFU Review SP-NEG corpus - Evaluation of recognition of both negation scope (PCS, F1s, F1t) and negation triggers (P: Precision, R: Recall, F1): Comparison of obtained results by the proposed model with results from other state-of-the-art approaches

cases, the system makes mistakes such as ignoring the presence of multi-term expressions. The post-processing process also generates certain errors, especially in cases of double negation. Regarding the results obtained evaluating with the corpus SFU SP-NEG, as far as we know, there are only results for negation triggers recognition and only using training and test evaluation which makes it difficult to reach conclusions about the state of the art improvements. However, the results obtained for both the detection of negation triggers and for the recognition of the negation scope, are comparable to those obtained for English using the Bioscope corpus. Some of the errors reported in works about negation triggers detection that use the SFU SP-NEG corpus have been corrected incorporating the BILOU format and using of character embeddings.

## 5 Conclusions and future work

This research has focused on the generation of a common model to deal with negation in both English and Spanish languages. In order to generalize its application to different languages, the model has been trained mainly using non-language dependent writing features. Results show that it is a robust architecture based on a single supervised learning model for both detection of negation triggers and recognition of their scope. Performance obtained for English is comparable to state of the art and results obtained for Spanish are only slightly lower than for English. Possible future work lines are the study of more non-language dependent features and the improvement of the extraction of relationships between terms introducing n-gram embeddings as a feature. Some detected errors related to multi-term expressions suggest that working with n-gram embeddings can improve current precision results.

## Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Innovation within the projects PROSA-MED (TIN2016-77820-C3-2-R) and EXTRA-E (IMIENS 2017).

## References

Bird, S. and E. Loper. 2004. Nltk: The natural language toolkit. In *Proceedings of the ACL 2004 on Interactive Poster and*

*Demonstration Sessions*, ACLdemo '04, Stroudsburg, PA, USA. ACL.

Cardellino, C. 2016. Spanish Billion Words Corpus and Embeddings, March.

Chapman, W. W., W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. 2001a. Evaluation of negation phrases in narrative clinical reports. In *Proceedings of the AMIA Symposium*, pages 105–109. AMIA.

Chapman, W. W., W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. 2001b. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310.

Chapman, W. W., D. Hilert, S. Velupillai, M. Kvist, M. Skeppstedt, B. E. Chapman, M. Conway, M. Tharp, D. L. Mowery, and L. Deleger. 2013. Extending the negex lexicon for multiple languages. *Studies in health technology and informatics*, 192:677.

Chowdhury, M. F. M. and A. Lavelli. 2013. Exploiting the scope of negations and heterogeneous features for relation extraction: A case study for drug-drug interaction extraction. In *Proceedings of the 2013 Conference of the North American Chapter of the ACL: Human Language Technologies*, pages 765–771. ACL.

Cotik, V., V. Stricker, J. Vivaldi, and H. Rodríguez Hontoria. 2016. Syntactic methods for negation detection in radiology reports in spanish. In *Proceedings of the 15th Workshop on BioNLP 2016: Berlin, Germany, August 12, 2016*, pages 156–165. ACL.

Councill, I. G., R. McDonald, and L. Velikovich. 2010. What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. In *Proceedings of the workshop on negation and speculation in NLP*, pages 51–59. ACL.

Fabregat, H., L. Araujo, and J. Martinez-Romo. 2018. Deep learning approach for negation cues detection in spanish. In *NEGES 2018: Workshop on Negation in Spanish: Seville, Spain: September 19-21, 2018: proceedings book*, pages 43–48.

- Fancellu, F., A. Lopez, and B. Webber. 2016. Neural networks for negation scope detection. In *Proceedings of ACL (Vol. 1: Long Papers)*, volume 1, pages 495–504. ACL.
- Fancellu, F., A. Lopez, B. Webber, and H. He. 2017. Detecting negation scope is easy, except when it isn't. In *Proceedings of European Chapter of the ACL: Vol. 2, Short Papers*, volume 2, pages 58–63. ACL.
- Goldin, I. and W. W. Chapman. 2003. Learning to detect negation with 'not' in medical texts. In *Proceedings of the Workshop on Text Analysis and Search for Bioinformatics, ACM SIGIR*. ACM SIGIR.
- Jiménez-Zafra, S. M., R. Morante, M. Martín, and L. A. U. Lopez. 2018. A review of spanish corpora annotated with negation. In *Proceedings of the Conference on Computational Linguistics*, pages 915–924.
- Konstantinova, N., S. C. de Sousa, N. P. Cruz, M. J. Maña, M. Taboada, and R. Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. ELRA.
- Lafferty, J. D., A. McCallum, and F. C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA. ICML.
- Li, H. and W. Lu. 2018. Learning with structured representations for negation scope extraction. In *Proceedings of ACL (Vol. 2: Short Papers)*, volume 2, pages 533–539. ACL.
- Loharja, H., L. Padró, and J. Turmo Borras. 2018. Negation cues detection using crf on spanish product review texts. In *NEGES 2018: Workshop on Negation in Spanish: Seville, Spain: September 19-21, 2018: proceedings book*, pages 49–54.
- Padró, L. and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*. LREC.
- Pyysalo, S., F. Ginter, H. Moen, T. Salakoski, and S. Ananiadou. 2013. Distributional semantics resources for biomedical text processing. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine*, pages 39–44. LBM.
- Ratinov, L. and D. Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on CoNLL*, pages 147–155. ACL.
- Santos, C. D. and B. Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *Proceedings of the International Conference on Machine Learning*, pages 1818–1826. ICML.
- Sarawagi, S. and W. W. Cohen. 2005. Semi-markov conditional random fields for information extraction. In *Advances in Neural Information Processing Systems 17*, pages 1185–1192. NIPS.
- Savova, G. K., J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the AMIA*, 17(5):507–513.
- Skeppstedt, M. 2011. Negation detection in swedish clinical text: An adaption of negex to swedish. *Journal of Biomedical Semantics*, 2(3):S3, Jul.
- Vincze, V., G. Szarvas, R. Farkas, G. Móra, and J. Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(S11):S9.

# Hacia una generación de resúmenes sin sesgo a partir de contenido generado por el usuario: Un enfoque preliminar

## *Towards Unbiased Automatic Summarization from User Generated Content: A Preliminary Approach*

Alejandro Reyes<sup>1</sup>, Elena Lloret<sup>2</sup>

Departamento de Lenguajes y Sistemas Informáticos

Universidad de Alicante

Apdo. de Correos 99

E-03080, Alicante, Spain

<sup>1</sup>ara65@alu.ua.es, <sup>2</sup>elloret@dlsi.ua.es

**Resumen:** En este trabajo se propone un enfoque novedoso de generación automática de resúmenes capaz de sintetizar grandes cantidades de información generada por diferentes tipos de usuarios en Internet y producir un nuevo texto coherente que presente la información de forma objetiva, es decir, evitando proporcionar información parcial o sesgada, a la par que aportando múltiples perspectivas sobre el tema en cuestión. En concreto, el escenario en el que se enmarca esta investigación es el ámbito turístico, centrándonos en las opiniones sobre distintos aspectos de varios hoteles proporcionadas por 5 tipos de perfil de usuario. La evaluación realizada con usuarios demuestra que i) la calidad de los resúmenes generados es adecuada y ii) que este tipo de resúmenes ayudaría a los usuarios a tomar mejores decisiones.

**Palabras clave:** Generación de resúmenes, contenido generado por el usuario, resúmenes abstractivos, resúmenes multi-perspectiva, información sin sesgo

**Abstract:** In this paper a novel approach for automatic summarization is proposed. This approach is able to synthesize huge amounts of information generated by different types of users on the Internet and produce a new coherent text that presents the information in an objective way, i.e., avoiding biased information and giving multiple perspectives for an specific aspect/topic. This study is mainly focused on the tourism sector, especially on the opinions about different topics existing in multiple hotels and given by 5 user types. The user evaluation conducted proves that i) the generated summaries have an appropriate quality, and ii) they would really help users to make better decisions.

**Keywords:** Automatic Summarization, User Generated Content, Abstractive Summarization, Multi-perspective Summarization, Unbiased information

## 1 Introducción

La aparición de la Web 2.0 ha supuesto un enorme aumento de la cantidad de datos a disposición de las personas. Este hecho, que a priori podría considerarse como una ventaja —a mayor información, mejor toma de decisiones—, en realidad supone una desventaja al no poder gestionar de forma eficaz y eficiente la sobrecarga de información disponible (Luo et al., 2013).

Por ejemplo, en el sector del turismo, nos encontramos con foros, redes sociales y otras plataformas que incluyen información

que puede ser relevante para un usuario a la hora de decidir sobre su futuro viaje. Las opiniones, valoraciones proporcionadas por otros usuarios —“Contenido Generado por el Usuario” (CGU)— que han visitado el mismo lugar, así como otra información adicional en estos medios, suponen una gran ayuda a los usuarios a la hora de tomar decisiones. Sin embargo, resulta muy difícil sintetizar tal cantidad de información y, generalmente, el usuario acaba leyendo unas sólo unas cuantas opiniones, usualmente las primeras, pudiendo obtener así una idea sesgada o parcial, elabo-

Hotel	#Coment. (03/2018)	#Coment. (11/2018)
Luxor Hotel & Casino Las Vegas	32,007	35,897
Melia Alicante	4,752	5,428

Tabla 1: Evolución en el número de comentarios de dos hoteles entre marzo y noviembre de 2018 (fuente de datos: TripAdvisor)

rada en base a un subconjunto de opiniones.

A modo ilustrativo, la Tabla 1 muestra el incremento en el número de comentarios generados por los usuarios de la página TripAdvisor<sup>1</sup> para dos hoteles elegidos al azar entre las fechas 5 de marzo y 5 de noviembre de 2018. Como se puede observar, el volumen de comentarios es considerablemente elevado, lo que hace inviable para una persona leer y analizar cada uno de ellos de forma manual.

Además, el hecho de encontrar opiniones diversas y contradictorias para un mismo aspecto hace que sea complicado para el usuario escoger cuáles de ellas se adecúan más a su perfil o necesidades. Esto se debe a que los comentarios que proporcionan los usuarios están basados en su experiencia, por lo que un aspecto específico puede ser malo para un usuario, mientras que para otro no. Por ejemplo, podemos ver que, dependiendo del tipo de viajero, algunos comentarios se contradicen para un mismo hotel: *“Mala ubicación, ya que está lejos del centro del Strip.”*, *“Su ubicación y comunicación con otros centros de diversión [...] le hacen estratégica.”*<sup>2</sup>.

Por todo ello, la hipótesis de nuestro trabajo es que un enfoque que sea capaz de resumir objetivamente toda esta información, captando los distintos puntos de vista sobre, por ejemplo, los diferentes aspectos de un hotel específico de acuerdo al perfil del viajante, sería de gran ayuda para los usuarios que desean saber de forma fácil y sencilla cuál es la mejor opción de alojamiento para su viaje. En base a esta hipótesis, nos planteamos como objetivo principal el análisis y uso de técnicas de Procesamiento de Lenguaje Natural (PLN) para desarrollar un enfoque de generación de resúmenes abstractivo, multi-documento y multi-perspectiva, capaz de producir resúmenes que no contengan información sesgada o parcial. Los resultados obteni-

dos verifican la hipótesis de partida y demuestran que los resúmenes generados, además de ayudar en la toma de decisiones, son coherentes y están bien escritos.

## 2 Estado de la cuestión

En la actualidad, los métodos y sistemas de resúmenes existentes, como por ejemplo el propuesto en (Paulus, Xiong, y Socher, 2017), se centran en la extracción y/o abstracción de la información más relevante de un texto, teniendo en cuenta varios factores, entre los que destacan la detección de la redundancia y la detección de polaridad, sobre todo cuando se generan resúmenes de opiniones o a partir de información subjetiva. Sin embargo, casi ninguno de ellos tiene en cuenta si existe información contradictoria sobre un aspecto específico ni tampoco el perfil de usuario para el que va dirigido el resumen. Por ello, en este artículo nos centraremos únicamente en aquellos enfoques que intentan ofrecer distintos puntos de vista sobre opiniones (resúmenes multiperspectiva), ya que son los más recientes relacionados con resúmenes a partir de CGU.

Lloret (2016) propone un método para la explotación de los metadatos existentes en la información procedente de CGU para enfocar la generación de resúmenes de distinta manera dependiendo de las necesidades del usuario en concreto. Para ello, realizó un estudio en el que terminó subdividiendo el proceso de generación de resúmenes en 3 fases: i) extracción de información básica; ii) identificación del tópico y de la polaridad; iii) selección de información relevante para la generación del resumen. La principal limitación de este trabajo es que solamente se queda a nivel de propuesta y no existe una implementación ni evaluación al respecto para poder determinar cuán útiles y buenos son los resúmenes generados.

Por otra parte, en el trabajo presentado por Esteban y Lloret (2017a) también se propone un sistema similar al de este trabajo, llamado TravelSum<sup>3</sup>, que consiste en una aplicación web de la cual los usuarios pueden obtener un resumen generado automáticamente en español a partir de CGU (Esteban y Lloret, 2017b). La principal diferencia entre el enfoque propuesto en este artículo y TravelSum, es que TravelSum únicamente agrupa

<sup>1</sup><https://www.tripadvisor.es/>

<sup>2</sup>[https://www.tripadvisor.es/Hotel\\_Review-g45963-d111709-Reviews-Luxor\\_Hotel\\_Casino-Las\\_Vegas\\_Nevada.html](https://www.tripadvisor.es/Hotel_Review-g45963-d111709-Reviews-Luxor_Hotel_Casino-Las_Vegas_Nevada.html)

<sup>3</sup><http://travelsun.gplsi.es/>

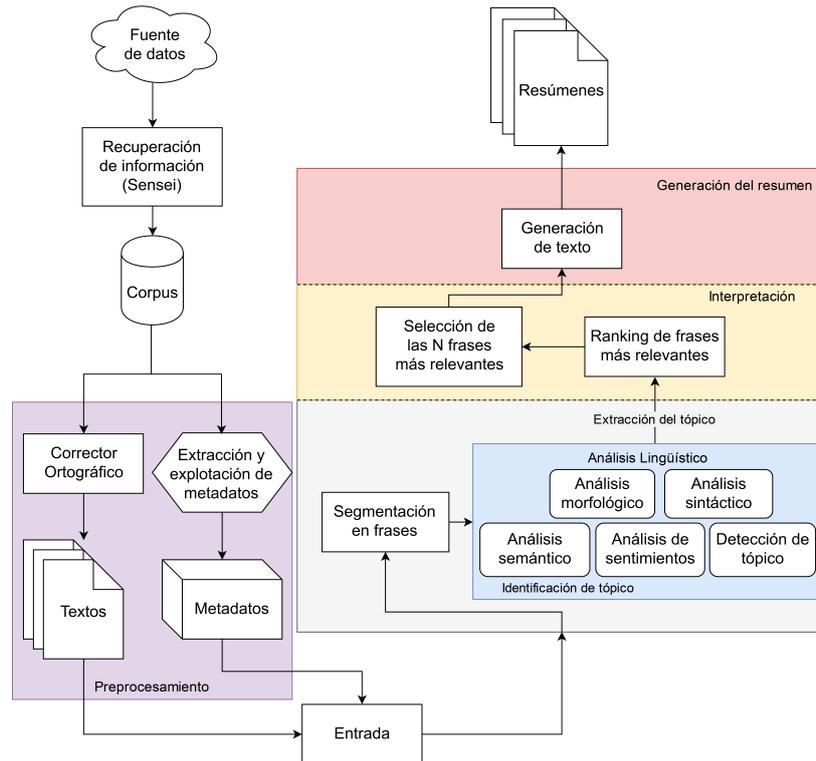


Figura 1: Arquitectura para el enfoque de resúmenes propuesto

frases por polaridad, mientras que el nuestro, además de aplicar un proceso de minería de opiniones, tiene en cuenta también la variedad y diferencias de opiniones de los usuarios para un mismo aspecto, así como el perfil de usuario que realiza un comentario (p. ej. si ha viajado solo, con amigos, etc.).

Entre las principales novedades de nuestro trabajo respecto a los comentarios, destacan: i) mejora e implementación completa de un enfoque de resúmenes basada en la idea conceptual de Lloret (2016); ii) la explotación e integración de los metadatos de los textos para el cálculo de la información relevante; y iii) análisis de la contradicción entre los distintos tópicos, para que los resúmenes generados no estén sesgados y/o se puedan adaptar a un perfil de usuario.

### 3 Enfoque de generación de resúmenes sin sesgo

La Figura 1 muestra la arquitectura del enfoque propuesto. Primero, se determina la fuente de información y se recopila el conjunto de documentos que servirán como entrada al enfoque de resúmenes. En nuestro caso, se decidió utilizar reseñas de hoteles, ya que las opiniones de los usuarios de una zona turística-

ca son variadas, y se recopiló, de forma automática utilizando el crawler SENSEI<sup>4</sup>, un corpus compuesto por comentarios en inglés de 10 hoteles distintos, de una ciudad europea elegida al azar, junto con todos los metadatos disponibles (p.ej., utilidad del comentario, fecha de publicación, número de contribuciones o la valoración general del servicio). Finalmente, se obtuvo un total de 3615 comentarios, con una media de 340 comentarios por hotel.

Una vez recopilado el corpus, el siguiente paso es el preprocesado del corpus para corregir los errores ortográficos derivados del CGU. Para ello, se utiliza un diccionario en inglés<sup>5</sup> ligeramente modificado para la corrección de las erratas. También se extraen y guardan los metadatos del corpus.

Con esto, ya tendríamos la entrada preparada para el enfoque de generación de resúmenes, para cuyo diseño y desarrollo se tomó como base las tres fases propuestas por Hovy (2003), utilizando Python 2.7 como lenguaje de programación. A continuación, se explicará más en detalle como se ha aborda-

<sup>4</sup><https://gplsi.dlsi.ua.es/gplsi13/es/node/301>

<sup>5</sup><https://introcs.cs.princeton.edu/java/data/words.utf-8.txt>

do cada una de las fases y qué herramientas de PLN se han usado.

### 3.1 Identificación del Tópico

En esta fase se llevan a cabo distintos tipos de análisis lingüísticos para la detección e identificación de tópicos.

#### 3.1.1 Análisis morfológico, sintáctico y semántico

Tras el preprocesado, se realiza una segmentación del texto en frases, a las cuales se les realiza un proceso de análisis morfológico y sintáctico, haciendo uso de la herramienta NLTK (Bird y Loper, 2004). Además, se realiza un análisis semántico, utilizando Wordnet (Miller, 1998), para agrupar las palabras por conceptos sinónimos (*synsets*), lo cual ayudará posteriormente a la detección de tópicos relevantes y al análisis de los sentimientos. Cabe mencionar que en esta propuesta no se realiza ningún proceso de desambiguación del sentido de las palabras, por lo que nos quedamos con el sentido más frecuente para cada palabra, puesto que se trata de un enfoque que, a pesar de su simplicidad, obtiene resultados bastante aceptables con respecto al estado de la cuestión para la tarea de desambiguación (McCarthy y Navigli, 2007).

#### 3.1.2 Análisis de sentimientos

Analizar el sentimiento de una frase nos va a permitir conocer la opinión que un usuario ha proporcionado sobre un determinado aspecto y si esta opinión es la misma que la de otras frases o no. La polaridad de una frase se calculará sumando las polaridades individuales de las palabras que la componen, previa eliminación de las *stopwords*, gracias a la información que proporciona la herramienta SentiWordnet (Esuli y Sebastiani, 2006), que utiliza los *synsets* previamente obtenidos en el análisis semántico. Se decidió utilizar esta herramienta tras una investigación preliminar de las diferentes opciones existentes para el análisis de sentimientos, mostrados en trabajos como el de Denecke (2009).

Sin embargo, el análisis de sentimientos no nos permite saber si hablan sobre el mismo aspecto o tópico, y para ello necesitamos incluir un procesamiento adicional para la detección del tópico.

#### 3.1.3 Detección de tópico

La detección del tópico de una frase es el proceso por el cual se obtiene de qué está hablando la frase. Generalmente, las frases escritas

por usuarios para mostrar una opinión o hacer una reseña de algún producto o servicio tienden a hablar de varias cosas en la misma frase. Es por ello, que nuestro enfoque tiene en cuenta este hecho y obtiene los tópicos relevantes para, finalmente, determinar cuál es el tópico predominante para cada frase del texto/s que se quiere/n resumir.

Para realizar la detección de tópicos relevantes, primero se eliminan las partes de la frase que podrían obstaculizar el correcto análisis de los datos, tales como *stopwords* y signos de puntuación. Posteriormente, a partir de los lemas obtenidos de cada una de las palabras del texto, se conforma un diccionario del documento y se pasa a un modelo Latent Dirichlet Allocation (LDA), obtenido con la librería de Python Gensim (Řehůřek y Sojka, 2010), para obtener la lista ordenada de los tópicos más frecuentes de una frase.

A partir de la lista generada y ordenada por frecuencia de aparición, se realiza un filtrado en el que se descartan las palabras que no son sustantivos, ya que los verbos, adjetivos y adverbios no se pueden utilizar para la definición de una característica, al expresar acciones o utilizarse para describir o modificar un sustantivo o verbo. De este modo podemos seleccionar, del conjunto filtrado y ordenado por frecuencia, la palabra más comentada como el tópico predominante de la frase, almacenándolo para usarlo en los siguientes apartados.

### 3.2 Interpretación

En esta fase se toma la información de la etapa anterior y se interpreta para detectar la información más relevante que deberá conformar el resumen.

#### 3.2.1 Ranking de relevancia

Para determinar la información relevante, se tuvieron en cuenta dos aspectos. Por un lado, se realizó un cálculo de la relevancia del autor del comentario en base a los metadatos del corpus utilizado, y por otro lado se aplicaron una serie de heurísticas que nos permitieron saber cuán relevante es una frase de una reseña con respecto a las otras.

La relevancia del autor la calculamos utilizando 3 metadatos principales: i) la fecha de publicación del comentario; ii) el número de contribuciones realizadas por el autor; y iii) la cantidad de comentarios de este autor que han sido considerados útiles por otros usuarios. En base a estos tres factores se es-

tablece una puntuación para cada metadato que se sumará para obtener un número positivo que representará la relevancia del autor en el corpus.

Para calcular la relevancia de una frase, se realizó previamente un análisis sobre los tópicos más relevantes para un tipo específico de usuario, así como un análisis de la contradicción existente entre varias frases que expresaban polaridades distintas al hablar sobre un tópico específico.

En concreto, para determinar la relevancia de un tópico, nos basamos en el número de apariciones de dicho tópico para un tipo de usuario concreto, de tal manera que, al final, tendremos una lista de tópicos relevantes para cada tipo de usuario individual, y una lista de tópicos relevantes en general, que aúna todos los anteriores.

Adicionalmente, se realizó un proceso de agrupación de los tópicos con el que obtuvimos una lista de palabras relacionadas utilizando el corpus de palabras vectorizadas de GloVe (Pennington, Socher, y Manning, 2014). Por ejemplo, para la palabra “bar” se detectaron palabras relacionadas como, por ejemplo, “café”, “restaurant”, etc.

Una vez obtenidas la lista de tópicos relevantes y la de tópicos agrupados, seleccionamos de esta última, aquellos tópicos que se encuentran dentro de la lista de los relevantes, de modo que podamos conocer si los tópicos de nuestras frases son relevantes o están relacionados con alguno de los relevantes.

Teniendo en cuenta la relevancia del tópico y la del autor, calculamos una puntuación de relevancia para una frase, la cual utilizaremos para ordenar las frases de mayor a menor según su relevancia para el usuario.

Además, gracias al análisis del sentimiento y la detección del tópico de las frases anteriores, podemos determinar cuándo dos frases están hablando sobre un tópico específico y diferenciar si la opinión hacia ese tópico es positiva o negativa, permitiéndonos así ajustar los resúmenes a varios perfiles de usuario y compararlos entre sí. Esto se realizó a partir de un conjunto de reglas, diseñadas de forma manual, que nos permitirán identificar opiniones contradictorias (con diferente polaridad) sobre un mismo tópico en los textos de entrada para el enfoque de resúmenes propuesto. Con estas reglas se confeccionaría un conjunto de tópicos y frases sobre los que no hay un consenso entre los usuarios, para utili-

zarlos a la hora de producir el resumen final.

### 3.2.2 Selección de frases relevantes

Una vez determinada la relevancia de cada frase, el siguiente paso consistió en seleccionar las más relevantes para incorporarlas al resumen final. Para ello, se decidió seleccionar las frases que se incluyeran en estas categorías, puesto que serían las secciones en las que se organizará el resumen a generar:

- Frase de introducción
- Opiniones positivas sobre el hotel
- Opiniones negativas sobre el hotel
- Aspectos más comentados del hotel
- Diversidad de opinión (Contradicción)
- Opiniones sobre aspectos específicos (Metadatos)

Dependiendo de la sección del resumen en la que nos encontremos la selección de frases variará. Algunas secciones utilizan la totalidad de las frases analizadas para sacar estadísticas sobre los tópicos, mientras que otras escogen entre las frases con mayor frecuencia basándose en el sentimiento que muestran, o devuelven las frases de los 3 tópicos relevantes más comentados junto con el porcentaje de aparición de estos en la totalidad del texto.

Por último, hay una parte concreta que selecciona frases contradictorias dentro del corpus para mostrar al usuario la existencia de diversidad de opiniones entre los datos recolectados y evitar así que el resumen esté sesgado.

Para todas estas selecciones se utilizan una serie de plantillas definidas para cada una de las partes del resumen, las cuales explicamos a continuación en la última fase del enfoque propuesto.

### 3.3 Generación del resumen

En esta última fase, utilizamos la información recolectada en la fase de Interpretación para generar un nuevo texto, basado en plantillas, que sintetice toda la información de los comentarios en un pequeño resumen abstractivo y proporcione la información de un modo claro y conciso.

Para ello, diseñamos y construimos varias plantillas, creadas a partir de reglas combinadas con lenguaje natural predefinido, y separadas por las secciones comentadas en el apartado anterior, que se montarán una tras

otra para la obtención del resumen final. En total se crearon 17 plantillas centradas en el ámbito de las opiniones de hoteles, que, combinadas, formarían un resumen adaptado a un tipo de usuario específico.

La primera parte del resumen consiste en una frase de introducción que nos presenta el tipo de usuario<sup>6</sup>, el nombre del hotel y la opinión general del hotel con respecto a un tipo concreto de usuario. A continuación, se muestra un ejemplo, donde *travellerType*,  $X$  e  $Y$  serían valores a rellenar según los datos calculados.

- (1) The *travellerType* who stayed at  $Y$  had a generally good opinion of this hotel, as the  $X\%$  of the comments are positive.

Seguidamente se introduce el tópico más comentado entre los positivos, así como algunos ejemplos del mismo, que sustituirán a “ $Y$ ”, “ $XYZ$ ” y “ $ZYX$ ”:

- (2) The users seem comfortable talking about “ $Y$ ” because they express positiveness on the  $X\%$  of the comments related to this topic. For example, you can see that on comments like “ $XYZ$ ” or “ $ZYX$ ”.

Del mismo modo, se muestra el tópico negativo con mayor aparición junto con frases en las que se ha detectado este tópico. Tras haber introducido, respectivamente, los tópicos positivo y negativo más comentados para el hotel y el tipo de usuario elegido, se comenta el número de frases analizadas ( $X$ ) y el total de tópicos detectados ( $Y$ ), así como los 3 tópicos más comentados entre todos los existentes ( $A$ ,  $B$ , y  $C$ ), mencionando también el porcentaje de su aparición en la totalidad de las frases analizadas ( $N$ ,  $M$  y  $P$ ):

- (3) We can also see that, from the  $X$  reviews crawled for this type of traveler and the  $Y$  topics detected, the most commented aspects for this hotel were the “ $A$ ”, noticeable on the  $N\%$  of the comments, “ $B$ ” with a  $M\%$  of appearance, and “ $C$ ” with a  $P\%$ .

Seguidamente introducimos la diversidad de opiniones (“ $A$ ”, “ $B$ ”) que existe entre frases que hablan sobre el mismo tópico para un hotel específico, mostrando algunos ejemplos

<sup>6</sup>En nuestro enfoque trabajamos con cinco perfiles de usuarios: personas que viajan en familia, en pareja, solos, con amigos o que viajan por trabajo.

(“ $ABC$ ”, “ $CBA$ ”) y el porcentaje de tópicos contradictorios ( $X\%$ ) que hay en el corpus:

- (4) This type of traveller has shown different opinions about some topics like “ $A$ ” and “ $B$ ”. For the topic “ $A$ ” they comment “ $ABC$ ” and also “ $CBA$ ”. The same occurs for the  $X\%$  of the previously commented topics on this hotel, so you cannot establish a definitive conclusion about them.

Por último, mostramos una lista de los aspectos específicos de un hotel con la puntuación media obtenida entre todos los usuarios, y que hemos extraído y calculado de los metadatos.

- (5) Finally, we list below the average scores of some specific aspects of the hotel that have been ranked by the users: Service: 2/5 Cleanliness: 3/5 Value: 3/5 Sleep Quality: 3/5 Rooms: 2/5 Location: 2/5.

De esta manera juntando todos los ejemplos anteriores, tendríamos el resumen final generado por el enfoque propuesto.

#### 4 Experimentación y evaluación

A partir del corpus recopilado y del enfoque de resúmenes propuesto, se generaron un total de 50 resúmenes en inglés, 5 resúmenes por cada perfil de usuario de cada hotel del corpus.

Con el objetivo de verificar si los resúmenes generados eran adecuados y útiles, se realizó una evaluación preliminar de forma manual, en la que participaron 10 evaluadores con altos conocimientos de inglés. Los evaluadores recibieron una encuesta<sup>7</sup> que debían rellenar tras haber leído los resúmenes y cuyas preguntas estaban relacionadas con sus hábitos a la hora de buscar información sobre un hotel y elegirlo, la coherencia, corrección ortográfica y gramatical y utilidad de los resúmenes generados. Las preguntas relacionadas con la evaluación directa de los resúmenes generados se respondieron siguiendo una escala de Likert de 5 puntos. Finalmente, en la última pregunta del cuestionario de evaluación quisimos conocer cómo de fácil sería descubrir si el resumen generado había sido realizado por una persona o una máquina, también conocido como Test de Turing.

<sup>7</sup><https://goo.gl/forms/HyXFKb8E6DGFawJx2>

Según los resultados obtenidos en esta evaluación, el 90 % de los usuarios utiliza Internet para informarse sobre los hoteles en los que se alojarán durante su viaje, frente a un 10 % que prefiere preguntar directamente a amigos o familiares.

Los evaluadores también consideraron útil la posibilidad de disponer de un resumen de los comentarios de TripAdvisor, ya que el 70 % de ellos consideran que la gran cantidad de información existente hace imposible su completa lectura y comprensión.

La Figura 2 muestra la puntuación que los usuarios dieron a los resúmenes generados por el sistema propuesto, utilizando en las estadísticas los 50 resúmenes generados. Puntuaciones más altas reflejan que el resumen generado tiene mayor coherencia. Podemos observar que la mayoría de los resúmenes han sido puntuados con más de un 3 en coherencia, lo que indica que los evaluadores los han encontrado entendibles y con una estructura adecuada.

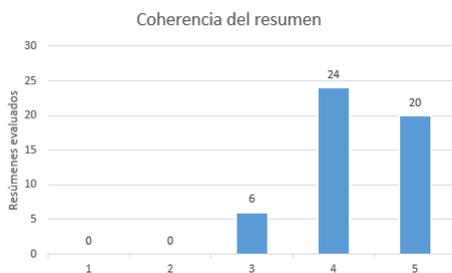


Figura 2: Grado de coherencia en los resúmenes según los evaluadores

Por otra parte, en la Figura 3 se muestra la evaluación de la falta de errores existentes en los resúmenes. A menor cantidad de errores, mayor puntuación para el resumen. Podemos observar también que la mayoría de los resúmenes han obtenido una buena puntuación con respecto a la cantidad de errores que presentan, lo que ayuda a que los usuarios los comprendan con mayor facilidad.

Sobre la utilidad de los resúmenes generados por el enfoque propuesto, los resultados de la Figura 4 indican que los evaluadores consideran los resúmenes adecuados para la toma de decisiones, lo que significa que a pesar de haber sido realizados de forma automática, serían de gran ayuda para sintetizar la información disponible y poder proporcionar una idea del hotel en cuestión.

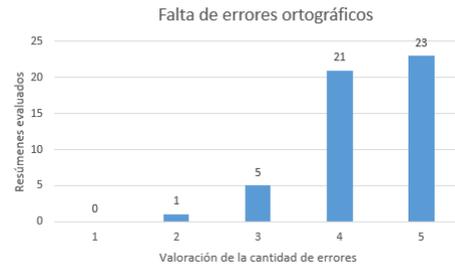


Figura 3: Ausencia de errores ortográficos según los evaluadores

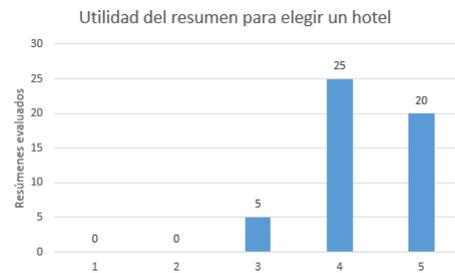


Figura 4: Grado de utilidad del resumen a la hora de elección de un hotel según los evaluadores

Sobre el Test de Turing realizado, en un 48 % de los casos, los evaluadores no consiguieron distinguir correctamente si el texto había sido escrito por un humano o no, hecho que nos indica que el enfoque de generación propuesto, en su estado actual, es bastante competitivo y tiene potencial para ser integrado en una aplicación real.

Sin embargo, al evaluar únicamente 5 resúmenes por cada hotel es difícil que los evaluadores no se den cuenta de los resúmenes han sido generados de forma automática, ya que todos siguen una estructura y un patrón similar.

Por último, remarcar que, además de la evaluación de los usuarios, también se realizó un análisis en detalle de cada uno de los resúmenes generados, para detectar las limitaciones de la propuesta y posibles áreas de mejora como, por ejemplo, la falta de detección de aspectos característicos que sean poco mencionados o la posibilidad de crear el texto sin el uso de plantillas.

## 5 Conclusión y trabajo Futuro

En este artículo se ha propuesto un enfoque de generación de resúmenes a partir de con-

tenido generado por usuario, con el fin de proporcionar al usuario un texto que aúne los diversos puntos de vista de las opiniones existentes y evitar, así, posibles sesgos que influyan a la hora de tomar decisiones. Concretamente, nos hemos centrado en opiniones sobre hoteles, recopilando un corpus de prueba, y aplicando técnicas de análisis lingüístico a distintos niveles junto con heurísticas para identificar información relevante y seleccionar aquella que podría ser más importante para el usuario.

Los resultados obtenidos a partir de la evaluación de los resúmenes generados han sido satisfactorios, por lo que se puede concluir que el enfoque propuesto es capaz de generar resúmenes de una calidad suficientemente buena para que los usuarios queden satisfechos. Sin embargo, el enfoque propuesto se podría mejorar integrando algunos aspectos que nos planteamos como trabajo futuro. El primer aspecto a corto plazo sería mejorar el análisis de sentimientos, teniendo en cuenta la negación, puesto que este fenómeno puede cambiar completamente la polaridad de una frase. A medio y largo plazo, nos gustaría adaptar el enfoque propuesto para el español, y probarlo y evaluarlo, en inglés y español, no sólo con reseñas de hoteles, sino también en otros ámbitos similares como opiniones sobre restaurantes, productos o servicios, de los que existen gran cantidad de información y mucha subjetividad en función de la experiencia.

### **Agradecimientos**

Este proyecto ha sido financiado parcialmente por la Generalitat Valenciana a través del proyecto PROMETEU/2018/089.

### **Bibliografía**

- Bird, S. y E. Loper. 2004. NLTK: the natural language toolkit. En *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, página 31. Association for Computational Linguistics.
- Denecke, K. 2009. Are SentiWordNet scores suited for multi-domain sentiment classification? En *Proceedings of 4th International Conference on Digital Information Management*, páginas 1–6. IEEE.
- Esteban, A. y E. Lloret. 2017a. Propuesta y desarrollo de una aproximación de generación de resúmenes abstractivos multigénero. *Procesamiento del Lenguaje Natural*, 58:53–60.
- Esteban, A. y E. Lloret. 2017b. TravelSum: A spanish summarization application focused on the tourism sector. *Procesamiento del Lenguaje Natural*, 59:159–162.
- Esuli, A. y F. Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. En *Proceedings of the 5th International Conference on Language Resources and Evaluation*, páginas 417–422.
- Hovy, E. 2003. Text summarization. En *The Oxford Handbook of Computational Linguistics 2nd edition*. Oxford University Press.
- Lloret, E. 2016. Introducing the key Stages for Addressing Multi-perspective Summarization. En *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, páginas 321–326.
- Luo, C., Y. Lan, C. Wang, y L. Ma. 2013. The effect of information consistency and information aggregation on eWOM readers' perception of information overload. En *Proceedings of Pacific Asia Conference on Information Systems*, página 180.
- McCarthy, D. y R. Navigli. 2007. Word sense disambiguation: An overview. *Proceedings of the 4th International Workshop on Semantic Evaluations*, páginas 7–12.
- Miller, G. 1998. *WordNet: An electronic lexical database*. MIT press.
- Paulus, R., C. Xiong, y R. Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.
- Pennington, J., R. Socher, y C. D. Manning. 2014. Glove: Global vectors for word representation. En *Proceedings of Empirical Methods in Natural Language Processing*, páginas 1532–1543.
- Řehůřek, R. y P. Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. En *Proceedings of the LREC Workshop on New Challenges for NLP Frameworks*, páginas 45–50.

# A different description of orientation in sign languages

## *Una descripción distinta de la orientación en las lenguas de signos*

Antonio F. G. Sevilla<sup>1,2</sup>, Jose María Lahoz-Bengoechea<sup>3</sup>

<sup>1</sup>Instituto de Tecnología del Conocimiento

Facultad de Psicología, Lateral 2, Campus de Somosaguas 28223 Pozuelo de Alarcón, Spain

<sup>2</sup>Departamento de Ingeniería del Software e Inteligencia Artificial

Facultad de Informática, c/ Prof. José García Santesmases 9, 28040 Madrid, Spain

<sup>3</sup>Departamento de Lengua Española y Teoría de la Literatura

Facultad de Filología, edificio D, c/ Prof. Aranguren s/n, 28040 Madrid, Spain

Universidad Complutense de Madrid

<sup>1</sup>afgs@ucm.es, <sup>3</sup>jmlahoz@ucm.es

**Abstract:** Sign languages are a very interesting object of linguistic study, posing challenges not present in oral languages. One of these challenges is describing and transcribing the internal structure of the language in a way that is adequate to its characteristics but also compatible with existing linguistic practice. The phonology of sign languages is of special interest. We focus on one phonological feature: that of hand orientation. We propose an interpretation and description system that better captures underlying meaning and structure, and that is more appropriate for its formal and computational treatment.

**Keywords:** Sign language, transcription, orientation, phonology

**Resumen:** Las lenguas de signos son un objeto de estudio lingüístico de gran interés. Presentan retos y dificultades distintos a los de las lenguas orales, como describir y transcribir la estructura interna de la lengua de una manera adecuada a sus características únicas pero también compatible con la práctica lingüística actual. El caso de la fonología es especialmente interesante. Nos centramos en un rasgo fonológico concreto: la orientación de la mano. Proponemos una interpretación y un sistema de descripción que capturan mejor la semántica y la estructura subyacente, y que además permiten un tratamiento formal y computacional más adecuado.

**Palabras clave:** Lengua de signos, transcripción, orientación, fonología

### 1 Introduction

Recently, sign languages are of increasing interest for the linguistic community, as well as society in general. As languages for the Deaf and hearing-impaired, social goals of accessibility and equal opportunity make the general public more aware of their existence and overall characteristics. From a scientific point of view, sign languages completely ignore sound –for obvious reasons– and their essentially multimodal nature presents challenges and opportunities in their understanding and formal description. The fact that they are natural languages, evolved within their communities by natural processes, also make these linguistic inquiries important from a general linguistic or psycho-linguistic point of

view. Discoveries and advances made in sign language studies may bring along advances for language in general and the language capacity in humans. See Senghas and Coppola (2001) for an interesting example.

Nowadays, linguistics rely heavily on computers, and the digital treatment of linguistic data. Formal accounts of language are expected to be computationally treatable, even if only in theory. And the engineering side of the issue is of ever increasing importance, the ability of computers to understand and process natural language gaining both efficiency and public awareness every year. However, sign languages present a sizeable challenge in this department. Most of NLP technologies and algorithms are based on the assumption that there exists an accurate, or at least rea-

sonable, representation of language using sequences of characters, which is not (yet) true for sign language.

Much existing work to computational processing of sign languages takes a word-based approach. It works as in languages where the written form of a word has moderate to low relation to the oral form, like those with ideographic writing (Chinese, Japanese), or, maybe not so extremely, English. The idea is to observe the physical realization of a word (or sign) and transcribe it to a character-based representation, see Starner, Weaver, and Pentland (1998) or Karayilan and Kiliç (2017). In the case of sign languages, this approach is very limited. The visual signal is more complex than the audible one, and this also allows for information to sometimes appear in disjoint, parallel manifestations (think of two hands working together to perform a single sign). Therefore, these approaches are often limited to the recognition of a fixed vocabulary of signs, and eschew any structure or inflection they might present.

Signs have a very rich inner complexity. This structure, which can sometimes parallel that of oral languages, has not only descriptive importance, but is often lexically or grammatically significant. Even at the phonological level, sign languages are organized in a manner not immediately equivalent to that of oral ones. In the literature, signs are classified phonologically using a number of features, including location, shape, contact, movement, and orientation (Liddell and Johnson, 1989). These describe the configuration and movement of the hand in space, and while there are additional non-manual features of sign language, the hands are the most salient and important elements.

In this paper we focus on orientation. Orientation is a simple but essential feature of sign language at the phonological level. It indicates the rotation of the hand as a whole, without regard for its shape and the individual positions of the fingers. Some examples can be seen in Figure 1, along with its “traditional” notation in SignWriting. Orientation is essential in the sense that it is necessary, it cannot be omitted and it often contributes to meaning in a significant way. There exist minimal pairs that are distinguished only by orientation: an example from Spanish Sign Language<sup>1</sup> is that of the number “1”, the letter “g” in fingerspelling, and the sign “today”

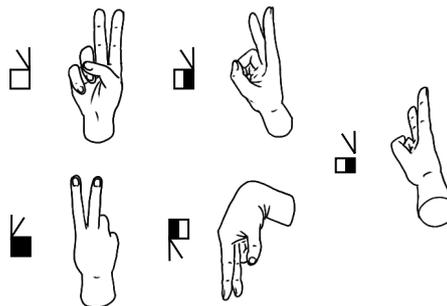


Figure 1: Some orientations of the hand, with the shape of the number two. The SignWriting representation is given to the left of each drawing

(Herrero Blanco, 2009). As a feature, orientation can be assimilated, for example in the formation of compound words, and is subject to substitution in speech errors (Sandler and Lillo-Martin, 2006, Chap. 10).

However, the “traditional” description of orientation in sign language is, to us, not sufficient. It presents a number of problems, both from a linguistic and a computational point of view. In this paper, we present a different description of the feature of orientation, which provides some improvements in its understanding and how it contributes to meaning. Our proposal is also very computer-friendly, meaning its formalization and computational treatment are also strong points for its adoption. We accompany our theoretical analysis with a proposed notation system, which allows its better input and storage in digital media.

In Section 2 we present the SignWriting approach. In Section 3, our theoretical interpretation is explicated, while Section 4 presents the provisional notation for its representation. Finally, in Section 5 we draw some conclusions and discuss some possible extensions of this work.

## 2 SignWriting approach

Existing approaches for the description of orientation tend to look at the hand by itself, and observe how it is rotated. In SignWrit-

<sup>1</sup>Throughout this article, we use drawings of hands in space to illustrate some points. Most of the time, these hand configurations are not signs, and we believe that the feature under discussion is “low-level” enough that our proposal is language-independent. However, most examples come from Spanish Sign Language, and there may be language-dependent phonotactic constraints we are unaware of.

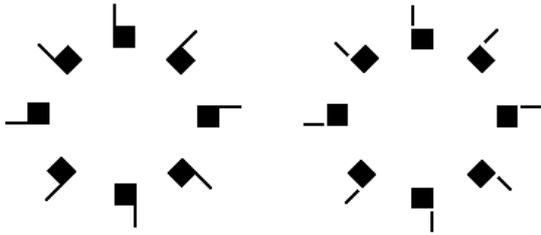


Figure 2: Eight possible rotations on the vertical plane and on the horizontal plane, respectively

ing (Sutton, 2009), a widely spread notation system, a color code is used. White identifies the palm of the hand, black the reverse. This allows us to distinguish signs by noting which side of the hand the signer can see, and if both, the exact profile is determined by placing them in an iconic way that transcribes how the hand is seen. Additionally, SignWriting makes a difference between hand configurations located on the vertical plane and those on the horizontal plane (by convention, the latter case is represented with the fingers detached from the hand root). Furthermore, rotation is allowed on each of those planes (Figure 2). In sum, the hand can rotate almost freely in the X, Y and Z axes and, in order to fully specify orientation, SignWriting approaches usually draw the hand in an iconic sign space, which has its own tricks to transform the three dimensions into two.

This approach is useful in that it is enough to fully capture the possible realizations, and its iconicity makes it easily understandable and transparent. However, it has some limitations. First, it is tied to the graphical representation. A more abstract approach that can be transcribed with Latin characters would be useful, especially for enabling computational treatment.

Second, sometimes it is possible to represent the exact same handshape and orientation in two different ways, just depending on whether the perspective is from the vertical or the horizontal plane (both configurations in Figure 3 are identical). This lack of biunivocity in the transcription unnecessarily increases computational complexity, inasmuch as it requires to store more configurations, or to add a layer of interpretation to undo the duplication. From the viewpoint of generation, it yields underspecification of expected



Figure 3: The palm facing the signer, with the fingers pointing left. These are equivalent SignWriting transcriptions for the same handshape and orientation. If the signer looks forward, it is a white located on the vertical plane. If the gaze is directed from top down, it can also be interpreted as a black-white situated on the horizontal plane

results.

It should also be noted that the graphical approach does not capture the underlying meaning structure properly. It is too tied to the realization, which makes it cumbersome in some cases.

For example, in Spanish Sign Language, there is a sign with the general meaning of “to ask”. It is a simple sign, with the hand in the shape of a “Q”<sup>2</sup>, and two repetitions of a slight movement whose direction depends on the identity of the grammatical subject and the (indirect) object. For the signer, this is not a complication, but rather makes a lot of sense. The sign “comes from” the asker, and “goes to” the one being asked. But it is not only the movement that follows this pattern; the orientation does too. The “Q” hand is pointed horizontally, and the finger tips point toward the askee, as in Figure 4.

This means that for “I ask you”, the sign is black, and for “you ask me”, it is white. One could say that it is normal for inflected forms to have to be listed separately, but what with “I ask him”, or “She asks you”? Orientation is more subtle here, requiring the full expressive power of the graphical description to be transcribed, and then, only capturing individual utterances.

It is clear that in the “ask” sign itself the syntactic spaces of subject and object are embedded in the orientation of the hand. The following proposal for transcription of orientation can capture this, and has some additional advantages.

While other approaches for sign language notation are not as heavily reliant on the

<sup>2</sup>The shape of the letter “Q” in the Spanish Sign Language fingerspelling alphabet. It is a configuration of all the fingers flexed at the first falanx, and the thumb lying against their tip. A drawing can be seen in Figure 4.

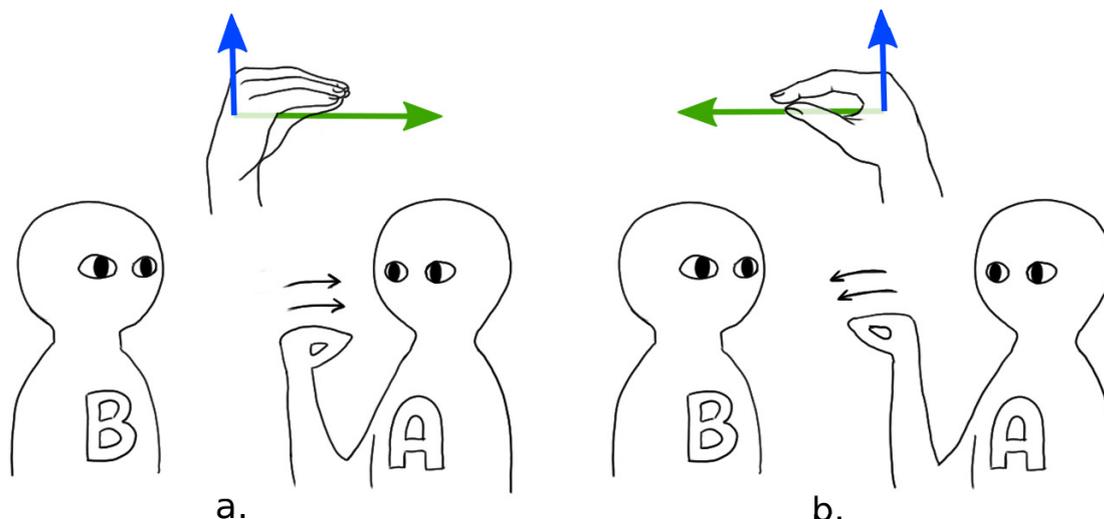


Figure 4: Two situations, where A is talking to B. In a., A says “You-Ask-I”, in b., A says “I-Ask-You”. This sign, for the verb “to ask” can then be understood as inflecting, and having different forms with different orientation. In our proposal, this can be succinctly expressed by saying that the palmar vector is oriented from the subject to the (indirect) object of the verb

picture or drawing of the hand orientation, similar criticism as with SignWriting can be made. In the HamNoSys notation system (Hanke, 2004), features are transcribed separately, but still with pictograms and symbols. Hand orientation is also treated holistically and by itself, with no reference to space, so the issues with the “ask” sign remain. In Stokoe notation (Stokoe, Casterline, and Croneberg, 1976), a linguistically motivated transcription system, orientation is described with a fixed set of symbols, again limiting the expression of the inner structure.

### 3 Proposed interpretation

Barring physical constraints that limit the range somewhat, the hand has free rotation in space, being able to rotate around the three directions. But the palm is basically flat and, in order to describe the position and orientation of a flat object in a 3D space, mathematically it is enough with two vectors. We can ignore the fingers for now, since their position is taken into account by the hand shape, a different feature.

#### 3.1 Hand vectors

The hand has a number of natural vectors that we can identify. In this approach, the three most natural and useful are selected. The first two, both of equal and great importance, are the “distal” and “palmar” vectors.

The distal vector is the one that goes from the wrist to the fingers, parallel to the palm. It points to where the index finger points when fully extended.

The palmar vector is perpendicular to the distal one, and in general perpendicular to the plane of the palm itself. It points to the inside of the closed hand, or what we as humans generally consider the palm to “point” to.

The lateral vector is perpendicular to the previous two. It points where the thumb points when the hand is open or, for example, in the thumbs-up gesture. In Figure 5 and in Figure 6, the three vectors are represented in a few different situations.

Since the left hand is a mirror image of the right hand, if the orientation of any two vectors is shared between both hands, the respective third vectors will point opposite to each other. For example, if we assume unit vectors and we take the cross product of distal and palmar, it will be equal to the lateral vector for the right hand, and opposite to it for the left hand:

$$\overrightarrow{distal} \times \overrightarrow{palmar} = \begin{cases} \overrightarrow{lateral} & \text{right hand} \\ -\overrightarrow{lateral} & \text{left hand} \end{cases}$$

But if two vectors are enough to describe the orientation, why define three? The mathematical answer is that this way we may fully describe our variation space (three dimen-

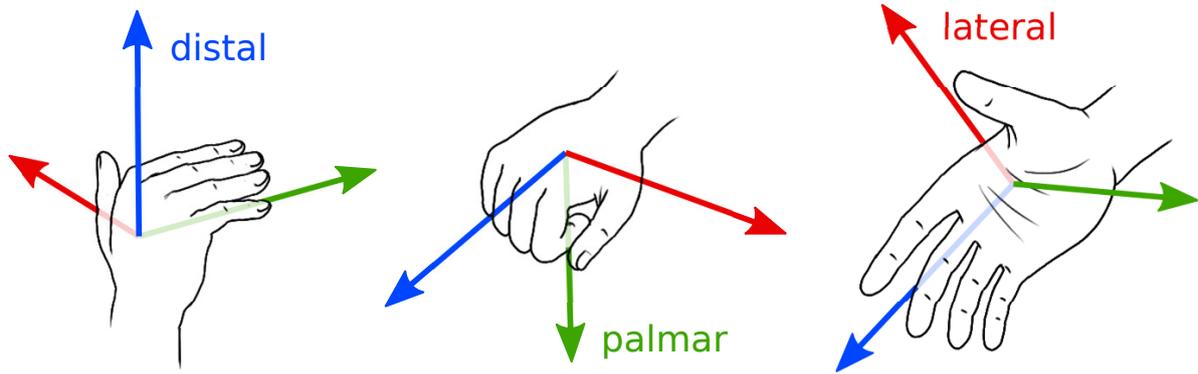


Figure 5: The “hand vectors” we propose to use for describing orientation. The distal vector points in the direction of the extended fingers, the palmar one in the direction of the palm, and the lateral one towards the extended thumb. Notice that flexion of the fingers does not affect the direction of the vectors

sions, three vectors). The linguistic answer is that it allows us to describe some signs in a more succinct and semantic way.

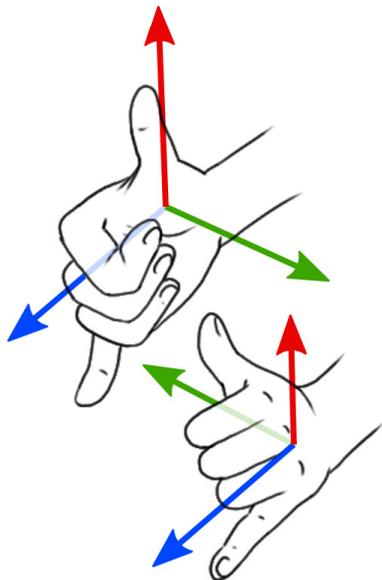


Figure 6: Hand vectors in the two hands. Notice that the vector systems are not equivalent, one can not be rotated to match the other

### 3.2 Underspecification

For most signs, specifying the distal and palmar vectors will be enough. Together, they both solve the spatial rotation of the hand and the black/white distinction, in a natural way. Numbers, for example, are distal “up” and palmar “front” (see Figure 1 again). In

Spanish Sign Language, some numbers are palmar “back”, however. This distinction is as easy to make with this approach as with the white/black one. Furthermore, the distal vector may be left out in some informal accounts, since it is the most natural when producing cardinal numbers (provided they are not integrated with some other morpheme). The only salient information would in that case be the distinction between palmar “front” or “back”, and only that vector needs to be noted.

And here lies one of the strong points of this approach. Underspecification is an extremely important characteristic of the phonology of languages (Steriade, 1995). With the graphical approaches, however, we are forced to commit to a particular realization of the sign, including redundant information and even slight, uninformative phonetic details.

With the proposed approach, only the necessary information has to be specified. To take an example from a gesture in popular culture, rather than a sign, in the “thumbs-up” gesture, the important orientational information is that the lateral vector (thumb) points up. Distal and palmar vectors are irrelevant, and so they can just be omitted. Upon realization, the producer can select the orientation that best fits the context, or the more comfortable one for the situation.

This “lateral” extra axis is also useful for signs presenting some movement of the hand that affects orientation. In the “there is not” sign of Spanish Sign Language, for exam-

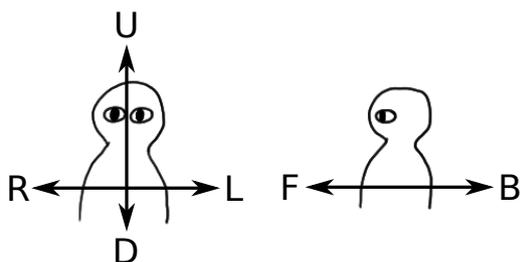


Figure 7: Points in the topographical sign-space, corresponding to the general directions Up, Down, Left, Right, Front and Back. Note that space is always transcribed from the point of view of the signer

ple, both distal and palmar vectors change. The lateral vector remains static, however, pointing backwards. We can thus understand the orientation and movement combined as a “minus quarter” rotation around the lateral, which points back, or alternatively distal from up to left, lateral back: (U,L)::B<sup>3</sup>.

This is the underspecified notation for (U,L):(L,D):B, and it could also have been written (U,L):(L,D), but the common practice should be to underspecify in the most economic way possible. Note that this apparent diversity of transcriptions is different from the previously mentioned issue of Sign-Writing regarding lack of biunivocity. The full vectorial transcription of orientation always entails three vectors and is unique for each specific configuration. However, given that the vector system only has two degrees of freedom (the position of one vector is always determined by that of the other two), then the most convenient vector can be omitted, following the principles of underspecification.

#### 4 Proposed notation

Accompanying the different approach to orientation that has been described, we also propose a succinct and expressive notation that can be used to transcribe it. We rely on a theoretical existing notation for space, which is out of the scope of this paper. For now, we will use a simple one, specifying gen-

<sup>3</sup>Here, we separate the vectors with ‘:’ in the order distal, palmar, lateral, but we leave out the palmar since it is predictable given the other two. The values inside the parentheses represent the consecutive orientations.

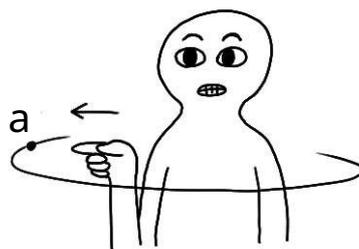


Figure 8: A 3rd person “placed” in syntactic sign space. The facial features are the mark of a 3rd person pronoun, with the finger pointing to the antecedent’s location in sign-space (or postcedent if it will be specified later)

eral directions<sup>4</sup> with uppercase letters (Up, Down, Left, Right, Front, Back), as in Figure 7, and pronominal spaces (“placed” constituents) with lowercase letters not belonging to any general direction (a, b, c, x, y... ). In other words, lowercase letters represent variables that are subject to deictic or anaphoric interpretation. Figure 8 shows an example of a 3rd person in sign-space.

Orientation can then be notated using colons to separate the spaces to which the hand vectors point, starting with the distal vector, following with the palmar and finishing with the lateral. To leave a vector unspecified, the corresponding colon is used, but no letters in the place where they should be. As a shorthand, if distal and palmar vectors are specified, the second colon and lateral space can be omitted.

Some relevant examples can be seen in Table 1. In Table 2, a more complete enumeration of different orientations and their vector transcription is given.

#### 5 Discussion and Conclusions

Description and transcription are two fundamental endeavors in linguistics, and intimately related. A good description enables natural and expressive transcription, and proper transcription shows the strengths and weaknesses of a particular conceptual model. In a feature seemingly simple such as hand orientation in sign language, we have shown that changing the way it is interpreted allows to discover underlying structure and meaning.

<sup>4</sup>Points in the “topographical” sign-space.

Examples	SignWriting <sup>i</sup>	Vectors <sup>ii</sup>
Letter a, numbers		U:F
Past		(U,B)::R
More, “thumbs-up”, work		::U
Bicycle, Sit		F:L
Child		F:D
Letter g		L:B
I ask a, I understand a	–	U:a
a (anaphora) <sup>iii</sup>	–	a:

<sup>i</sup> Only orientation feature, plus the shape from the first listed example (the shape cannot be omitted in this notation).

<sup>ii</sup> The parenthesis and comma are a shorthand for bisyllabic transitions.

<sup>iii</sup> As in Figure 8, indicating a 3rd person by pointing. The shape is “point” or “1” (extended index), and the only relevant orientation is the distal vector.

Table 1: Examples of the orientation feature in different signs, and its transcription in SignWriting and vector notation

Our representation highlights the pieces of information that are phonologically relevant. Redundant or predictable phonetic aspects must be filled in at a later stage in the derivation, and may be subject to particular realizational constraints or variants. This move brings the analysis of sign languages closer to that of oral languages, by capturing general principles of phonology, like underspecification.

In addition, we want to point out the very natural treatment of this representation that can be done with computers. The abstract or symbolic notation is much more amenable to digital processing than graphical representations, and the use of vectors and space locations may make mathematical treatment of animated agents easier, or even help with computer-vision recognition of signed language.

One issue remains that we can see. In some signs, the vector which matters is that of the index finger. Think for example of the

basic pointing sign, where an element of the sign space is referred to. Here, the important direction is the tip of the index finger. If this finger is fully extended, that coincides with our distal axis. However, in a natural realization of the sign, the finger can be flexed, taking it out of alignment with the distal vector.

Some possibilities arise. The easiest, of course, is to think that the ideal sign is the one with the index fully extended, and the distal vector pointing properly, but “laziness” and other realization constraints mean that the actual hand shape is often more relaxed.

Other possibility is to add another axis, corresponding to the extended fingertip. It would then have to be seen if others fingers also would benefit from this treatment, how many of them would be needed, and how to describe this phenomenon with a concise but comfortable notation.

In order to decide this issue, more data and study would be needed. However, the underlying technique of using vectors pointing to space locations to specify orientation seems to be sound, useful and expressive.

### Acknowledgements

This research is partially supported by the IDiLyCo project (TIN2015-66655-R) funded by the Spanish Ministry of Economy, Industry and Competitiveness.

### References

- Hanke, T. 2004. HamNoSys-Representing sign language data in language resources and language processing contexts. In *Proceedings of the Workshop on Representation and Processing of Sign Language, Workshop to the fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 1–6.
- Herrero Blanco, Á. L. 2009. *Gramática didáctica de la lengua de signos española (LSE)*. SM, Madrid.
- Karayilan, T. and Ö. Kiliç. 2017. Sign language recognition. *2nd International Conference on Computer Science and Engineering, UBMK 2017*, pages 1122–1126.
- Liddell, S. K. and R. E. Johnson. 1989. American Sign Language: The phonological base. *Sign Language Studies*, 64(1):195–277.

SignWriting	Vector	SignWriting	Vector	SignWriting	Vector	SignWriting	Vector
	U:B		U:L		U:F		U:R
	L:B <sub>a</sub>		L:D <sub>b</sub>		L:F <sub>c</sub>		L:U <sub>d</sub>
	D:B		D:R		D:F		D:L
	R:B <sub>e</sub>		R:U <sub>f</sub>		R:F <sub>g</sub>		R:D <sub>h</sub>
	F:U		F:L		F:D		F:R
	L:U <sub>d</sub>		L:B <sub>a</sub>		L:D <sub>b</sub>		L:F <sub>c</sub>
	B:U		B:R		B:D		B:L
	R:U <sub>f</sub>		R:F <sub>g</sub>		R:D <sub>h</sub>		R:B <sub>e</sub>

Table 2: Main variations of orientation and their corresponding vector notation, using distal and palmar vectors. Subscripts signal orientations that have a twofold representation in SignWriting (subscript *a* is the same as the example previously presented in Figure 3). The generalization is that all orientations with distal either Left or Right have two possible SignWriting notations, because they refer to points where the vertical and the horizontal planes converge

Sandler, W. and D. Lillo-Martin. 2006. *Sign language and linguistic universals*. Cambridge University Press, Cambridge.

Senghas, A. and M. Coppola. 2001. Children creating language: How Nicaraguan Sign Language acquired a spatial grammar. *Psychological Science*, 12(4):323–328.

Starner, T., J. Weaver, and A. Pentland. 1998. Real-time American Sign Language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375.

Steriade, D. 1995. Underspecification and markedness. In J. A. Goldsmith, editor, *The handbook of phonological theory*. Blackwell, Oxford, pages 114–174.

Stokoe, W. C., D. C. Casterline, and C. G. Croneberg. 1976. *A dictionary of American Sign Language on linguistic principles*. Linstok Press, Washington, DC.

Sutton, V. 2009. *SignWriting: Sign languages are written languages!* The SignWriting Press, La Jolla, CA.

## **Detección de plagio translingüe con grafos semánticos: experimentando con recursos en abierto**

### *Detection of translingual plagiarism with semantic graphs: experimenting with open resources*

**Ana García-Serrano, Antonio Menta Garuz**

UNED, Universidad Nacional de Educación a Distancia

ETSI Informática, C/ Juan del Rosal 16, 28040 Madrid

[agarcia@lsi.uned.es](mailto:agarcia@lsi.uned.es), [mentared@gmail.com](mailto:mentared@gmail.com)

**Resumen:** Hoy en día el idioma ha dejado de ser una barrera para plagiar documentos disponibles en Internet. Tras enfoques probabilísticos ya clásicos que no alcanzan buenos resultados con documentos multilingües con paráfrasis (Barrón-Cedeño, 2012), aparecen trabajos que, utilizando grafos de conocimiento, aumentan la capacidad semántica del análisis de las oraciones y mejoran los resultados de detección de plagio. Además, actualmente hay recursos lingüísticos, basados en el conocimiento, o de desarrollo de software que están disponibles para la experimentación, una vez decidido cuál de ellos elegir, cuáles están realmente disponibles en abierto, qué eficiencia aportan si se integran en la experimentación planteada, o qué tipo de características debe tener el ordenador o el servidor necesario para la investigación. Este trabajo plantea una investigación experimental para la detección de plagio translingüe siguiendo una línea de investigación y utilizando recursos disponibles en abierto. Los resultados alcanzan el estado del arte, y esperamos que el planteamiento seguido, el análisis justificado y las dificultades técnicas reportadas, acercará a los lectores la metodología necesaria en este tipo de experimentaciones y permitirá planificar sus trabajos futuros. El software desarrollado está disponible en abierto.

**Palabras clave:** Plagio translingüe, recursos lingüísticos, recursos en la red, experimentación, desarrollo de software

**Abstract:** Today the language has ceased to be a barrier to plagiarize documents available on the Internet. After classic probabilistic approaches that do not achieve good results with multilingual documents with paraphrasing (Barrón-Cedeño, 2012), there are works that, using knowledge graphs, increase the semantic ability in the analysis of sentences and improve the results of plagiarism detection. In addition, currently in linguistic engineering there are linguistic or knowledge-based resources, or software development resources that are available to experimentation once decided, which ones to choose, which ones are available, what efficiency they provide if they are integrated into the proposed experimentation, or what kind of features the computer or server should have to the investigation. This work proposes an experimental investigation into a concrete problem, the detection of translingual plagiarism following a line of research and using open resources. The results reach the state of the art, and we hope that the followed approach, the justified analysis and the technical difficulties reported, will bring readers closer to the methodology needed in this type of experimentation and will allow planning their future works. The software developed is available in open.

**Keywords:** Translingual plagiarism, linguistic resources, linked data, experimentation, software development

## 1 Introducción

Hay una diferencia importante entre inspirarse en obras de terceros y copiar el contenido intencionadamente, “*Plagiar es reusar las ideas, procesos, resultados o palabras de alguien más sin mencionar explícitamente a la fuente y a su autor*” (Barrón-Cedeño et al., 2013).

Comas y Sureda (2008) constatan que el 61% de los universitarios españoles confiesa haber copiado de internet y el 3,3% incluso haber comprado documentos. Diez años después, en el mundo la proporción ha aumentado al 85% en estudiantes (Eaton et al., 2017) y el idioma ha dejado de ser una barrera. Son necesarias herramientas automáticas capaces de detectar los posibles casos de plagio, aunque la decisión final del mismo debería ser tomada por expertos en la materia.

Tras enfoques probabilísticos ya clásicos como la utilización de n-gramas a nivel de carácter o de corpus paralelos para cada idioma que no alcanzan buenos resultados con documentos multilingües con paráfrasis (Barrón-Cedeño, 2012), aparecen trabajos que, utilizando grafos de conocimiento, aumentan la capacidad semántica en el análisis de las oraciones y mejoran los resultados de detección de plagio. Además, actualmente en ingeniería lingüística hay un gran número de recursos lingüísticos, basados en el conocimiento *Linked Data* (LD) y recursos para desarrollo de software, que están disponibles para llevar a cabo las tareas de experimentación exigibles en este campo de investigación.

Pero hay preguntas que resolver a lo largo del proceso de desarrollo, como cuál elegir, cuáles están realmente disponibles, qué eficiencia aportan si se integran en la experimentación planteada, o qué tipo de características debe tener el ordenador o servidor, y todas ellas, sin conocer si los resultados serán los esperados.

Los objetivos de este artículo son tanto avanzar en la investigación de la detección translingüe, como mostrar las decisiones que exige una experimentación científica usando recursos en abierto.

En lo que sigue, se identifican los tipos de plagio, se presenta una breve revisión del estado del arte, así como la serie de recursos PAN y algunas herramientas automáticas actuales para

detección de plagio, que muestran que el problema sigue abierto.

A continuación, siguiendo la línea de investigación realizada por Franco-Salvador et al. (2016a) y Franco-Salvador (2017), se describe la propuesta de detección de plagio translingüe utilizando recursos disponibles como son *Freeling* para el análisis lingüístico de textos multilingües, *BabelNet*, un diccionario semántico multilingüe que promete abstraer del idioma de los conceptos que aparecen en los textos, o recursos de software como *GraphStream*, para la gestión de grafos de conocimiento (Menta, 2018).

Finalmente se incluyen los detalles de las pruebas y resultados obtenidos con el prototipo desarrollado y las conclusiones.

## 2 Trabajos relacionados

La detección de plagio admite clasificaciones desde diferentes puntos de vista. En (Martin, 2004) se clasifican según el objetivo del plagio: (1) plagio de ideas; (2) plagio palabra por palabra (sin el texto con comillas); (3) plagio de fuentes y (4) plagio de autoría.

El estado del arte también se puede clasificar según las aproximaciones utilizadas, orientadas al estudio de características internas del documento (longitud de palabras, frecuencia de uso, número de adjetivos, etc.) o al estudio de características externas (cálculo de la similitud con fragmentos de terceros, distribución similar de palabras, etc.) (Meyer et al., 2007). El primer tipo se conoce como detección intrínseca y el segundo como detección externa, al recurrir a un conjunto de documentos externos.

O bien se pueden clasificar según el modo de copia, exacta o modificada (paráfrasis) (Barrón-Cedeño et al., 2013). En el segundo caso se modifica el texto sustituyendo ciertas partes con significado similar o eliminando algún elemento, y la detección se centra en la semejanza entre fragmentos de texto.

En Franco-Salvador et al. (2012 y 2016b), se clasifican los enfoques para detección de plagio translingüe en modelos basados en: (1) diccionarios, reglas y tesauros lingüísticos que traducen de forma aislada conceptos y palabras; (2) la sintaxis de cada documento; (3) corpus comparables, con documentos en diferentes idiomas que describen de forma aproximada el mismo contenido y (4) corpus paralelos que contienen documentos con el contenido exacto en diferentes idiomas.

Este trabajo de detección de plagio translingüe (EN-ES) es clasificable como detección de plagio de ideas, externa, de copia modificada o paráfrasis y detección semántica de plagio utilizando un diccionario enciclopédico del LD que permite comparar los documentos sospechosos con los del corpus.

## 2.1 Herramientas para la detección automática de plagio

La diferencia en la calidad, cobertura y precio de las herramientas automáticas es grande, aunque el modelo de negocio sea muy similar, y depende de la eficiencia y la calidad del corpus. En la mayoría de ellas se desconocen las técnicas que se utilizan para descubrir el plagio.

En Nahas (2017) se describen varias herramientas en la red, tanto gratuitas como de pago, que se comparan con los siguientes aspectos: (As1) Utilizable para la prevención de casos de plagio académicos; (As2) Utilizable sin necesidad de ser descargada por el usuario final (integrada en los servidores de correo, página web y otras); (As3) Corpus de internet o/y propio (como otros trabajos académicos) y (As4) Aporta otras funcionalidades.

Los detalles y comentarios de las cinco herramientas seleccionadas son los siguientes:

**Urkund**<sup>1</sup>. Es As1, As2, As3. Actualmente disponen de más de 23 millones de documentos.

**Turnitin**<sup>2</sup>. Es As1, As3, As4. La aplicación busca coincidencias entre 61.000 millones de páginas indexadas, más de 600 millones de trabajos de estudiantes y 150 millones de artículos, libros y periódicos. Destaca en la buena detección de copia exacta (analiza grupos de ocho a diez palabras). Prácticamente es la única herramienta capaz de detectar plagio translingüe, traduciendo los contenidos al inglés y realizando detección monolingüe en inglés.

**Unplag**<sup>3</sup>. Es As1, As3, As4. Centrado en trabajos académicos, realiza una búsqueda de fragmentos del documento con los índices web de Bing y Yahoo. Permite al usuario analizar hasta cinco documentos de forma simultánea.

**PlagiarismCheck**<sup>4</sup> Es As1, As3, As4. Muy efectiva tanto para académicos como para propietarios de sitios web, redactores, bloggers y otros, y revisa los textos web para ver si se

roban en línea. Indica que detecta todos los tipos de plagio.

**Plagiarisma**. Es As1, As2, As3, As4. Página web que permite subir archivos o pegar directamente el texto (max. de 1000 caracteres en cada prueba) (Krizkova et al., 2016).

Por lo tanto, la detección de plagio translingüe con paráfrasis es un tema abierto también entre la mayoría de las herramientas comerciales.

## 2.2 Recursos PAN

La iniciativa PAN (*Uncovering Plagiarism Authorship and Social Software Misuse*), es un referente internacional en el área y se creó en 2007 como evento científico para promover la investigación forense de textos digitales. En 2010 se desarrolló un *framework*, que facilita la reproducibilidad, al incluir un corpus y una métrica de evaluación (Potthast et al., 2010 y 2011b).

En la primera edición se desarrolló un corpus basado en un conjunto de libros del proyecto Gutenberg, grandes trabajos de la literatura clásica, la mayoría en inglés. Como estos libros no contienen casos de plagio conocido, se crearon manual y artificialmente los casos de plagio. Del total de libros del proyecto Gutenberg, el corpus PAN-PC-10 contiene 22.000 libros en inglés, 520 en alemán y 210 en español.

En la segunda edición de la competición se mejoró el corpus con la inclusión de fragmentos sospechosos en documentos con temas similares, en vez de sólo incluirlos de forma aleatoria en cualquier tipo de documento. Para hacerlo, el contenido del corpus PAN-PC-10 se dividió en clústeres según temas y la mitad de los fragmentos artificiales creados son insertados en el mismo clúster. También se añadieron nuevos tipos de plagio con el fin de aumentar el realismo del corpus. La novedad en 2011 fue la inclusión de casos basados en paráfrasis, porque en la mayoría aparece algún tipo de modificación en la estructura o en los elementos que forman las frases, y sobre todo cuando el plagio es translingüe. En Potthast et al. (2011a) se resume el contenido del corpus PAN-PC-11, con 26.939 documentos y un total de 61.064 casos de plagio, de los que cerca del 70% presenta paráfrasis.

Muchas de las investigaciones desde entonces han utilizado la versión PAN-PC-10 del corpus para la parte de desarrollo de su

<sup>1</sup> <https://www.urbund.com/es/>

<sup>2</sup> <http://www.turnitin.com/>

<sup>3</sup> <https://es.unplag.com/>

<sup>4</sup> <https://plagiarismcheck.org>

aplicación y PAN-PC-11 para comprobar los resultados, como es el caso de este trabajo.

En la actualidad para temas como la ciberseguridad, análisis de redes, forense y otras, además de la detección de plagio también se investiga en la detección de autoría de un texto (Kestemont et al., 2018), aunque queda fuera del objeto de este trabajo.

### 2.3 Técnicas de detección automática de plagio translingüe

El modelo *Cross-language character n-gram* (CL-CNG) utiliza n-gramas a nivel de caracteres para dividir los documentos en fragmentos comparables. Una vez normalizados los términos, es habitual representar el documento como un vector con todas las posibilidades de aparición conjunta de los elementos del alfabeto (McNamee et al., 2004).

El modelo *Cross-language alignment-based similarity* (CL-ASA) utiliza un corpus paralelo de textos (Barrón-Cedeño, 2012). Cada idioma tiene un factor de longitud diferente, determinado por la media y la desviación estándar de la longitud de los caracteres de una traducción de un idioma a otro (Tabla 1).

Parameter	en-de	en-es	en-fr	en-nl	en-pl
$\mu$	1.089	1.138	1.093	1.143	1.216
$\sigma$	0.268	0.631	0.157	1.885	6.399

Tabla 1 - Factor de longitud

El elemento fundamental de este modelo es el diccionario estadístico para la traducción, que incluye la probabilidad de traducción de una palabra de un idioma a otro para calcular la similitud entre dos documentos, mostrando un elevado rendimiento a bajo coste computacional (Franco-Salvador et al., 2012).

En Franco-Salvador et al. (2013) se describe la aproximación *Cross-language knowledge graphs analysis* (CL-KGA), utilizando grafos de conocimiento, donde los nodos son conceptos del documento, las aristas unen dos conceptos relacionados, y el peso de la arista muestra la importancia de la relación. Una vez obtenidos los grafos de los documentos, sospechoso y originales, se aplica una medida de la similitud entre ellos.

La técnica *Plagiarism detection using linguistic knowledge* (PLDK) combina la información sintáctica con la semántica entre los conceptos del documento (Abdi et al., 2015). Calcula la similitud a partir de un vector

de similitud en el orden de las palabras y otro de similitud semántica mediante consultas a *WordNet*, buscando el ancestro común y su distancia, que servirá de índice de semejanza.

En Franco-Salvador et al. (2016a) se propone una aproximación híbrida: utilizar grafos de conocimiento para calcular la similitud semántica por medio de recursos de redes semánticas como *WordNet* o *BabelNet*, y un modelo basado en el espacio vectorial para capturar los aspectos sintácticos que las redes semánticas no son capaces de detectar.

En los últimos años se ha producido un giro hacia las técnicas basadas en *Deep Learning*, frecuentemente utilizando una representación de palabras basada en *word embeddings*, para aportar conocimiento semántico (Suleiman et al. 2017), (Gupta, 2017).

## 3 Propuesta

El objetivo de este trabajo es encontrar una solución al problema de detección de plagio translingüe, partiendo de la técnica CL-KGA, utilizando herramientas disponibles de *Linked Data*, porque (1) la utilización de grafos de conocimiento permite abstraer tanto el problema del idioma a la hora de calcular la semejanza de dos documentos, como del orden de aparición de las palabras en las oraciones (clave para resolver paráfrasis) y (2) pueden utilizarse recursos semánticos para obtener otra información semántica del contenido.

Como se detalla en los apartados siguientes, esta propuesta exige: (1) La selección del corpus, PAN-PC-10, para entrenar el modelo en busca de los mejores parámetros posibles, y PAN-PC-11 para la fase de prueba, (2) el procesado lingüístico de los documentos y (3) la selección de las métricas de similitud entre los grafos de conocimiento y entre los documentos.

### 3.1 Procesamiento lingüístico

Las tareas de procesamiento de lenguaje necesarias se realizan con *Freeling*<sup>5</sup> y son: detección de idioma, tokenización del documento, división del documento en oraciones, análisis y etiquetación morfológica, análisis sintáctico de la oración y lematización.

Se selecciona *Freeling* por las funcionalidades que incluye en varios idiomas,

<sup>5</sup> <http://nlp.lsi.upc.edu/freeling/node/1>

sus diccionarios morfológicos de calidad, y porque puede integrarse con otras aplicaciones.

Como recurso semántico se ha escogido *BabelNet*<sup>6</sup> porque es un diccionario enciclopédico multilingüe con más de 14 millones de entradas que conecta entidades y conceptos en 271 idiomas (v. 3.7). Cada entrada o *synset*, representa un sentido de un concepto y contiene sus sinónimos.

Para seleccionar los fragmentos de los documentos, siguiendo a Franco-Salvador et al. (2016a), por defecto son de 5 sentencias consecutivas y con un salto posterior de 2 sentencias. Se preprocesan las palabras del fragmento para el formato (palabra\_lemma, etiqueta) de consulta en *BabelNet*, y la información extraída se organiza en grafos.

Para conseguir todos los sentidos de cada uno de los fragmentos de texto, que son los nodos iniciales del grafo, se piden los *synsets* vecinos y se estudia su coincidencia con algún *synset* ya almacenado como nodo, con la excepción de aquellos que pertenecen a la misma palabra. En caso de encontrar otro *synset* existente en el grafo se añade la relación entre los dos nodos. Este proceso es recursivo hasta un número máximo de saltos (parámetro).

### 3.2 Similitud entre grafos de conocimiento

El cálculo de la similitud entre dos grafos,  $G$  y  $G'$ , en este trabajo se basa en los nodos comunes y en el número de aristas que los une. Se define por  $S_c(G, G')$ , basada en el coeficiente de Dice entre los pesos de los nodos.

$$S_c(G, G') = \frac{2 \cdot \sum_{c \in V(G) \cap V(G')} w(c)}{\sum_{c \in V(G)} w(c) + \sum_{c \in V(G')} w(c)},$$

Donde  $V(G)$  y  $V(G')$  son los dos grafos por comparar y  $w(c)$  es el número de aristas incidentes en el vértice. El valor de semejanza entre dos fragmentos se encuentra en  $[0, 1]$ . En caso de que los dos grafos contengan sentidos y conceptos similares, el grafo de intersección entre los dos contendrá un alto número de nodos respecto al grafo unión. Por lo tanto, la medida se alejará de 0 y estará más cercana a 1. Por el contrario, si el grafo de intersección está vacío, será 0. Cada vez que la comparación de dos fragmentos supere un valor umbral (que se define de forma empírica), se marca como un

resultado positivo (parcial) de ser plagio. En caso de obtenerse un resultado positivo en tres fragmentos consecutivos del documento, el algoritmo confirma el plagio.

### 3.3 Métrica de evaluación

La métrica para evaluar el resultado de la detección de casos de plagio es el *F1-score* en este trabajo, aunque en las competiciones PAN se utiliza *plagdet*, que también se basa en la *precisión*, *exhaustividad* y, además, en la granularidad de casos positivos.

$$plagdet(S, R) = \frac{F_\alpha}{\log_2(1 + gran(S, R))},$$

Donde  $S$  es el conjunto de casos reales de plagio,  $R$  es el conjunto de detecciones,  $gran(S, R)$  es un valor de granularidad (el número de casos positivos de un mismo plagio) y  $F_\alpha$  es el valor *F1-score*. *F1-score* se diferencia de *pladget* en la granularidad, para evitar redundancia. Como en este trabajo solo se tiene en cuenta un caso positivo, *plagdet* se convierte en el valor de *F1-score*.

### 3.4 Diferencias

Hay cuatro diferencias fundamentales entre la aproximación descrita en Franco-Salvador et al. (2016a y 2016b) y la presentada: (1) Respecto a la métrica de similitud entre grafos, en este trabajo se utiliza el grado incidente en los vértices y en el suyo una interpolación entre aristas y vértices; (2) Para la generación de los grafos de conocimiento, se exige un máx. de tres saltos y ellos reinician el número de saltos al encontrar un vértice; (3) Para determinar plagio, son suficientes 3 grafos consecutivos positivos y ellos, para cada fragmento eligen los 5 más parecidos y aplican una función de convergencia uniendo fragmentos hasta superar un límite; (4) La hipótesis de trabajo de utilizar granularidad uno, hace que la métrica *plagdet* se convierta en *F1-score*.

## 4 Prototipo experimental

Descargados los dos corpus PAN<sup>7</sup>, el prototipo se encarga de crear los fragmentos de texto, el procesado lingüístico con *Freeling*, asociar a cada fragmento un grafo de conocimiento utilizando *BabelNet*, la serialización a disco y la recuperación de cada documento con sus fragmentos y grafos de conocimiento asociados,

<sup>6</sup> <https://babelnet.org/>

<sup>7</sup> <http://pan.webis.de/>

calcular la similitud entre grafos y de la decisión de plagio cuando en la comparación de fragmentos haya tres resultados positivos.

#### 4.1 Otros detalles técnicos

Se ha elegido la librería de grafos *GraphStream*<sup>8</sup>, por su facilidad de uso, la disponibilidad de los algoritmos más habituales, la capacidad dinámica para añadir tanto nodos como relaciones, su visualización de los grafos y la capacidad de serialización de la librería.

Debido al alto consumo de memoria al generar los grafos, el preproceso con *Freeling* se ha desplegado en un Ubuntu 14.04 LTS, con un wrapper Java alrededor de la aplicación C++ y con JAX-RS se ha obtenido un endpoint de consulta (Apache 7.0)<sup>9</sup>. Para una máquina diferente, modificar la *url* y *localhost* por su *ip*.

El prototipo se encuentra disponible en <https://github.com/Hisarlik/CrossLanguagePlagiarism/>.

#### 4.2 Dificultades integrando recursos semánticos

Las principales dificultades relacionadas con el uso de recursos en abierto han sido:

**Instalación y utilización de *Freeling***, que es laboriosa por la gran cantidad de librerías de C a compilar y las variables de entorno y sistema necesarias (cada distribución Linux tiene diferentes configuraciones de carpetas).

**Configuración librerías Java**, para acceder a las funcionalidades de *Freeling*, en cuyo repositorio de github hay código para un API Java. Para este trabajo se desarrolló una aplicación API Rest compleja, para (por ejm.) referenciar las funcionalidades por separado.

Modificación de los **Timeouts en la API Rest** desarrollada para el procesado de los documentos especialmente largos.

**Instalación de *BabelNet***. La versión 4.0 dispone de una API para unas pocas peticiones. Gracias a la política para investigación del startup babelscape.com, se descargó en local la base de datos (16GB archivos Lucene).

**Uso de memoria de los grafos**. Además del tiempo de computación, como casi todos los grafos pasan de mil nodos y el prototipo está en un Mac de 16GB de RAM, de los cuales 4GB

dedican a *Freeling*, si se guardaban los grafos en memoria, se quedaba sin ella, y se decidió serializar con *GraphStream* (interfaz en Java, de *Serializable* para las clases).

**Lógica recursiva de búsqueda de los nodos del grafo**. La mayor dificultad debida a la capacidad del equipo personal usado ha sido implementar la búsqueda de relaciones entre los nodos iniciales del grafo y su expansión con *synsets* intermedios entre ellos. En un equipo como el descrito, el tiempo de creación de un grafo suele ser de 5m y procesar un documento 100m, así que en un tiempo razonable solo era asumible probar en un subconjunto del corpus.

### 5 Pruebas y evaluación

Una vez desarrollado el prototipo, hay que configurar los parámetros del algoritmo para documentos en español e inglés. Estos son: (P1) Profundidad de peticiones a *BabelNet* o número de saltos permitidos para encontrar una relación entre dos conceptos; (P2) Umbral de similitud entre fragmentos y (P3) Número de fragmentos similares consecutivos para identificar plagio.

#### 5.1 Pruebas con PAN-PC-10 y 11

La primera prueba sobre un subconjunto monolingüe (inglés) con 96 documentos sospechosos, 35 originales y 35 casos reales de plagio obtuvo el mejor resultado con la configuración: (P1) 2, (P2) 0.35 y (P3) 3. Se detectan 20 posibles casos de plagio, de los cuales 9 falsos positivos, con 24 casos no detectados. El algoritmo alcanzó Precisión 0.55; Exhaustividad (o *recall*) 0.31 y *F1-score* 0.4.

Los mejores resultados de todas las pruebas realizadas (Tabla 2), muestran que la ampliación del número de grafos consecutivos para decidir plagio no mejora la detección, y con valores de similitud entre grafos mayor que 0.3, aumentan los falsos positivos.

Configuración	Casos Reales Detect	Detecciones Total	Casos Reales Total	Precisión	Recall	F1-score
sim > 0.3 y 3 frag. consecuti.	15	78	35	0.19	0.43	0.27
sim > 0.4 y 3 frag. consecuti.	8	9	35	0.89	0.23	0.36
sim > 0.3 y 4 frag. consecuti.	9	30	35	0.3	0.26	0.28
sim > 0.35 y 4 frag. cons.	9	12	35	0.75	0.26	0.38
sim > 0.35 y 3 frag. cons.	11	20	35	0.55	0.31	0.40

Tabla 2 – Resultado: solo inglés

<sup>8</sup> <http://graphstream-project.org/>

<sup>9</sup>

[https://github.com/Hisarlik/CrossLanguagePlagiarism\\_API\\_Freeling](https://github.com/Hisarlik/CrossLanguagePlagiarism_API_Freeling)

Las siguientes pruebas fueron para detección multilingüe, entre textos en español e inglés. El subconjunto del corpus tiene 35 documentos sospechosos en inglés, 40 originales en español y 47 casos reales de plagio.

El mejor resultado se ha obtenido con la configuración: (P1) 2, (P2) 0.32 y (P3) 3. Se han detectado 32 posibles casos de plagio de los cuales 28 lo eran y 4 no y los otros 19 casos posibles no han sido detectados. El algoritmo alcanza una Precisión 0.875; Exhaustividad 0.595 y *F1-score* 0.708. La principal conclusión es que se identifican pocos falsos positivos.

Aparte de este resultado, se han realizado otras pruebas cuyos mejores resultados se muestran en la Tabla 3, concluyéndose que si el valor de similitud es 0.3 o mayor, el número de falsos positivos aumenta, pero los resultados generales mejoran.

Configuración	Precisión	Recall	F1-score
similitud > 0.3 y 3 grafos consecutivos	0.53	0.49	0.58
similitud > 0.35 y 3 grafos consecutivos	0.96	0.66	0.65
similitud > 0.25 y 4 grafos consecutivos	0.12	0.68	0.20
similitud 0.32 y 3 grafos consecutivos	0.87	0.59	0.70

Tabla 3 – Resultado español - inglés

Una vez encontrados los parámetros óptimos, estos se han utilizado para procesar el corpus PAN-PC-11 (34 documentos sospechosos en inglés, 33 originales en español y 40 casos reales). De los 40 casos reales de plagio, en la mejor prueba se detectan 16, con 7 falsos positivos, alcanzando una Precisión 0.70, *Recall* 0.40 y *F1-score* 0.51, resultados inferiores a las pruebas con PAN-PC-10, mostrando la dificultad del plagio translingüe con paráfrasis.

## 5.2 Comparación con otros trabajos

Antes de comparar los resultados con los de las dos competiciones PAN, es necesario recordar que, en este trabajo (1) no se ha utilizado la totalidad del corpus, solo aquellos casos en que los documentos originales están en español y los sospechosos en inglés y (2) se utiliza el *F1-score* porque el algoritmo propuesto es de granularidad 1, recordando que algunos de los trabajos presentados también utilizan esta granularidad.

Los mejores resultados de los 11 trabajos en PAN 2010, publicados en<sup>10</sup>, son: *pladget* 0,79; 0,70; 0,69. El mejor en este trabajo es *F1-score* 0,70. Los mejores resultados oficiales de los 9 de PAN 2011, publicados en<sup>11</sup>, son: *pladget* 0,55; 0,41; 0,34. Comparando con el enfoque CL-KGA (DCW) con *pladget* 0.65 y CL-KGA (WSD concepts y/o weighting) con *pladget* 0.64, los resultados obtenidos son inferiores a estos últimos. El mejor en este trabajo es *F1-score* 0,51, similar a los mejores obtenidos en la competición.

## 6 Conclusiones

De acuerdo con los resultados obtenidos, se puede afirmar que la selección de un modelo basado en grafos de conocimiento y utilizando recursos semánticos (en abierto) para la detección de posibles casos de plagio translingüe (español e inglés), presenta resultados comparables con los obtenidos en las competiciones PAN 2010 y 2011.

Otro objetivo del trabajo era justificar la selección de recursos en abierto y cómo resolver las dificultades. Los resultados muestran la adecuación de los recursos para la gestión de los grafos, el procesado lingüístico y la aproximación seguida para definir los parámetros óptimos del algoritmo.

Sin embargo, es una solución costosa e intensiva en tiempo, memoria y CPU en un ordenador personal como el utilizado, por lo que en las pruebas solo se trabajó con documentos en inglés y español. En estas condiciones, se podría mejorar el rendimiento, ya sea paralelizando la creación de los grafos o guardando temporalmente los nodos que aparecen en la mayoría de los grafos.

Sería interesante estudiar modificaciones en el cálculo de la similitud entre los grafos, otras formas de ponderar la importancia de los conceptos para aumentar la calidad de los resultados o probar con otras aproximaciones para el problema planteado.

## Agradecimientos

Este trabajo ha sido parcialmente financiado por los proyectos Musaces (S2015/HUM3494) y VEMODALEN (TIN2015-71785-R).

<sup>10</sup> <https://pan.webis.de/clef10/pan10-web/plagiarism-detection.html>

<sup>11</sup> <https://pan.webis.de/clef11/pan11-web/plagiarism-detection.html>

**Bibliografía**

- Abdi, A., N. Idris, R. Aliguliyev y R. M. Aliguliyev. 2015. PDLK: Plagiarism detection using linguistic knowledge. *Expert Systems with Applications*, 42(22): 8936-8946.
- Barrón-Cedeño, A. 2012. On the Mono and Cross-Language Detection of Text Re-Use and Plagiarism. *Ph.D. thesis*, DSIC, UPV.
- Barrón-Cedeño, A., M. Vila y P. Rosso. 2013. Plagiarism meets Paraphrasing: Insights for the Next Generation in Autom. Plagiarism Detection. In: *Computational Linguistics*, 39(4) 917-947.
- Comas R. y J. Sureda. 2008. Academic cyberplagiarism: tracing the causes to reach solutions. *The Humanities in the Digital Era*, 10:1-7.
- Eaton S., M. Guglielmin y B. Otoo. 2017. PLAGIARISM: Moving from punitive to proactive approaches. In *Selected Proc. of the IDEAS Conference: Leading Educational Change*, páginas 28-36.
- Franco-Salvador, M., P. Gupta y P. Rosso. 2012. Detección de plagio translingüe utilizando el diccionario estadístico de BabelNet. *Computación y Sistemas*, 16(4): 383-390.
- Franco Salvador, M., P. Gupta y P. Rosso. 2013. Cross-language plagiarism detection using multilingual semantic network. *Proc. ECIR Springer*, páginas 710-713.
- Franco-Salvador M., P. Gupta, P. Rosso y E. Banchs. 2016a. Cross-language plagiarism detection over continuous space and knowledge graph-based representations of language. *Knowledge-based systems* 111, páginas 87-99.
- Franco-Salvador M., P. Rosso y M. Montes 2016b. A Systematic Study of Knowledge Graph Analysis for Cross-language Plagiarism Detection. *Information Processing & Management*, 52(4): 550-570.
- Franco-Salvador M. 2017. A Cross-domain and Cross-language Knowledge-based Representation of Text and its Meaning. *Ph.D. thesis*, DSIC, UPV.
- Gupta P. 2017. Cross-View Embeddings For Information Retrieval. *Ph.D. thesis*, DSIC, UPV.
- Kestemont, M., M. Tschuggnall, E. Stamatatos, W. Daelemans, G. Specht, B. Stein y M. Potthast. 2018. Overview of the Author Identification Task at PAN-2018 Cross-domain Authorship Attribution and Style Change Detection. *Proc. CLEF*, CEUR 2125.
- Krizkova, S., H. Tomaskova y M. Gavalec. 2016. Preference comparison for plagiarism detection systems. *Fuzzy Systems (FUZZ-IEEE)*, páginas 1760-1767.
- Martin, B. 2004. Plagiarism: policy against cheating or policy for learning? Australia. <https://ro.uow.edu.au/artspapers/78/>.
- McNamee, P. y J. Mayfield. 2004. Character n-Gram Tokenization for European Language Text Retrieval. *Information retrieval*, 7(1-2): 73-97.
- Menta, A. 2018. Detección de plagio multilingüe mediante recursos semánticos. *Tesis de Máster*. ETSI Informática, UNED.
- Meyer, S., B. Stein y M. Kulig. 2007. Plagiarism Detection without Reference Collections. In *Advances in data analysis*, Springer, páginas 359-366.
- Nahas, M. 2017. Survey and Comparison between Plagiarism Detection Tools. *American J. of Data Mining and Knowledge Discovery*, 2(2): 50-53.
- Potthast M., A. Barrón-Cedeño, B. Stein y P. Rosso. 2010. An Evaluation Framework for Plagiarism Detection. In *proc. COLING-2010*, páginas 997 -1005
- Potthast M., A. Eiselt, A. Barrón-Cedeño, B. Stein y P. Rosso. 2011a. Overview of the 3rd International Competition on Plagiarism Detection. In: Petras V., Forner P., Clough P. (Eds.), *Notebook Papers of CLEF 2011 LABs and Workshops*. In CEUR workshop proceedings, Vol. 1177.
- Potthast M., A. Barrón-Cedeño, B. Stein y P. Rosso. 2011b. Cross-Language Plagiarism Detection. In: *Languages Resources and Evaluation*. Special Issue on Plagiarism and Authorship Analysis, 45(1): 45-62.
- Suleiman, D., A. Awajan y N. Al-Madi. 2017. Deep Learning Based Technique for Plagiarism Detection in Arabic Texts. In *New Trends in Computing Sciences (ICTCS) IEEE*, páginas 216-222.

# Análisis comparativo de las características computacionales en los sistemas modernos de análisis de sentimiento para el español

## *Comparative analysis of the computational characteristics in modern sentiment analysis systems for Spanish*

Edgar Casasola,<sup>1</sup> Alejandro Pimentel,<sup>2</sup> Gerardo Sierra,<sup>2</sup>  
Eugenio Martínez Cámara,<sup>3</sup> Gabriela Marín<sup>1</sup>

<sup>1</sup>Universidad de Costa Rica

<sup>2</sup>Universidad Nacional Autónoma de México

<sup>3</sup>Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI). Universidad de Granada, España.

edgar.casasola@ucr.ac.cr, {apimentela, gsierram}@iingen.unam.mx,  
emcamara@decsai.ugr.es, gabriela.marin@ucr.ac.cr

**Resumen:** Existen múltiples sistemas para análisis de sentimiento con diseños heterogéneos y niveles variados de desempeño. En este artículo se presenta un modelo de generación de especificaciones computacionales de los sistemas para identificación de polaridad tendientes a facilitar comparaciones más profundas y detalladas de las técnicas que se utilizan. Buscamos crear conciencia entre los investigadores de la información necesaria para la construcción de especificaciones que permitan replicar sistemas. A su vez, discutimos las dificultades que se tiene al evaluar y al hacer comparaciones congruentes entre sistemas e ir más allá del resultado que se puede obtener para una tarea específica sobre un conjunto de datos particular. Estamos convencidos de que una estructuración completa y clara de todos los procesos y de los ajustes a que son sometidos los trabajos presentados en las competencias es crucial para enriquecer el conocimiento del uso de estrategias hacia la mejora general de los sistemas.

**Palabras clave:** Análisis de sentimiento, especificación de sistemas, revisión de literatura, corpus en español

**Abstract:** There are multiple systems for sentiment analysis with heterogeneous designs and varying levels of performance. In this paper, we propose a model of computational specifications of polarity identification systems. The model makes easier the comparison of the different techniques used. We seek to create awareness among researchers of the necessary information for the elaboration of specifications that allow the replication of systems. Additionally, we discuss the difficulties of evaluating and conducting consistent comparisons among systems, and going beyond the result that can be obtained for a specific task on a particular data set. We are convinced that a complete and clear framework that encompasses all the modules of the systems that participate in competitions is crucial to enrich the knowledge for improving the state-of-the-art in sentiment analysis.

**Keywords:** Sentiment analysis, systems specification, literature review, corpus in Spanish

## 1 Introducción

En la Web 2.0 la comunidad de usuarios conectados a nivel mundial a través de plataformas como *YouTube*, *Facebook*, y *Twitter* están produciendo información que posibilita la creación de nuevos tipos de aplicaciones de inteligencia colectiva (O'Reilly y Bat-

telle, 2009). Una de estas aplicaciones es el análisis de sentimiento, y una tarea específica particular que ha recibido mucha atención ha sido la identificación automática de la polaridad de textos cortos publicados en estas redes sociales. Sin embargo, el desarrollo de estas tecnologías del lenguaje varía entre idio-

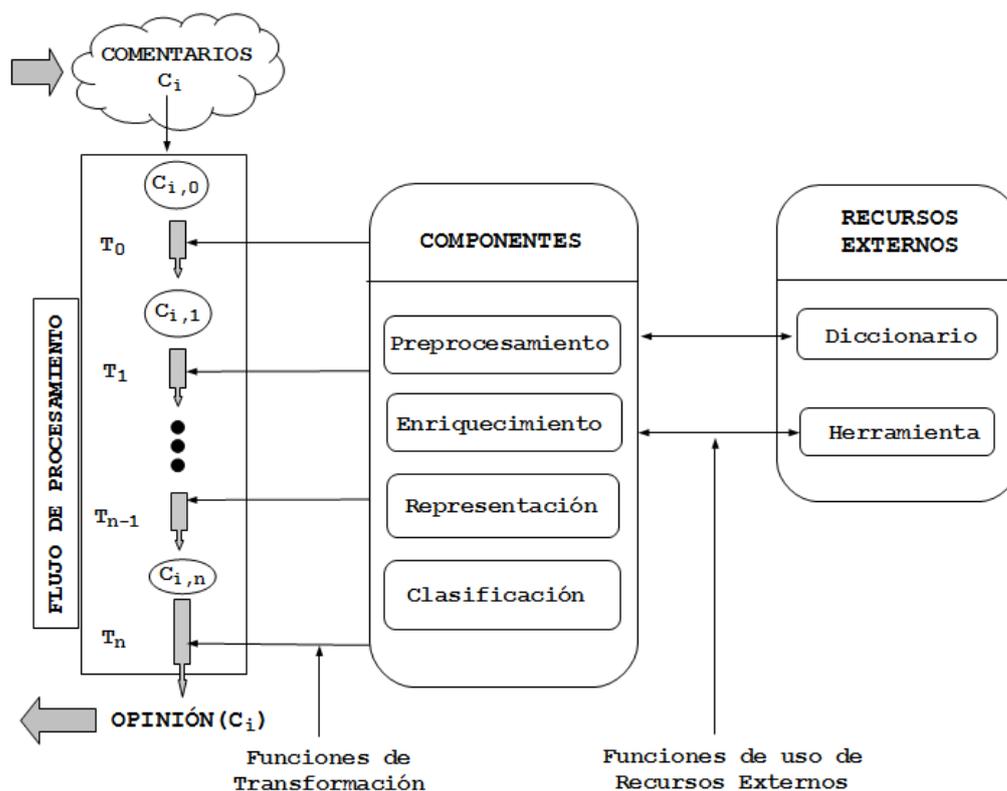


Figura 1: Diagrama general del modelo conceptual

mas y en el año 2012 se hizo evidente la existencia de una brecha entre el nivel alcanzado para idiomas como el inglés con respecto al español (Melero et al., 2012). Este trabajo se desarrolla orientado a este tipo de sistemas para el idioma español con el interés particular en disminuir esa brecha entre idiomas. En esa misma línea y con el fin de promover la interacción entre grupos de investigación de diferentes países de habla hispana se desarrolló un modelo base que proporciona una especificación conceptual uniforme para discutir y poder comparar las implementaciones de sistemas de análisis semántico. Las particularidades de este modelo serán presentadas en la sección 2.

En su primera versión, el modelo se enfocó en los sistemas para análisis de sentimiento y particularmente para los que llevan cabo identificación de polaridad a partir de comentarios de texto en idioma español. El modelo consiste de una abstracción de referencia centrada en los componentes generales de los sistemas, una descripción conceptual mediante el uso de una ontología en lenguaje OWL de los términos utilizados y la re-

lación entre ellos, e incluyen un conjunto de plantillas para la especificación de los componentes computacionales de los sistemas. En este trabajo mostramos el uso de esas plantillas para llevar a cabo una revisión de los sistemas presentados en el taller TASS 2017 (Díaz-Galiano et al., 2018) de la Sociedad Española para Procesamiento de Lenguaje Natural. TASS (Taller de Análisis de Semántico en la SEPLN) es un taller que se lleva a cabo desde el año 2012 con la finalidad de fomentar el desarrollo de técnicas para el tratamiento de opinión en español. Por lo tanto, se utilizaron las plantillas del modelo común para describir las especificaciones de cada uno de los sistemas.

## 2 Modelo base para la especificación de sistemas los sistemas

Para efectos de análisis se utiliza como base el modelo SAM que se muestra en la Figura 1 (Casasola-Murillo, 2018) este modelo conceptual permite especificar en términos computacionales las características implementadas en los sistemas para análisis de sentimiento.

En la Figura 1 se muestra la relación existente entre los elementos del modelo. En el lado izquierdo del diagrama se puede apreciar el *flujo de procesamiento* típico de aplicaciones de procesamiento de lenguaje natural. Sin embargo, el modelo incluye la especificación de los componentes computacionales involucrados en este procesamiento. En la parte central del diagrama se pueden observar los diferentes componentes del modelo según su rol dentro del sistema. Se incluyen: los componentes de *preprocesamiento del texto*, los componentes utilizados para *enriquecimiento* que agregan información extra al texto de cada comentario analizado, los componentes de *representación* que modifican la estructura de representación del comentario con fines computacionales, y finalmente los componentes utilizados para llevar a cabo la clasificación de los comentarios. Adicionalmente, a la derecha del diagrama, se muestran los *recursos externos*, tales como herramientas computacionales y diccionarios o lexicones. Este trabajo se centra en la especificación de estos componentes de preprocesamiento, enriquecimiento de texto, representación, clasificación y los recursos externos reportados en cada artículo de investigación analizado. Para el análisis de los componentes de los sistemas se utilizaron las plantillas para especificación de sistemas del modelo SAM. Uno de los aportes más relevantes del trabajo de Casasola-Murillo (2018) fue la construcción de una plantilla taxonómica para especificación de sistemas de análisis de sentimiento. El modelo **SAM** permite construir una especificación utilizando un grupo de plantillas para los diferentes componentes según su rol. Como parte del modelo se incluye un documento con la especificación detallada del modelo y está acompañado con una ontología en formato OWL con la descripción y organización jerárquica de los conceptos.

### 3 Metodología

En la edición de TASS 2017 se propusieron dos tareas: la clasificación de polaridad a nivel de documento; y la clasificación de polaridad a nivel de aspecto. En la primera tarea participaron 10 equipos con el corpus InterTASS (Díaz-Galiano et al., 2018). Esa tarea corresponde a la clasificación de polaridad en un *tweet*. Para dicha tarea se ofreció el recurso del InterTASS, un corpus de más de 2,000 *tweets* anotados con cuatro categorías

de opinión: *positive*, *neutral*, *negative* y *none*. Este trabajo se limita al análisis de los sistemas que reportaron resultados para esta tarea particular.

ID	Sistema	M-F1	Acc.
2	ELiRF-UPV	0.493	0.607
9	RETUYT	0.471	0.596
7	ITAINNOVA	0.461	0.576
6	Jacerong	0.460	0.602
1	INGEOTEC	0.526	0.595
3	Tecnolengua	0.456	0.582
10	SINAI	0.442	0.575
5	LexFAR	0.432	0.541
4	OEG	0.367	0.386
8	GSI	0.327	0.558

Tabla 1: Resultados obtenidos por cada sistema en TASS 2017

En la Tabla 1 se presentan los 10 sistemas con los resultados obtenidos. De igual manera se presenta un identificador numérico que será utilizado como índice de referencia para cada trabajo en lo que resta del artículo. Si bien hubo sistemas que obtuvieron mejores resultados al ser evaluados con otro corpus conocido como el corpus general *General Corpus*, en la tabla comparativa se eligieron los sistemas que presentaron evaluaciones utilizando el corpus InterTASS ya que para este corpus hubo una mayor participación.

La especificación del modelo y las plantillas fueron entregadas a un grupo de evaluadores previo a la especificación de artículos. Posteriormente, cada artículo fue revisado por al menos tres evaluadores, quienes revisaron un máximo de dos artículos cada uno. El perfil de estos evaluadores era el de estudiantes de Maestría en Computación y Lingüística con conocimiento de Procesamiento de Lenguaje Natural. Cada evaluador generó una especificación del sistema descrito utilizando las plantillas suministradas. Las plantillas fueron llenadas en forma individual y posteriormente se unificaron por consenso entre los tres evaluadores y el investigador a cargo. Posteriormente el proceso de especificación para todos los trabajos fue replicado por un investigador internacional y mediante la comparación de ambos resultados se obtuvieron los resultados que aquí se reportan. Cabe mencionar que todas las especificaciones fueron generadas a partir de los textos de los artículos publicados.

Componentes de preprocesamiento	Artículos									
	1	2	3	4	5	6	7	8	9	10
acentos	✓				✓	✓				
recodificación						✓				
emoticons	✓		✓	✓		✓				
hashtag		✓	✓			✓			✓	
mayúsculas	✓		✓		✓		✓		✓	✓
mención	✓	✓	✓	✓	✓	✓	✓		✓	
negación	✓		✓	✓	✓	✓			✓	
numeral	✓	✓		✓						
diccionario			✓	✓	✓	✓			✓	
puntuación		✓	✓	✓	✓	✓				
repetición	✓			✓	✓	✓			✓	✓
stopwords	✓			✓	✓		✓		✓	
url	✓	✓	✓	✓	✓	✓	✓		✓	
fechas		✓								

Tabla 2: Componentes de preprocesamiento reportados en cada trabajo

#### 4 Descripción de los componentes de cada sistema

Para cada grupo de componentes se presentan los resultados comparativos según la organización propuesta en el modelo.

##### 4.1 Pre-procesamiento

En la Tabla 2 se muestran las diferentes prácticas que se utilizaron en los sistemas que participaron en el TASS 2017.

El preprocesamiento se refiere a las transformaciones que se efectúan sobre los datos crudos con el objetivo de obtener algo uniforme y sin elementos que se puedan considerar ruidosos para el mejor manejo del texto.

Como ejemplos podemos observar la eliminación de puntuación, caracteres y ciertas palabras que se consideran sin contenido ni beneficio en su consideración; la normalización de diferentes formas de escribir expresiones informales, la eliminación de números y ligas, entre otras.

La etapa de preparar los datos para un procesamiento más sencillo y efectivo no solo se ha vuelto una práctica básica, sino que ha llegado a tener un carácter tan cotidiano que muchas veces no se le da la importancia y especificación suficientes y necesarios para llevar a cabo experimentación empírica. En la Tabla 2 se hace evidente cómo se le da un peso considerable a elementos propios del área de análisis de sentimiento y que se han reportado con buenos resultados como la consideración de la negación. O datos característicos del corpus (*tweets*) como las menciones o los *hashtags*. Sin embargo, la mayoría de las consideraciones que se hacen

para el pre-procesamiento se dejan sin detallar o simplemente no se mencionan aunque se hayan tomado en cuenta.

Cada *framework* tiene un conjunto de acciones de pre-procesamiento propias de las que se puede hacer uso. Esto ayuda a establecer una línea de comparación entre aquellos sistemas que hagan uso de dichos *frameworks*. No obstante, muchos de los trabajos anexan capas extra de pre-procesamiento, no utilizan los *frameworks* o no reportan con exactitud los parámetros que utilizan.

Antes no se tenía una herramienta para capturar los diferentes módulos y procesos que se llevaban a cabo en cada una de las etapas de un sistema de SA. Gracias al modelo propuesto es posible dar una estructuración a este tipo de características, tanto como una herramienta para la normalización de la información reportada como para la concientización de los autores para la retroalimentación de la información y la eficiente réplica experimental.

##### 4.2 Enriquecimiento

El enriquecimiento se refiere al anexo de datos mediante diferentes técnicas que resultan en un aporte de información que se puede obtener de la misma información mediante sistemas que hacen uso de conocimiento de dominio. Por ejemplo, los etiquetadores sintácticos, lematizadores y las técnicas de ponderación de la importancia de las palabras como el *tf-idf* entre otros.

Con el modelo de comparación que estamos utilizando, podemos observar para este módulo ciertos elementos populares a través

Técnica de enriquecimiento	Artículos									
	1	2	3	4	5	6	7	8	9	10
categoría gramatical		✓	✓	✓	✓	✓				
parsing entidad		✓		✓						
frecuencia lema		✓		✓	✓		✓		✓	
stemming	✓			✓						
stemming'	✓									
ponderación	✓	✓				✓				✓
polaridad				✓	✓	✓	✓	✓	✓	
embedding									✓	
aprendizaje a distancia	✓	✓								

Tabla 3: Técnicas de enriquecimiento reportadas en cada trabajo

de los diferentes sistemas como es el caso del etiquetado de partes de la oración (POS), la lematización y el enriquecimiento mediante el uso de lexicones de polaridad. El aprendizaje a distancia o *distance supervision* (Mintz et al., 2009) se menciona como técnica de enriquecimiento de los conjuntos de datos de entrenamiento.

El modelo comparativo que estamos utilizando es particularmente valioso para notar técnicas compartidas entre diferentes sistemas. En la Tabla 3 se puede observar que el enriquecimiento mediante lexicones de polaridad es una técnica bastante popular, y un elemento utilizado por varios de los sistemas más sobresalientes como el 9, el 7 y el 6. No obstante, el sistema 2, que es el más sobresaliente, no los utiliza, e incluso en su trabajo se puede leer (Hurtado, Pla, y González, 2017):

“Tampoco se han utilizado lexicones de polaridad debido a que no mejoraban los resultados sobre validación...”

Misma conclusión a la que llegaron Rosá et al. (2017) del sistema 9, aunque ellos si lo utilizaron:

“En particular, se observa que el aporte del léxico subjetivo es poco relevante.”

Esto plantea nuevas implicaciones sobre la eficacia de las técnicas reportadas por trabajos sobresalientes del estado del arte. Los sistemas son una serie de herramientas en conjunto. La aportación que tiene cada una de ellas en el resultado que se obtiene es desconocido, e imposible de averiguar si no se tiene una estructuración de las diferentes opciones que se tienen para cada módulo.

En el estado del arte se reportaron buenos resultados con ayuda de los lexicones de polaridad (Moreno-Ortiz y Hernández, 2017). Pero esto no implicaba que fuera gracias a los lexicones de polaridad que se obtuvieran los resultados sobresalientes. No obstante, se siguen utilizando casi indiscriminadamente sin llegar a hacerse pruebas específicas de la ganancia que están produciendo. Es posible que haya técnicas junto con las que funcionan bien y otras con las que no. Es debido a consideraciones de esta naturaleza por lo que se vuelve evidente la necesidad de construir un mapa de eficacia para cada una de las herramientas que se pueden utilizar.

En contraposición con lo anterior, para el sistema 7, Montañés-Salas et al. (2017) defienden que sus resultados si se vieron beneficiados con la inclusión de información de polaridad:

“Como se puede observar, la utilización del diccionario afectivo hace que tengamos mejores resultados, aunque el algoritmo también obtiene resultados bastante satisfactorios sin ellos.”

En las dos secciones siguientes ahondaremos más acerca de los procesos que se usaron para incluir la información de polaridad y las posibles razones que pueden estar detrás de estas discrepancias en los resultados.

### 4.3 Representación

La representación se refiere a la configuración en la que se ingresarán los datos para su procesamiento y clasificación. Es decir, qué características en particular se toman en cuenta para definir los vectores de entrada de un sistema computacional.

Método de representación	Artículos									
	1	2	3	4	5	6	7	8	9	10
bolsa de palabras				✓	✓			✓	✓	
n-grama	✓					✓				
q-grama	✓					✓				
skip-grama	✓									
features		✓	✓	✓	✓	✓			✓	✓
embeddings		✓					✓		✓	✓

Tabla 4: Métodos de representación del texto utilizadas en cada trabajo

En la Tabla 4 se listan una serie de características que cada uno de los sistemas reporta como haber tomado en cuenta. Es posible

observar una tendencia hacia la agrupación en un vector de características (*features*).

Por otro lado, un detalle que resalta a la vista es que los tres sistemas más sobresalientes (2,7 y 9) emplean *word embeddings* para la representación de sus datos. Esto es de particular interés para la comparación de sistemas en general, pues en todos los módulos de estudio, no se encontró otra característica que identificara tan bien a estos tres sistemas.

En adición a lo anterior, es importante resaltar dos experimentos similares que llevaron a cabo los sistemas 2 y 7 referente a la representación de sus palabras en combinación con información de polaridad (Hurtado, Pla, y González, 2017; Montañés-Salas et al., 2017).

Como se mencionó en la sección 4.2, el sistema 2 (Hurtado, Pla, y González, 2017) descartó el uso de diccionarios de polaridad debido a la ausencia de mejoras en el desempeño de dicho enriquecimiento. No obstante, hacen experimentos con *sentiment specific word embeddings* (SS-WE), para los cuales aplican *distant supervision* con heurísticas de emoticones. Esto podría verse como un uso indirecto de información de polaridad para la creación de *embeddings*.

Por su parte, para el sistema 7, Montañés-Salas et al. (2017) obtiene sus propios SS-WE mediante una técnica distinta. Se basa en el uso de varios diccionarios afectivos para introducir sinónimos basados en emociones y con esto reduce el vocabulario del que obtiene *embeddings* especiales.

De sus experimentos al utilizar *word embeddings* entrenados especialmente tomando en cuenta sentimientos, Hurtado, Pla, y González (2017) no obtiene mejoras mientras que Montañés-Salas et al. (2017) sí. Lo que supone que la información que se obtiene de diccionarios afectivos necesita consideraciones especiales para poder ser aprovechada. A continuación, la sección 4.4 presenta el análisis de los componentes de clasificación.

#### 4.4 Clasificación

La clasificación se refiere a los algoritmos que se utilizan para obtener el resultado final de la categoría a la que pertenece la polaridad de los comentarios. En la Tabla 5 se muestran los diferentes sistemas que reportan los participantes del TASS 2017

De manera similar a los componentes de representación, para este tipo de componen-

Estrategia de clasificación	Artículos									
	1	2	3	4	5	6	7	8	9	10
Bayes				✓						
Ensamble	✓	✓				✓		✓	✓	
Red neuronal		✓					✓	✓	✓	✓
Logística			✓			✓				
SVM	✓				✓	✓			✓	✓
Genético	✓									

Tabla 5: Estrategias de clasificación reportadas en cada trabajos

tes destacan como característica común a los sistemas más sobresalientes (2, 9 y 7) el uso de redes neuronales.

Por otro lado, se puede notar que al igual que con los componentes de representación, se tiene la presencia del sistema 10, sin embargo, como se observa en las tablas, este sistema se encuentra pobremente especificado y no nos es posible analizar por qué, a pesar de tener estas cosas en común con los sistemas más sobresalientes, no tuvo un mejor desempeño.

Un elemento muy interesante dentro de este rubro que no puede ser capturado con el modelo actual es la presencia de indicaciones dentro de los artículos que demuestran una tendencia acerca del mejor funcionamiento de algunos sistemas sobre otros, como se mencionan Moreno-Ortiz y Hernández (2017):

“We also tried a SVM classifier on the same feature sets, but we consistently obtained poorer results compared to the logistic regression classifier.”

Lo anterior se puede traducir como:

“También tratamos con clasificadores SVM sobre el mismo conjunto de características, pero obtuvimos de manera consistente peores resultados comparados con los de un clasificador logístico.”

Este tipo de aportación puede resultar muy valioso para la réplica, la identificación de las características que pueden combinar bien con ciertos algoritmos de clasificación y el fomento en el cambio de estrategias. Lo anterior tiene el potencial de enriquecer los trabajos mediante una experimentación dirigida y, al mismo tiempo, facilitar la comparación entre sistemas inclusive si no comparten el resto de los módulos ni características.

## 4.5 Diccionarios

Recurso	Artículos									
	1	2	3	4	5	6	7	8	9	10
polarización	✓		✓	✓	✓	✓	✓	✓	✓	
emoticones				✓						
stoplist				✓			✓		✓	
lexicon				✓			✓			
embedding									✓	✓

Tabla 6: Recursos estáticos reportados como herramientas en cada trabajo

Los diccionarios se refieren a aquellas herramientas estáticas de las que hacen uso los participantes para añadir información a sus sistemas. Esto incluye lexicones de polaridad, diccionarios de emoticones, listas de palabras de paro, embeddings pre-entrenados entre otros. Todas aquellas herramientas que están listas para usarse como consulta.

En la Tabla 6 se muestra el uso que tuvieron los participantes del TASS 2017 de distintas herramientas estáticas. Como se puede observar, la mayoría de los sistemas utilizó lexicones de polarización. Como ya se había mencionado en la sección 4.2, Hurtado, Pla, y González (2017) y Rosá et al. (2017) llegan a la conclusión de que esta estrategia en particular no logra un mejor rendimiento en los sistemas. Mientras que Montañés-Salas et al. (2017) llega a la conclusión contraria.

En la sección 4.3 discutimos un cambio en la manera en la que se incorporó la información de polaridad entre los sistemas 2 y 7 como una posible explicación para la diferencia de los resultados. No obstante, no podemos descartar la posibilidad de que la diferencia, o parte de ella, provenga del recurso utilizado. El sistema 2 no utiliza diccionarios afectivos, pero el sistema 9 sí a pesar de que también reporta que no obtuvo mejoras con ello (Rosá et al., 2017).

Podemos notar una interesante diferencia entre los sistemas 7 y 9 en cuanto a los recursos de polaridad que utilizan. Rosá et al. (2017) (sistema 9) utilizan tres léxicos subjetivos disponibles para el español (Cruz et al., 2014; Saralegi y San Vicente, 2013; Brooke, Tofiloski, y Taboada, 2009). Mientras que Montañés-Salas et al. (2017) (sistema 7) utilizan el *Dictionary of Affect in Language* (Cynthia M. Whissell et al., 1986) y el *Affective Norms for English Words* (ANEW; Bradley et al., 1999). A pesar de que Montañés-Salas et al. (2017) utilizan diccionarios

en inglés, obtienen mejoras en sus resultados mientras que Rosá et al. (2017) no las obtienen con diccionarios en español.

Como mencionamos en la sección 4.3, la diferencia en el rendimiento puede venir de la representación. Esto supone un espacio para experimentar, la representación por sinónimos y *embeddings* con diccionarios en español, o bien, la anexión de información de polaridad mediante el uso de diccionarios en inglés. Esto es tan solo un ejemplo de los nichos que se pueden localizar mediante la estructuración de las características de los sistemas con ayuda de modelos como los que estamos utilizando. Finalmente, el último tipo de componentes por mencionar son los tipos de herramientas, y herramientas específicas que son utilizadas en cada uno de los trabajos.

## 4.6 Herramientas

Las herramientas se refieren a la paquetería computacional de la que se apoyan los participantes para llevar a cabo cada uno de los módulos antes mencionados, si es el caso.

Herramienta	Artículos									
	1	2	3	4	5	6	7	8	9	10
POS tagger		✓		✓	✓	✓	✓			
clasificador	✓			✓	✓		✓		✓	
ortografía						✓				
tokenizador		✓		✓						
embeddings		✓							✓	
ponderación	✓									
sentimientos			✓							
framework		✓	✓	✓						

Tabla 7: Herramientas computacionales reportadas en cada trabajo

En la Tabla 7 se muestran las diferentes tareas para las que los sistemas que participaron en el TASS 2017 reportaron utilizar paquetería especializada. Este rubro es particularmente útil para la especificación y estandarización de las librerías y sus respectivos parámetros de operación. Al identificar de manera precisa los recursos que se utilizan en cada etapa del proceso, se facilita en gran medida la réplica, la experimentación, el reporte de resultados y la construcción de paquetes de software que coordinen todo de manera automática.

## 5 Conclusiones

Es claro que la alta heterogeneidad que se presenta en los sistemas hace que no se pueda establecer una relación causa efecto entre

los componentes de cada sistema y su efecto sobre el rendimiento.

Muchos de los procesos que se aplican en los diferentes módulos se hacen debido a resultados exitosos reportados en otros trabajos. No obstante no se tiene una verdadera perspectiva de comparación para identificar en qué medida las técnicas replicadas son las que tuvieron un aporte en los sistemas reportados en el estado del arte.

Son escasos los trabajos que cuentan con la experimentación para determinar la efectividad de las herramientas periféricas que utilizan. De allí la importancia de generar conciencia tendiente a producir una adecuada declaración explícita de cada etapa. Además, resaltar la importancia de generar especificaciones para capturar las pruebas satisfactorias e insatisfactorias que se reporten. Si un sistema similar logra mejoras al variar algún elemento particular esto puede servir de base para replicación y observación del efecto de las variantes.

Un modelo comparativo de esta naturaleza permite agilizar la experimentación tendiente a determinar el efecto de los componentes más polémicos y sus combinaciones. Lo anterior, para encontrar aquellos casos en que funcionen, la manera en la que funcionan, o bien, descartar prácticas que en realidad no están aportando un beneficio. De esta forma, se puede determinar el panorama general asociado al desarrollo los estos sistemas con el fin de lograr mejoras en su desempeño.

### ***Agradecimientos***

Se agradece el apoyo de MICITT y CONICIT del Gobierno de Costa Rica y al proyecto IN403016 DGAPA-PAPIIT de la UNAM. Eugenio Martínez Cámara fue financiado por el programa Juan de la Cierva Formación (FJCI-2016-28353) del Gobierno de España.

### ***Bibliografía***

Casasola-Murillo, E. 2018. *Desarrollo de un modelo computacional para la especificación de sistemas de análisis de sentimiento con comentarios de redes sociales en español*. Ph.D. tesis, Escuela de Ciencias de la Computación, U. de Costa Rica.

Díaz-Galiano, M. C., E. Martínez-Cámara, M. García-Cumbreras, M. García-Vega, y Villena-Román. 2018. The democratization of deep learning in tass 2017. *Procesamiento del Lenguaje Natural*, 60:37–44.

Hurtado, L.-F., F. Pla, y J.-Á. González. 2017. Elirf-upv en tass 2017: Análisis de sentimientos en twitter basado en aprendizaje profundo. En *Proceedings of TASS 2017: Workshop on Sentiment Analysis at SEPLN*, volumen 1896, páginas 29–34, Murcia, Spain, September. SEPLN, CEUR Workshop Proceedings. ISSN: 1613:0073.

Melero, M., A.-B. Cardús, A. Moreno, G. Rehm, K. de Smedt, y H. Uszkoreit. 2012. *The Spanish language in the digital age*. Springer.

Mintz, M., S. Bills, R. Snow, y D. Jurafsky. 2009. Distant supervision for relation extraction without labeled data. En *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, páginas 1003–1011. Association for Computational Linguistics.

Montañés-Salas, R.-M., d.-H.-M. Alonso-Rafael, J. Veá-Murguía, R. Aznar-Gimeno, y F. J. Lacueva-Pérez. 2017. Fasttext como alternativa a la utilización de deep learning en corpus pequeños. En *Proceedings of TASS 2017: Workshop on Sentiment Analysis at SEPLN*, volumen 1896, páginas 65–69, Murcia, Spain, September. SEPLN, CEUR Workshop Proceedings. ISSN: 1613:0073.

Moreno-Ortiz, A. y C. P. Hernández. 2017. Tecnolengua lingmotif at tass 2017: Spanish twitter dataset classification combining wide-coverage lexical resources and text features. En *Proceedings of TASS 2017: Workshop on Sentiment Analysis at SEPLN*, volumen 1896, páginas 35–42, Murcia, Spain, September. SEPLN, CEUR Workshop Proceedings. ISSN: 1613:0073.

O'Reilly, T. y J. Battelle. 2009. *Web squared: Web 2.0 five years on*. O'Reilly Media.

Rosá, A., L. Chiruzzo, M. Etcheverry, y S. Castro. 2017. Retuyt en tass 2017: Análisis de sentimientos de tweets en español utilizando svm y cnn. En *Proceedings of TASS 2017: Workshop on Sentiment Analysis at SEPLN*, volumen 1896, páginas 77–83, Murcia, Spain, September. SEPLN, CEUR Workshop Proceedings. ISSN: 1613:0073.

# TASS 2018: The Strength of Deep Learning in Language Understanding Tasks

## *TASS 2018: La Potencia del Aprendizaje Profundo en Tareas de Comprensión del Lenguaje*

Manuel Carlos Díaz-Galiano,<sup>1</sup> Miguel Á. García-Cumbreras,<sup>1</sup>  
Manuel García-Vega,<sup>1</sup> Yoan Gutiérrez,<sup>2</sup> Eugenio Martínez-Cámara,<sup>3</sup>  
Alejandro Piad-Morffis,<sup>4</sup> Julio Villena-Román<sup>5</sup>

<sup>1</sup>Centro de Estudios Avanzados en Tecnologías de la Información y de la Comunicación (CEATIC). Universidad de Jaén, España

<sup>2</sup>University of Alicante, España

<sup>3</sup>Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI). Universidad de Granada, España

<sup>4</sup>University of Havana, Cuba. <sup>5</sup>MeaningCloud, España.

<sup>1</sup>{mcdiaz,magc,mgarcia}@ujaen.es, <sup>2</sup>ygutierrez@dlsi.ua.es

<sup>3</sup>emcamara@decsai.ugr.es, <sup>4</sup>apiad@matcom.uh.cu

<sup>5</sup>jvillena@meaningcloud.com

**Abstract:** The edition of TASS in 2018 was the edition of the evolution of TASS to a competitive evaluation workshop on semantic and text understanding tasks. Consequently, TASS has enlarged the coverage of tasks, and it goes beyond sentiment analysis. Thereby, two new tasks focused on semantic relation extraction in the health domain and emotion classification in the news domain were added to the two traditional tasks of TASS, namely sentiment analysis at tweet level and aspect level. Several systems were submitted, and most of them are based on state of the art classification methods, which highlight those ones grounded in Deep Learning. As addition contribution, TASS 2018 released two new corpora, specifically the ones of the two new tasks.

**Keywords:** Sentiment analysis, emotion classification, digital health

**Resumen:** La edición de 2018 de TASS ha sido la de la evolución de TASS hacia un taller de evaluación competitiva sobre tareas de análisis semántico y de entendimiento del lenguaje, ampliando así su cobertura de tareas más allá del análisis de opiniones. De este modo, a las dos tareas clásicas de clasificación de la polaridad a nivel de tuit y a nivel de aspecto, se ha añadido una tarea de extracción de relaciones semánticas en el dominio médico, y otra de clasificación de emociones en el dominio periodístico. Son numerosos los sistemas que se evaluaron en TASS 2018, y hay que destacar que la mayoría de ellos están a la vanguardia en el uso de técnicas de clasificación, destacando los sistemas basados en Aprendizaje Profundo. Como contribución adicional, TASS 2018 ha publicado dos nuevos *corpora*, correspondientes a las dos nuevas tareas.

**Palabras clave:** Análisis de opiniones, análisis de emociones, salud digital

### 1 Introduction

The edition of TASS 2018 was the edition of the evolution of TASS into a competitive evaluation workshop on semantic analysis and text understanding tasks. Accordingly, TASS has enlarged the coverage of Natural Language Processing (NLP) tasks, and it goes beyond Sentiment Analysis.

In this paper, we describe the edition of TASS 2018, in which four tasks were orga-

nized. Two of them were the usual tasks of TASS, namely sentiment analysis at tweet level (TASK 1) and sentiment analysis at aspect level (TASK 2). Two new tasks were additionally organized, the first one focused on semantic relation extraction in health data (TASK 3), and the second one emotion classification in the the news domain (TASK 4).

Eighteen research teams participated in the four tasks of TASS 2018, and they sub-

	P	N	NEU	NONE	Total
ES	1,115	1,402	418	474	3,409
PE	756	820	594	758	2,928
CR	677	912	297	447	2,333

Table 1: InterTASS 2.0: tweets subsets

mitted state-of-the-art systems, which is a relevant contribution for the research community of NLP in Spanish. As additional contribution, two new corpora were released, namely the two ones of the two new tasks.

The paper is organized as follows: Section 2 describes the corpora provided in TASS 2018. The four organized tasks are detailed in Section 3. Section 4 exposes the main conclusions of TASS 2018 and the future work related to TASS.

## 2 Resources

TASS 2018 provided five datasets to the participants for the evaluation of their systems. Only two of them were already used in previous editions.

### 2.1 InterTASS 2.0 Corpora

In 2018, a new version of InterTASS Corpus arised, with new subsets of training and test data. *InterTASS 2.0* is composed of three subsets, with tweets written in different varieties of Spanish (for Spain, Peru and Costa Rica). It exhibits a large amount of lexical and even structural differences in each variant, and tweets were annotated with 4 different polarity labels POSITIVE, NEGATIVE, NEUTRAL and NONE. Table 1 shows the tweets distribution for each subset (Spain:ES, Peru:PE and Costa Rica:CR).

All subsets are balanced, although they have more positive and negative tweets.

### 2.2 Social-TV and STOMPOL Corpora

The Social-TV and STOMPOL corpora were released in previous TASS editions (Martínez-Cámara et al., 2017). They have tagged at aspect level, with 3 levels of opinion: positive (P), neutral (NEU) and negative (N). Table 2 shows the tweets distribution for these data in training and test sets.

### 2.3 eHealth-KD Corpora

For evaluation purposes of the eHealth-KD challenge, a corpus of health-related sen-

	training	Test	Total
Social-TV	1,773	1,000	2,773
STOMPOL	1,284	784	2,068

Table 2: Social-TV and STOMPOL Corpora

tences in Spanish was manually built and tagged. The corpus consists of a selection of articles collected from the MedlinePlus<sup>1</sup> website. MedlinePlus is the United States National Institutes of Health’s website. This platform freely provides large health textual data from which a selection was made for constituting the eHealth-KD corpus in Spanish language. Table 3 shows the distribution of this corpus.

Metric	Total	Trial	Training	Dev.	Test
<i>Files</i>	11	1	6	1	3
<i>Sentences</i>	1173	29	559	285	300
<i>Annotations</i>	13113	254	5976	3573	3310
<b>Entities</b>	7188	145	3280	1958	1805
- Concepts	5366	106	2431	1524	1305
- Actions	1822	39	849	434	500
<b>Roles</b>	3586	71	1684	843	988
- subject	1466	33	693	339	401
- target	2120	38	991	504	587
<b>Relations</b>	2339	38	1012	772	517
- is-a	1057	18	434	370	235
- part-of	393	3	149	145	96
- property-of	836	15	399	244	178
- same-as	53	2	30	13	8

Table 3: Size of the eHealth-KD v1.0 corpus

#### 2.3.1 SANSE corpus

The Spanish brANd Safe Emotion (SANSE) corpus comprises 15,152 news headlines from newspapers of some Spanish speaking countries: Spain, Argentina, Chile, Colombia, Cuba, USA, Mexico, Peru and Venezuela. The aim was to build a representative corpus of the use of Spanish in headlines.

The corpus was randomly splitted into two sets: the L1 subset with 2,000 headlines, and the L2 subset with the rest 13,152 headlines.

L1 was manually annotated by two annotators, and a third annotator undid the tie in those cases with no agreement. A SAFE headline was defined as an utterance that arises a positive or neutral emotion and is not related to any controversial topic such as religion or

<sup>1</sup><https://medlineplus.gov/xml.html>

extreme wing political news.<sup>2</sup> Otherwise utterances were considered as UNSAFE. The agreement of the annotation was 0.58 according to  $\pi$  and  $\kappa$  (Cohen, 1960), which may be considered as moderate according to Landis and Koch (1977), though close to be substantial. This is justified considering the strong subjective nature of emotions.

Finally the L1 set was splitted into three subsets: *training* (1,250 headlines), *development* (250) and *test* (500). The L2 subset was automatically annotated by a voting system built upon the outputs of the systems submitted for S1 subtask (see Section 3.4.1).

### 3 Tasks

TASS 2018 organized four tasks, the usual tasks 1 (see Section 3.1) and 2 (see Section 3.2) about sentiment analysis at tweet and aspect levels, and two new tasks. TASK 3 (see Section 3.3) is focused on the extraction of semantic relations on health data. TASK 4 (see Section 3.4) proposes the emotional classification of news headlines in order to identify their level of safety for publishing spot ads.

#### 3.1 Task 1. Sentiment analysis at tweet level

This task was focused on the evaluation of polarity classification systems at tweet level of tweets written in Spanish.

The submitted systems had to classify short tweets written in an informal language, many of them with misspelling or emojis, even onomatopoeias. But this year, systems had to solve a new problem: Multilinguality. One of the corpus was expanded with tweets written in Spanish from Peru and Costa Rica.

This extended corpus was the International TASS Corpus (InterTASS), a corpus released in 2017 with text written in Spanish from Spain and its description can be found in (Martínez-Cámara et al., 2017), while the varieties from Peru and Costa Rica have been released this year and their descriptions are shown in (Martínez-Cámara et al., 2018).

Currently, it exhibits a large amount of lexical and even structural differences in each variant. The main purpose of compiling and using an inter-varietal corpus of Spanish for the evaluation tasks is to challenge participating systems to cope with the many faces of this language worldwide.

<sup>2</sup>These topics may arise strong conflicting emotions in some readers.

However, the General Corpus of TASS was provided in the same way as previous editions. Further details in Martínez-Cámara et al. (2017).

Datasets were annotated with 4 different polarity labels (POSITIVE, NEGATIVE, NEUTRAL and NONE), and systems had to identify the orientation of the opinion expressed in each tweet in any of those 4 polarity levels.

Four sub-tasks were proposed:

**Subtask-1:** Monolingual ES. *training* and *test* were the InterTASS ES datasets.

**Subtask-2:** Monolingual PE. *training* and *test* were the InterTASS PE datasets.

**Subtask-3:** Monolingual CR. *training* and *test* were the InterTASS CR datasets.

**Subtask-4:** Cross-lingual. The Spanish of *training* set had to be different from the evaluation one, in order to test the dependency of systems on a language.

Accuracy and the macro-averaged versions of Precision, Recall and F1 were used as evaluation measures. Systems were ranked by the Macro-F1 and Accuracy measures.

For TASK 1 five system were presented. Most of them make use of deep learning algorithms, combining different ways of obtaining the word embeddings: INGEOTEC, RETUYT-InCo (Chiruzzo and Rosá (2018)), ITAINNOVA (Montanés, Aznar, and del Hoyo (2018)), ELiRF-UPV. (González, Hurtado, and Pla (2018b)) and ATALAYA (Luque and Pérez (2018)).

The first three teams classified for each monolingual subtask are shown in Table 4.

For the cross-lingual runs, the participants selected an InterTASS dataset to train their systems and a different one to test the dependency of systems on a language. Table 4 shows the results of the first 3 teams classified in these cross-lingual subtasks.

To read a complete information about the systems, runs and results see (Martínez-Cámara et al., 2018).

#### 3.2 Task 2. Aspect-based Sentiment Analysis

Aspect-based Sentiment Analysis challenge was focused on aspect-based polarity classification systems. The datasets to evaluate the approaches were similar to previous editions (Martínez-Cámara et al., 2017): Social-TV and STOMPOL.

This year only one group has participated, ELiRF (González, Hurtado, and Pla, 2018b)

Run	Task	M. F1	Acc.
elirfEsRun1	monoES	0.503	0.612
retuytLstmEs1	monoES	0.499	0.549
atalayaUbav3	monoES	0.476	0.544
retuytLstmCr2	monoCR	0.504	0.537
elirfCrRun2	monoCR	0.482	0.561
atalayaCrLr502	monoCR	0.475	0.582
retuytCnnPe1	monoPE	0.472	0.494
atalayaPeLr502	monoPE	0.462	0.451
ingetecRun1	monoPE	0.439	0.447
retuytSvmEs2	multiES	0.471	0.555
ingetecRun1	multiES	0.445	0.530
atalayaMlp300	multiES	0.441	0.485
ingetecRun1	multiPE	0.447	0.506
retuytSvmPe2	multiPE	0.445	0.514
atalayaMlp300	multiPE	0.438	0.523
retuytSvmCr1	multiCR	0.476	0.569
ingetecRun2	multiCR	0.454	0.5382
itainnovaClBase	multiCR	0.409	0.440

Table 4: TASK 1: Best teams per task

that has submitted three experiments for each collection. They explored different approaches based on Deep Learning. Specifically, they studied the behaviour of the CNN, Attention Bidirectional Long Short Term Memory (Att-BLSTM) and Deep Averaging Networks (DAN), similar to the proposal of the team for TASK 1. In order to study the performance of the different models, they carried out an adjustment process. Table 5 show the results obtained in their experiments.

For evaluation, exact match with a single label combining “aspect-polarity” was used. Similarly to TASK 1, the macro-averaged version of Precision, Recall, F1, and Accuracy were considered, and Macro-F1 was used for a final ranking of proposed systems.

Run	Corpus	M. F1	Acc.
ELiRF-run1	Social-TV	0.485	0.627
ELiRF-run3	Social-TV	0.483	0.628
ELiRF-run2	Social-TV	0.476	0.625
ELiRF-run2	STOMPOL	0.526	0.633
ELiRF-run1	STOMPOL	0.490	0.613
ELiRF-run3	STOMPOL	0.447	0.576

Table 5: Macro f1 (M. F1) and accuracy (Acc.) in TASK 2 Social-TV corpus and STOMPOL corpus results

To read a complete information about

the systems, runs and results see (Martínez-Cámara et al., 2018).

### 3.3 Task 3. eHealth-KD

eHealth Knowledge Discovery (eHealth-KD) challenge proposed the identification of **Concepts** and **Actions**, for linking them later as a form of capturing the semantics of a broad range of health related text. Concepts are key-phrases that represent actors or entities relevant in a domain, while Actions represent how these Concepts interact with each other. For this challenge two types of relations: **Subject** and **Target**, were used to link Concepts and Actions (an special type of Concept), describing the main roles that a Concept can perform. In addition, other four specific semantic relations between Concepts were defined. A detailed description of the eHealth-KD challenge can be found in its web site<sup>3</sup> and in (Martínez-Cámara et al., 2018).

#### 3.3.1 Subtasks and evaluation scenarios

eHealth-KD proposed two evaluation strategies: at subtask level (i.e. subtasks A, B and C) and per scenario. The subtasks were: **subtask A** concerned with the extraction of the relevant key phrases; **subtask B** concerned with classifying the key phrases identified in subtask A as either **Concept** or **Action**; and **subtask C** concerned with discovering the semantic relations between pairs of entities.

The evaluation scenarios were: **Scenario 1** which involved subtasks A, B and C sequentially; **Scenario 2** which involved subtasks B and C sequentially; and **Scenario 3** which only involved subtask C.

#### 3.3.2 Participants

Six teams evaluated their systems on eHealth-KD 2018 challenge. These are listed next and classified regarding the characteristics that they used, referring the tag labels described in the next paragraph: [**Team UC3M**] [SDEN] (Zavala, Martínez, and Segura-Bedmar, 2018); [**Team SINAI**] [KRN] (López-Ubeda et al., 2018); [**Team UPF-UPC**] [SKN] (Palatresi and Hontoria, 2018); [**Team TALP**] [DEN] (Medina and Turmo, 2018); [**Team LaBDA**] [DE] (Suarez-Paniagua, Segura-Bedmar, and

<sup>3</sup><http://www.sepln.org/workshops/tass/2018/task-3/>

Martínez, 2018); and **[Team UH]** [RN], which is described in (Martínez-Cámara et al., 2018). The tag labels designed to provide an overview of the characteristics of each system: **[S]** Used shallow supervised models such as CRF, logistic regression, SVM, decision trees, etc; **[D]** Used deep learning models, such as LSTM, convolutional networks, etc; **[E]** Used word embeddings or other embedding models trained with external corpora; **[K]** Used external knowledge bases, either explicitly or implicitly (i.e. through third-party tools); **[R]** Used hand crafted rules based on domain expertise; and **[N]** Used natural language processing techniques or features, i.e., POS-tagging, dependency parsing, etc.

### 3.3.3 Results

A variety of approaches dealt effectively with the health knowledge discovery problem. However, there are still issues to resolve. Classic supervised learning, deep learning and knowledge-based techniques were the best performing submissions, in general. The official results can be found in (Martínez-Cámara et al., 2018) and in the TASS 2018 web site<sup>4</sup>. From them the top results per subtask were:

- Subtasks A and B: Team UC3M which was based on a CRF model with pre-trained embeddings as features. This team got F1 87.2% and Acc 95.9% in both subtasks respectively.
- Subtask C: in concordance with Scenario 3, did not exceed 45% in F-score. This can be considered as the most difficult subtask to deal with, even after having applied novel approaches (i.e. TALP with F1 44.8% and LaBDA with F1 42%) based on convolutional neural networks.

Regarding the top results per scenario. In:

- Scenario 1, the top performing strategy belonged to UC3M with an 74.4% of F1, pretty close to SINAI with 71.0%. These teams got a high F1 basically because their results in the subtasks A and B were high bringing on advantage in the average measure.
- Scenario 2 and 3: the top performing strategy belonged to TALP with an F1 of

72.2% and F1 of 44.8% respectively per scenario. This team reduced its advantage in the overall score, due to this did not submit results for the Scenario 1.

An interesting phenomenon is that the best systems in subtask A were not correlated with the best systems in subtask C. This suggests that the optimal approach for either subtask is different, giving rise to an interesting research line that would explore integrated approaches to simultaneously solving these three subtasks.

### 3.3.4 Analysis of the results

The analysis of the results revealed that subtasks A and B were easier than subtask C for mostly participant teams. In subtask A, around 70% of the annotations in the test set were correctly identified by at least 3 of the participant systems. Likewise, in subtask B, 71% of the annotations were correctly classified by at least 4 systems. On the contrary, 64% of the relations in subtask C were not recognized by any system.

In general, the most competitive approaches in individual tasks were dominated by state-of-the-art machine learning. In the particular case of subtask C, modern deep learning approaches seemed to outperform classic techniques. However, adding domain-specific knowledge, mostly in the form of knowledge bases with health-related concepts, provided a significant boost, even when less powerful learning techniques were used, particularly for key phrase extraction (subtask A). Most participants used NLP features, either explicitly, or implicitly captured in word embeddings and other representations. The best overall systems did not generalize across the three subtasks, while systems that did generalize did not outperform the baseline in general.

## 3.4 Task 4: Good or Bad News?

Emotions are usually related to subjective data. However, the reading of facts, like the ones described in news, may also arise emotions. TASK 4 is motivated by the industrial interest on the identification of the emotions that news headlines may arise on a reader, as they have an indubitable impact in the perception of ads placed along with those articles. The goal of TASK 4 was defined as a binary classification problem: systems had to identify SAFE news (positive emotion) and

<sup>4</sup><http://www.sepln.org/workshops/tass/2018/task-3/index.html#results>

UNSAFE news (negative emotions).

### 3.4.1 Subtasks

Two subtasks were proposed: subtask 1 (S1) was set up as a monolingual classification task, and subtask 2 (S2) as a multilingual classification task.

S1 proposed the classification of headlines into SAFE or UNSAFE without taking into account the Spanish version. Participants were provided with the *training* and *development* subsets of the L1 SANSE corpus, and two *test* sets for the evaluation: the L1 SANSE test subset and the L2 SANSE corpus.

The aim of S2 was to assess the generalization capacity of the submitted systems. Participants were provided with SANSE subsets with headlines written only in the Spanish language spoken in Spain for training. The *test* set was composed of headlines written in the Spanish language spoken in different countries of America. Due to space constraints, the statistics of the SENSE corpus for S2 are shown in the web page of the task.<sup>5</sup>

### 3.4.2 Participants and Results

Seven research groups participated in TASK 4. Four of them submitted the results of their systems on the two subtasks, and three of them on the two runs of S1.

Each group was allowed to submit up to three systems. From all the submitted systems, we highlight the following: (1) most of the submitted systems were grounded in deep learning methods; (2) although most of the neural network systems were based on the use of Recurrent Neural Networks (RNN), specifically the Long Short-Term Memory (LSTM) architecture, Herrera-Planells and Villena-Román (2018) (MEANINGCLOUD) proposed the use of Convolutional Neural Networks (CNN), which reached good results in S1 L1; (3) the top performance system (INGEOTEC) in S1 and S2 was based on the optimization of a set of base linear classification systems using a genetic programming system; and (4) only one group (Plaza del Arco et al., 2018) (SINAI) proposed the incorporation of external knowledge to represent the headlines and subsequently used a linear classification system.

The evaluation measures used the macro-averaged version of the Precision, Recall and F1, as well as the Accuracy. Systems were

<sup>5</sup><http://www.sepln.org/workshops/tass/2018/task-4/>

ranked according to F1. The results of the best systems in the two subtasks are shown in Table 6. The main features of the three best systems are detailed next.

**INGEOTEC.** The system by Moctezuma et al. (2018) was the highest ranked one in S1 and S2. The system was an ensemble method built upon a genetic programming method, EvoDAG (Graff et al., 2017), for optimizing the contribution of each base system.

**ELiRF\_UPV.** The system of González, Hurtado, and Pla (2018a) was based on the Deep Averaging Network (DAN) model. The main contributions were (1) the use of a set of pre-trained vectors of word embeddings in Spanish, which was generated from a set of Spanish tweets; and (2) the conclusion that the language used in news headlines and Twitter must be similar, as the use embeddings trained on tweets is not harmful for the system.

**rbnUGR.** The main contribution of Rodríguez Barroso, Martínez-Cámara, and Herrera (2018) was the comparison of three architectures of RNN for the encoding of the input headlines: (1) taking the last vector state of a LSTM layer; (2) the concatenation of the two last vector states of a Bidirectional LSTM layer; and (3) the concatenation of the output vectors of a LSTM layer per each input token. Results showed that the use of single LSTM layers is more beneficial, and the use of all the output vectors (run\_3) allows to improve the generalization capacity, as it reached better results than the other two systems in S2.

### 3.4.3 Analysis

We conducted an analysis of the difficulty of the subtasks, which consisted on the study of the percentage of headlines correctly classified by the systems of the five groups that submitted a description paper.

We combined the output of the systems of each group<sup>6</sup> by a voting system, which resulted as the overall output of each group. The rate of headlines rightly predicted by the groups in each task is in Table 7. The analysis shows that S1 L2 is the least hard task because all the headlines were at least predicted by one group, as expected due to the fact that the annotation was performed by a voting system built upon the submitted systems.

<sup>6</sup>Three systems were allowed to submit as utmost.

System	S1 L1				S1 L2				S2			
	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc
INGEOTEC_run1	0.794	0.795	0.795 <sup>1</sup>	0.802	0.853	0.880	0.866 <sup>4</sup>	0.871	0.722	0.715	0.719 <sup>1</sup>	0.737
ELiRF_UPV_run2	0.787	0.794	0.790 <sup>2</sup>	0.794	0.850	0.884	0.867 <sup>3</sup>	0.865	0.747	0.657	0.699 <sup>2</sup>	0.722
ELiRF_UPV_run1	0.795	0.784	0.790 <sup>3</sup>	0.800	0.878	0.889	0.883 <sup>1</sup>	0.893	0.736	0.649	0.690 <sup>3</sup>	0.715
rbnUGR_run1	0.784	0.764	0.774 <sup>4</sup>	0.786	0.880	0.867	0.873 <sup>2</sup>	0.888	0.683	0.661	0.672 <sup>6</sup>	0.700
MEANING-CLOUD_run3	0.767	0.767	0.767 <sup>5</sup>	0.776	0.781	0.804	0.793 <sup>7</sup>	0.801	0.647	0.654	0.651 <sup>7</sup>	0.658
rbnUGR_run3	0.763	0.765	0.764 <sup>6</sup>	0.772	0.838	0.870	0.853 <sup>6</sup>	0.853	0.687	0.678	0.683 <sup>4</sup>	0.631
rbnUGR_run2	0.774	0.752	0.763 <sup>7</sup>	0.776	0.868	0.857	0.863 <sup>5</sup>	0.878	0.679	0.672	0.676 <sup>5</sup>	0.698

Table 6: Macro averaged precision (P), recall (R), F1 and accuracy (Acc) reached by each submitted system to each subtask of the groups that submitted a system description paper. The superscripts are the rank order of the submitted systems

In contrast, the 4.40% and 5.57% of the headlines were not predicted by any group in S1 L2 and S2 respectively. Also as expected, the multilingual task (S2) is substantially harder than the monolingual one (S1 L1), because 34.14% of the headlines were classified by as much two groups, whereas only 17% of the headlines in S1.

#	S1 L1 (Acc.)	S1 L2 (Acc.)	S2 (Acc.)
0	4.40	4.40	0
1	5.00	9.40	0.69
2	7.60	17.00	5.35
3	12.00	29.00	13.96
4	23.00	52.00	26.04
5	48.00	100.00	53.94

Table 7: The % of headlines rightly classified by the groups. Only four groups participated in S2. Acc. indicates the accumulative percentage. Column one is the number of groups

Regarding the results shown in Table 6 and the statistics of Table 7, there is still room for improvement. First, only high performance systems can predict the safety meaning of 20% of the headlines, hence more efforts should be done in the design of classification systems to understand the meaning of the headlines. Subsequently, the subtask S2 pointed out the differences among the spoken Spanish versions, showing the need of increasing the generalization capacity of machine learning methods, which is essential for text understanding tasks with documents in different versions of the same language.

## 4 Conclusions

The edition of 2018 of TASS contributed with the organization of two new tasks and the update of the InterTASS corpus and the release of two new ones (eHealth-KD and SANSE).

The systems presented at TASK 1, with the new version of the InterTASS corpora, obtains similar results, with F1 values near to 0.50. This means that systems can achieve better results with this new collection.

TASK 3 was mostly dominated machine learning approaches. Nevertheless, adding domain-specific knowledge, mostly in the form of knowledge bases with health-related concepts, provided a significant boost, even when less powerful learning techniques were used. As future works, new semantic relations will be considered.

TASK 4 showed an industrial application of emotion classification in a monolingual and multilingual environment. There is room for improvement in both environments, because there are headlines that were not rightly classified by any submitted system. As future work, the annotation of the SANSE corpus will be revised with the aim of improving the agreement score.

## Acknowledgments

This work has been partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER), the projects REDES (TIN2015-65136-C2-1-R, TIN2015-65136-C2-2-R), PROMETEU/2018/089 and SMART-DASCI (TIN2017-89517-P) from the Spanish Government. Eugenio Martínez Cámara was supported by the Spanish Government Programme Juan de la Cierva Formación (FJCI-2016-28353).

## References

Chiruzzo, L. and A. Rosá. 2018. RETUYT-InCo at TASS 2018: Sentiment analysis in spanish variants using neural networks and svm. In *Proceedings of TASS 2018*, volume 2172, Sevilla, Spain. CEUR-WS.

- Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.
- González, J.-A., L.-F. Hurtado, and F. Pla. 2018a. ELiRF-UPV en TASS 2018: Categorización emocional de noticias. In *Proceedings of TASS 2018*, volume 2172, Sevilla, Spain. CEUR-WS.
- González, J.-A., L.-F. Hurtado, and F. Pla. 2018b. ELiRF-UPV en TASS 2018: Análisis de sentimientos en twitter basado en aprendizaje profundo. In *Proceedings of TASS 2018*, volume 2172, pages 37–44, Sevilla, Spain. CEUR-WS.
- Graff, M., E. S. Tellez, H. Jair Escalante, and S. Miranda-Jiménez, 2017. *Semantic Genetic Programming for Sentiment Analysis*, pages 43–65. Springer Int. Publishing.
- Herrera-Planells, J. and J. Villena-Román. 2018. MeaningCloud at TASS 2018: News headlines categorization for brand safety assessment. In *Proceedings of TASS 2018*, volume 2172, Sevilla, Spain. CEUR-WS.
- Landis, J. R. and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.
- López-Ubeda, P., M. C. Díaz-Galiano, M. T. Martín-Valdivia, and L. A. Urena-Lopez. 2018. SINAI en TASS 2018 Task 3. Clasificando acciones y conceptos con UMLS en MedLine. In *Proceedings of TASS 2018*, volume 2172, Sevilla, Spain. CEUR-WS.
- Luque, F. M. and J. M. Pérez. 2018. Atalaya at TASS 2018: Sentiment analysis with tweet embeddings and data augmentation. In *Proceedings of TASS 2018*, volume 2172, Sevilla, Spain. CEUR-WS.
- Martínez-Cámara, E., Y. Almeida-Cruz, M. C. Díaz-Galiano, S. Estévez-Velarde, M. A. García-Cumbreiras, M. García-Vega, Y. Gutiérrez, A. Montejo-Ráez, A. Montoyo, R. Muñoz, A. Piad-Morffis, and J. Villena-Román. 2018. Overview of TASS 2018: Opinions, health and emotions. In *Proceedings of TASS 2018*, volume 2172, Sevilla, Spain. CEUR-WS.
- Martínez-Cámara, E., M. C. Díaz-Galiano, M. A. García-Cumbreiras, M. García-Vega, and J. Villena-Román. 2017. Overview of TASS 2017. In *Proceedings of TASS 2017*, volume 1896, Murcia, Spain. CEUR-WS.
- Medina, S. and J. Turmo. 2018. Joint classification of key-phrases and relations in electronic health documents. In *Proceedings of TASS 2018*, volume 2172, Sevilla, Spain. CEUR-WS.
- Moctezuma, D., J. Ortiz-Bejar, E. S. Tellez, S. Miranda-Jiménez, and M. Graff. 2018. INGEOTEC solution for task 4 in TASS'18 competition. In *Proceedings of TASS 2018*, volume 2172, Sevilla, Spain. CEUR-WS.
- Montanés, R., R. Aznar, and R. del Hoyo. 2018. Aplicación de un modelo híbrido de aprendizaje profundo para el análisis de sentimiento en twitter. In *Proceedings of TASS 2018*, volume 2172, Sevilla, Spain. CEUR-WS.
- Palatresi, J. V. and H. R. Hontoria. 2018. TASS2018: Medical knowledge discovery by combining terminology extraction techniques with machine learning classification. In *Proceedings of TASS 2018*, volume 2172, Sevilla, Spain. CEUR-WS.
- Plaza del Arco, F. M., E. Martínez-Cámara, M. T. Martín Valdivia, and A. Ureña López. 2018. SINAI en TASS 2018: Inserción de conocimiento emocional externo a un clasificador lineal de emociones. In *Proceedings of TASS 2018*, volume 2172, Sevilla, Spain. CEUR-WS.
- Rodríguez Barroso, N., E. Martínez-Cámara, and F. Herrera. 2018. SCI<sup>2</sup>S at TASS 2018: Emotion classification with recurrent neural networks. In *Proceedings of TASS 2018*, volume 2172, Sevilla, Spain. CEUR-WS.
- Suarez-Paniagua, V., I. Segura-Bedmar, and P. Martínez. 2018. ABDA at TASS-2018 task 3: Convolutional neural networks for relation classification in spanish ehealth documents. In *Proceedings of TASS 2018*, volume 2172, Sevilla, Spain. CEUR-WS.
- Zavala, R. M. R., P. Martínez, and I. Segura-Bedmar. 2018. A hybrid Bi-LSTM-CRF model for knowledge recognition from ehealth documents. In *Proceedings of TASS 2018*, volume 2172, Sevilla, Spain. CEUR-WS.

# La modelización de la morfología verbal bribri

## *Modeling the verbal morphology of Bribri*

Sofía Flores-Solórzano

Universidad de Costa Rica, Sede del Atlántico, 30501 Turrialba, Costa Rica

sofia.flores.s@gmail.com

**Resumen:** La modelización computacional de la morfología verbal bribri (lengua de la estirpe Chibchense hablada en el sudeste de Costa Rica) es posible mediante el uso de la tecnología de los transductores de estados finitos (FST). En este trabajo se desarrolla un analizador y generador morfológico a partir de las descripciones gramaticales existentes. Se han definido 11 paradigmas verbales correspondientes a 6 conjugaciones orales y 5 conjugaciones nasales, además de una serie de reglas de alteración fonológica y ortográfica. El sistema soporta casos complejos como los verbos intrínsecamente medios, defectivos o que presenten supletivismo en la raíz, así como la concatenación de sufijos y direccionales.

**Palabras clave:** Bribri, lenguas indígenas, morfología, análisis morfológico, transductor de estados finitos

**Abstract:** Computational modeling of the verbal morphology of Bribri (Chibchan family language currently spoken in southeastern Costa Rica) is made possible through the use of finite state transducers (FST). A morphological analyzer and generator have been developed from the existing grammatical descriptions. A total of 11 verbal paradigms corresponding to 6 oral conjugations and 5 nasal conjugations were defined and implemented in the transducer along with a number of phonological and orthographic alteration rules. The system covers complex cases such as middle verbs, defective verbs and suppletivism, as well as the concatenation of suffixes and directionals.

**Keywords:** Bribri, indigenous languages, morphology, morphological analysis, finite state transducer

## 1 Introducción

El bribri es una lengua de la estirpe chibchense que hablan en la actualidad aproximadamente unos 7.000 indígenas en el sudeste de Costa Rica (Instituto Nacional de Estadística y Censos, 2013). Hasta la fecha, es probablemente la lengua indígena del país más estudiada y descrita. Se encuentra por lo tanto en una situación favorable para llevar a cabo una investigación aplicada en el campo de la lingüística computacional o el procesamiento del lenguaje natural. Precisamente el analizador morfológico que presentamos constituye un primer intento de modelizar computacionalmente la morfología bribri<sup>1</sup>.

El sistema de escritura bribri emplea una serie de diacríticos para representar la tonalidad y la nasalidad; para su introducción se utilizó la herramienta Teclado Chibcha (Flores Solórzano, 2010) que consta de un mapa del teclado y una fuente Unicode que apila adecuadamente los diacríticos, estos se encuentran soportados en el analizador.

El lexicón fue elaborado con el formalismo *Lexc* (*Finite-State Lexicon Compiler*, Compilador Léxico de Estados Finitos.). Este es un lenguaje de programación de tipo declarativo de alto nivel que se usa específicamente para la elaboración de lexicones de lenguas naturales. La sintaxis de este lenguaje está diseñada para facilitar la definición de la estructura morfológica y el tratamiento de algunas irregularidades y ambigüedades ortográficas (Beesley y Karttunen, 2003).

La formulación y ordenamiento de las reglas en cascada, que contienen las alteraciones fonológicas y ortográficas necesarias, se realizó mediante el formalismo denominado *Replace Rules*, reglas de sustitución, desarrollado por Beesley y Karttunen (2003). Las reglas de sustitución son parecidas a las reglas de reescritura utilizadas por la fonología tradicional, incluidas en los trabajos de Chomsky y Halle (1968). Son básicamente expresiones regulares que utilizan unos operadores básicos; su notación es ideal para que los lingüistas puedan definir relaciones complejas de estados finitos de una manera bastante cómoda y familiar (Beesley y Karttu-

<sup>1</sup>La herramienta se encuentra disponible para ser evaluada en la URL <http://morphology.bribri.net>.

nen, 2003: 132–137).

Finalmente, con la herramienta *Foma* (Hulden, 2009) realizamos la composición del lexicón y las reglas de sustitución en una única red de estados finitos denominada transductor léxico. El transductor contiene la información morfológica de la lengua bribri: morfemas léxicos, derivativos, inflexivos, alteraciones, infijaciones, supletivismo, interdigitación, composición, etc. Funciona tanto para el análisis como para la generación, en otras palabras, es bidireccional.

Actualmente existen transductores léxicos y *taggers* (etiquetadores) para casi todas las lenguas de prestigio. También se han desarrollado analizadores morfológicos para un limitado número de lenguas indígenas de América; Mager et al. (2018) citan el quechua (Rios, 2010), el mapudungun (Chandía, 2012), el mohawk (Assini, 2013), el plains cree (Snoek et al., 2014), el odawa (Bowers et al., 2017), entre otras. De acuerdo con Mager et al. (2018), el modelo de estados finitos es el más extendido entre los casos que estos investigadores documentan.

En el siguiente apartado describiremos en detalle el modelado de un paradigma verbal bribri en un transductor de estados finitos.

## 2 La morfología verbal bribri

Uno de los principales retos de la investigación ha sido modelar el sistema verbal bribri. Dicha modelización se ha basado principalmente en las descripciones de Constenla, Elizondo, y Pereira (1998); Margery Peña (1982) y Jara (2013). Asimismo, hemos contado con la consultoría de los maestros bribris Julio Morales Campos y Franklin Morales Morales, ambos tienen una amplia experiencia en la escritura y enseñanza de su lengua. A continuación nos referiremos a los elementos más esenciales del sistema verbal bribri, y posteriormente a su modelización computacional.

El **tema verbal** se obtiene apartando la consonante glotal (saltillo) o el tono descendente de las formas del perfecto improspectivo activo<sup>2</sup> (Tabla 1). El perfecto improspectivo es la base del tema verbal y es la única forma que permite predecir las formas de la voz media de la conjugación. Por ejemplo, *apàka'* ‘narré, visité’ es el perfecto improspectivo activo del verbo *apakók* ‘narrar, visitar’, cuyo tema verbal es *apàka* (la forma improspectiva sin el saltillo).

El sistema verbal bribri se divide en **verbos orales** y **verbos nasales**. La última vocal del tema del verbo es su vocal temática. Los verbos orales son

Perfecto Improspectivo	Tema
<i>apaka'</i>	<i>apaka-</i> narrar
<i>të'</i>	<i>të-</i> cortar, punzar
<i>ali'</i>	<i>ali-</i> cocinar
<i>yulö'</i>	<i>yulö-</i> buscar
<i>u'</i>	<i>u-</i> moler
<i>ñq'</i>	<i>ñq-</i> comer alimentos suaves
<i>tse'</i>	<i>tse-</i> llevar, traer
<i>ini'</i>	<i>ini-</i> jugar
<i>tó</i>	<i>tó-</i> comprar, costar, valer
<i>ku'</i>	<i>ku-</i> encontrar, jalar
<i>sú</i>	<i>sú-</i> ver
<i>wö'ík</i>	<i>wö'ík-</i> soplar
<i>báts</i>	<i>báts-</i> unir
<i>apëit</i>	<i>apëit-</i> prestar, alquilar

Tabla 1: Obtención del tema verbal bribri

aquellos cuyo tema termina en una vocal oral y los verbos nasales, en una vocal nasal. Los temas verbales que terminan en consonante, como *wö'ík* ‘soplar’, *báts* ‘unir’ o *apëit* ‘prestar’, constituyen un paradigma aparte y generalmente son tratados como verbos orales, aunque hay excepciones; por ejemplo, hemos observado que el tema verbal *bi-kéits* ‘pensar’, el cual termina en consonante, es tratado por algunos hablantes como un verbo oral, i.e. *bikéitsók*<sup>3</sup>, mientras que por otros como nasal *bi-kéitsuk*<sup>4</sup>.

Los verbos orales se dividen en 6 conjugaciones determinadas por la vocal temática oral o la consonante final del tema, a saber: **-a**, **-ë**, **-i**, **-ö**, **-u** y **-Cons**. Los verbos nasales se subdividen a su vez en 5 conjugaciones según su vocal temática: **-a**, **-e**, **-i**, **-o** y **u**.

La **voz** es la categoría más importante del sistema verbal bribri. Los verbos transitivos e intransitivos suelen presentar **voz activa** y **voz media**. La voz media se usa generalmente para excluir la presencia de un agente. Hay algunos verbos que se refieren a procesos que carecen de conjugación activa y solo se conjugan en la voz media, son verbos defectivos medios i.e. *kiànuk* ‘querer’, *tsikìnuk* ‘nacer’, *sènuk* ‘vivir’ etc.

Hay al menos **cinco modos**: indicativo, imperativo, de finalidad, optativo y adversativo. El único modo que presenta oposiciones de aspecto y tiempo es el modo indicativo.

En bribri **el aspecto** es más determinante que el tiempo. Es importante indicar si la acción se ha completado (aspecto perfectivo) o si aún se está desarrollando, va a desarrollarse en el futuro o pue-

<sup>2</sup>Jara (2013) lo denomina perfecto remoto.

<sup>3</sup>El morfema *-ók* indica el infinitivo de los verbos orales.

<sup>4</sup>*-ík* indica el infinitivo de los verbos nasales.

Tema improspectivo transitivo ali- ‘cocinar’					
Voz activa					
Imperfecto primero	<i>al-</i>	<i>è</i> →	Imperfecto habitual	<i>alè-</i>	<i>ke</i>
Infinitivo	<i>al-</i>	<i>ók</i>	Imperfecto habitual negativo	<i>alè-</i>	<i>ku</i>
Modo imperativo	<i>al-</i>	<i>ó</i>	Imperfecto potencial	<i>alè-</i>	<i>mi</i>
Modo imperativo negativo	<i>al-</i>	<i>ar</i>	Imperfecto futuro	<i>alè-</i>	<i>dá ~ rá</i>
Modo de finalidad	<i>al-</i>	<i>ó</i>	Imperfecto futuro negativo	<i>alè-</i>	<i>pa</i>
Modo optativo	<i>al-</i>	<i>a'ky</i>			
Modo adversativo	<i>al-</i>	<i>a'</i>			
Perfecto prospectivo	<i>al-</i>	<i>é</i>			
Perfecto improspectivo	<i>al-</i>	<i>i'</i>			
		↓			
Voz Media					
Imperfecto primero	<i>alì-</i>	<i>r</i> →	Imperfecto habitual	<i>alìr-</i>	<i>ke</i>
Infinitivo	<i>alì-</i>	<i>nyk</i>	Imperfecto habitual negativo	<i>alìr-</i>	<i>ku</i>
Perfecto prospectivo	<i>alì-</i>	<i>na</i>	Imperfecto potencial	<i>alìr-</i>	<i>mi</i>
Perfecto improspectivo	<i>alì-</i>	<i>ne</i>	Imperfecto futuro	<i>alìr-</i>	<i>dá</i>
Modo de finalidad	<i>alì-</i>	<i>no</i>	Imperfecto futuro negativo	<i>alìr-</i>	<i>pa</i>
			Forma anterior	<i>alìr-</i>	<i>ule</i>

Tabla 2: Conjugación de los verbos con vocal temática -i

de desarrollarse y está incompleta (aspecto imperfectivo). Hay dos formas perfectivas: el perfecto improspectivo y el perfecto prospectivo. Y cinco formas imperfectivas: el imperfecto primero, el imperfecto habitual, el imperfecto potencial, el imperfecto futuro y la forma anterior. Los verbos bribris carecen de morfemas que especifiquen exclusivamente el tiempo. Las formas perfectivas se refieren generalmente al pasado remoto y al presente; y las formas imperfectivas se pueden referir al pasado, al presente y al futuro. En las traducciones al español se debe apelar al contexto para decidir muchas veces qué tiempo utilizar.

## 2.1 Paradigma de los verbos terminados en -i

Por razones de espacio no nos referiremos a todas las conjugaciones del sistema verbal bribri. Nuestra intención es explicar la morfotaxis de los verbos regulares mediante un paradigma verbal: la conjugación de los verbos transitivos terminados en -i, y así mostrar su implementación en un transductor de estados finitos.

Los verbos con la vocal temática -i se conjugan en su mayoría de manera regular (Tabla 2). Obsérvese como la vocal temática -i cae o desaparece en la voz activa, y se manifiesta solo en las formas medias y en el perfecto improspectivo. Este comportamiento de la vocal temática es típico de las conjugaciones regulares del sistema verbal bribri, y por consiguiente, de la conjugación de los verbos en -i<sup>5</sup>, como se observa en los siguientes ejemplos:

<sup>5</sup>Sin embargo, es importante tener en cuenta que hay al-

*alók* (ali') ‘cocinar’, *ichàkók* (ichàki') ‘preguntar’, *tsakók* (tsaki') ‘reventar algo’, *inúk* (inì')’, etc. La primera forma es el infinitivo y la forma entre paréntesis, el perfecto improspectivo.

En la conjugación de la voz activa se agregan una serie de desinencias (-è, -ók, -ó, -ar, -ó, -a'ky, -a', -é) a la raíz del tema verbal (Tabla 2). Entendemos por raíz verbal el tema de verbo sin la vocal temática.

Asimismo, nótese que a partir del imperfecto primero activo, se derivan todas las demás formas imperfectivas activas mediante la concatenación de las correspondientes terminaciones (-ke, -ku, -mi, -dá ~ -rá, -pa).

En la voz media, las terminaciones medias (-r, -nyk, -na, -no) se agregan al tema verbal –la raíz verbal más la vocal temática–, y la vocal temática cambia de tono bajo a tono alto o descendente.

La forma anterior y los imperfectos medios también se derivan del imperfecto primero medio, al cual se le concatenan las desinencias imperfectivas y la desinencia de la forma anterior (-ule).

gunos verbos cuyo tema termina en -i que mantienen la vocal temática en la raíz del tema verbal, es decir, no se produce la caída de la vocal, i.e. *kiók* (ki') ‘llamar’, *biók* (bi') ‘escarbar’, *tuúk* (tuí) ‘anochecer’. En este último ejemplo, *tuúk* (tuí), la conjugación es nasal a pesar de que el tema verbal sea oral. Como se ejemplifica en los casos expuestos, parece que si el tema es monosilábico y termina en una vocal cerrada anterior -i, hay una clara tendencia a conservar la vocal en la conjugación activa.

Declaración de los símbolos multi-carácter (etiquetas)	
<b>Lexicon Root</b>	Verbos ; Sustantivos ;
<b>Lexicon Verbos</b>	
morfema <i>upper:lower</i>	<i>continuum class</i> ; ! comentario

Tabla 3: Estructura de un archivo *Lexc*

## 2.2 Implementación de los verbos terminados en -i en un transductor de estados finitos

Como señalamos en la Introducción, el lexicón se elaboró con el Compilador Léxico de Estados Finitos (*Lexc*). La estructura de un archivo *Lexc* tiene tres secciones: a) La declaración de los símbolos multi-carácter o etiquetas; b) una declaración obligatoria que se denomina *Lexicon Root*, donde se enumeran las principales clases del lexicón (generalmente corresponden a las categorías gramaticales, i.e. verbos, sustantivos, adjetivos, pronombres etc.); y c) las diferentes clases del lexicón, *Lexicon classes*, con la descripción morfológica (Tabla 3).

La declaración de los símbolos multi-carácter es opcional y en ella se enumeran las etiquetas que se han definido en las clases del lexicón. La declaración es útil para que el transductor interprete los símbolos multi-carácter precisamente como símbolos o etiquetas y no como cadenas de símbolos, o sea, que interprete un símbolo como +Sust como un único símbolo multi-carácter y no como cinco símbolos: + S u s t.

La declaración *Lexicon Root* es obligatoria. Debajo de ésta se enumeran las principales clases del lexicón. Los nombres de las clases son de libre elección, pero deben finalizar con un punto y coma (;). Las clases enumeradas en esta sección no son las únicas clases presentes en el lexicón sino las que se encuentran en el punto de inicio de la red.

Las clases se introducen con la palabra LEXICON seguida del nombre de la clase. Debajo se describen las diferentes bases léxicas o morfemas; un morfema por línea con su clase continua (*continuum class*) o en su defecto, el símbolo numeral o almohadilla (#) que indica que la cadena no va a admitir más morfemas. Después un punto y coma (;) obligatorio que indica el fin de la cadena. Optativamente se puede añadir un comentario que se indica con el signo de exclamación (!). En nuestro caso, usamos el comentario para indicar la traducción al español de las bases léxicas o morfemas.

En el siguiente fragmento del Lexicón, escrito en *Lexc*, mostramos la modelización de la conjugación de los verbos de tema no monosilábico terminados en -i en un transductor léxico de estados finitos (ejemplo 1). Por razones de espacio no presentamos la clase continua de los direccionales y otros sufijos que pueden presentar los verbos bribris. En figura 1 se observa la visualización del transductor.

- (1) Multichar\_Symbols + V + Imp1Tran è + Imp2 ke + Imp2Neg ku + ImpPot mi + ImpFut dâ + ImpFutNeg pa + Inf ók + ModoImp ó + ModoImpNeg ar + ModoFin ó + ModoOpt a'ku + ModoAdvers a' + Prosp é + PerfImp + VozMedia uk a e + Imp1 ó + Anterior ule

LEXICON Root  
Verbos ;

LEXICON Verbos  
al -i ; ! cocinar tr.

LEXICON -i  
i TemaImprospectivo ;  
i:0 TemaVozActivaOral ;

LEXICON TemaVozActivaOral  
+ V:0 TiemposVozActivaOral ;

LEXICON TiemposVozActivaOral  
+ Imp1Tran:è TiemposImperfectoPrimero ;  
+ Inf:ók # ;  
+ ModoFin:ó # ;  
+ Prosp:é # ;  
+ ModoOpt:a'ku # ;  
+ ModoAdvers:a' # ;  
+ ModoImp:ó # ;  
+ ModoImpNeg:ar # ;

LEXICON TemaImprospectivo  
+ V:0 Improspectivo ;

LEXICON Improspectivo  
+ VozMedia:d TiemposVozMedia ;  
+ PerfImp:' # ;

LEXICON TiemposVozMedia  
+ Prosp:a # ;  
+ PerfImp:e # ;  
+ Inf:uk # ;  
+ Imp1:0 TiempoAnterior ;  
+ ModoFin:ó # ;

LEXICON TiempoAnterior  
+ Anterior:ule # ;  
0 TiemposImperfectoPrimero ;

LEXICON TiemposImperfectoPrimero	
+ Imp2:ke	# ;
+ Imp2Neg:ku	# ;
+ ImpFut:dâ	# ;
+ ImpFutNeg:pa	# ;
+ ImpPot:mi	# ;
0	# ;

De acuerdo con la morfotáctica definida en el ejemplo 1, la base o raíz verbal<sup>6</sup>, en este caso, *al*, tiene como clase continua la clase *-i* (correspondiente a la vocal temática). A su vez la clase *-i* presenta dos recorridos posibles: si la cadena de entrada termina con la vocal *-i*, su clase continua es la clase *TemaImprospectivo*; si la entrada no lleva la vocal temática (de ahí el símbolo 0), le corresponde la clase *TemaVozActivaOral*.

En las clase *TemaVozActivaOral*, la cadena de entrada se contrasta con el elemento “+V”, símbolo multi-carácter que se encuentra en el lado superior del red, y concatena este símbolo en el lado superior del recorrido.

En la clase *TiemposVozActivaOral*, la cadena de entrada se vuelve a contrastar con el lado inferior de la red, donde se encuentran los morfemas del imperfecto primero transitivo, el infinitivo, el modo de finalidad, el prospectivo, el modo optativo, el modo adversativo, el modo imperativo y el modo imperativo negativo. En caso de existir correspondencia, se concatena el símbolo correspondiente en el lado superior. Por razones prácticas, hemos llamado a estas clase *TiemposVozActivaOral*, aunque somos conscientes de que la mayoría de los morfemas son de modo.

Tras concatenar alguna de las formas léxicas identificadas con los símbolos multi-carácter +Imp1Tran, +Inf, +ModoFin, +Prosp, +ModoOpt, +ModoAdvers, +ModoImp o +ModoImpNeg, todas las cadenas finalizan el recorrido, como lo indica el símbolo almohadilla, con excepción de la cadena que presente *è* en el lado inferior, que continúa hacia la clase siguiente: *TiemposImperfectoPrimero*.

Por otro lado, si en la forma superficial de la cadena de entrada está presente la vocal temática, a la cadena le corresponde la morfotáctica de las clases *TemaImprospectivo*, *Improspectivo* y *TiemposVozMedia*.

Finalmente, apuntar que *-d* es el morfema de la voz media bribri. En el nivel fonológico /d/ tiene dos alófonos en distribución complementaria, [r] y [n]. Mediante tres reglas de alteración aplicadas en paralelo, hemos modelado la realización del fonema /d/:

(2) define Alófonos

<sup>6</sup>La raíz verbal es el tema verbal sin la vocal temática.

d (->) r	+V tono   V'   V _ [ .#.   Vo   Cons ] ,,
d (->) n	V tono   V _ [ Vn .#.   Vn ] ,,
d (->) n	Vn tono _ .#. ;

De acuerdo con la primera regla del ejemplo 2, *d* puede optativamente realizarse como *r* antes de una vocal con o sin tono o saltillo y seguida de posición final de palabra o vocal oral o consonante. Según la segunda regla, *d* se realiza como nasal, *n*, antes de una vocal con o sin tono seguida de una vocal nasal que puede estar en posición final. La tercera regla indica que la realización nasal de /d/ puede darse antes de una vocal nasal con tono y seguida de posición final.

### 3 Resultados

La composición del lexicón junto con las reglas de alteración produce un transductor de estados finitos compuesto por 10.609 estados y 38.515 arcos. El análisis morfológico automático (*apply up*) de la conjugación del verbo *alók* en dicho transductor arroja los siguientes resultados:

- (3) alè  
ali + V + Imp1Tran  
+ Adv [duda]
- alók  
ali + V + Inf
- aló  
ali + V + ModoImp
- alar  
ali + V + ModoImpNeg
- aló  
ali + V + ModoFin
- ala'ku  
ali + V + ModoOpt
- ala'  
ala' + Sust  
ali + V + ModoAdvers
- alé  
ali + V + Prosp
- alèke  
ali + V + Imp1Tran + Imp2
- alèku  
ali + V + Imp1Tran + Imp2Neg
- alèmi  
ali + V + Imp1Tran + Suf [incoativo]  
ali + V + Imp1Tran + Dir

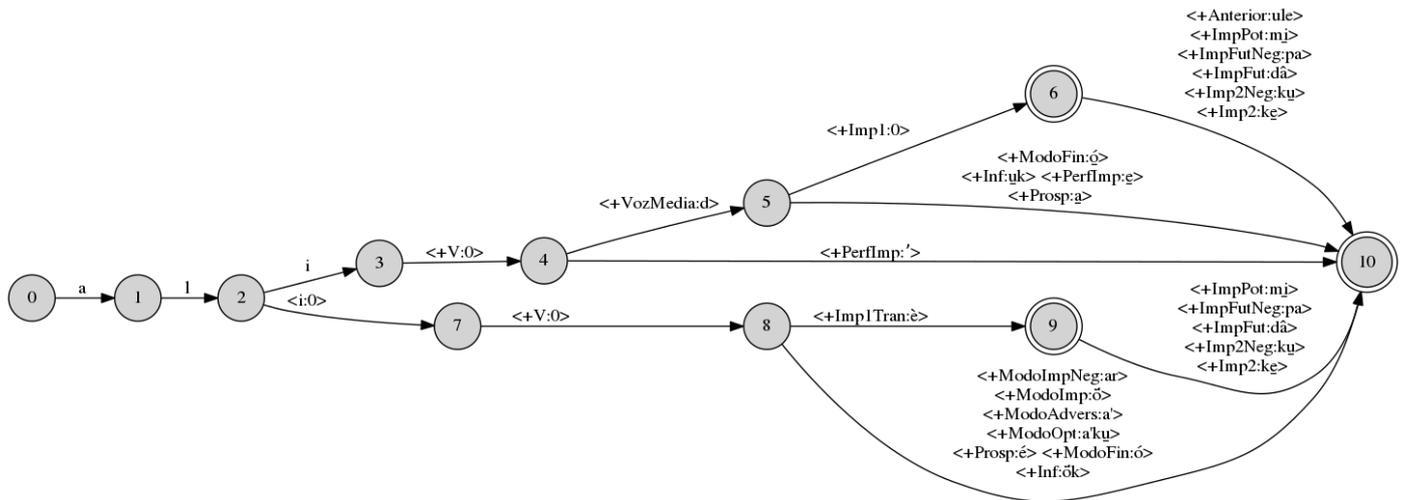


Figura 1: Ejemplo de la modelización en una red de estados finitos de tipo transductor de la conjugación del verbo *alók* ‘cocinar’ (Gráfico generado automáticamente con *Foma* y el paquete *GraphViz*)

ali + V + Imp1Tran + ImpPot

alèdâ

ali + V + Imp1Tran + ImpFut

alèrà

ali + V + Imp1Tran + ImpFut

alèpa

ali + V + Imp1Tran + ImpFutNeg

alìr

ali + V + VozMedia + Imp1

alìnuḱ

ali + V + VozMedia + Inf

alìna

ali + V + VozMedia + Prosp

alìne

ali + V + VozMedia + PerfImp  
 ali + V + PerfImp + Suf[reiterativo]  
 ali' + Sust + Suf[iterativo]

alìnó

ali + V + VozMedia + ModoFin

alìrke

ali + V + VozMedia + Imp1 + Imp2

alìrku

ali + V + VozMedia + Imp1 + Imp2Neg

alìrmi

ali + V + VozMedia + Imp1 + Suf[incoativo]  
 ali + V + VozMedia + Imp1 + Dir  
 ali + V + VozMedia + Imp1 + ImpPot

alìrdâ

ali + V + VozMedia + Imp1 + ImpFut

alìrpa

ali + V + VozMedia + Imp1 + ImpFutNeg

alìrule

ali + V + VozMedia + Imp1 + Anterior

El analizador nos devuelve la cadena que se encuentra en el lado superior de la red (*upper-side*) más una serie de etiquetas que se componen de un operador “+” y una categoría gramatical. Como puede comprobarse, las “etiquetas” contienen la información morfosintáctica.

Los resultados del análisis automático del ejemplo 3 son satisfactorios, pues todas las formas léxicas –aquellas que se encuentran en el lado superior de la red– existen en la lengua bribri. En el ejemplo 3, la mayoría de los casos presentan el tema verbal, *ali*, más el símbolo multi-carácter +V, que indica que se trata de una forma verbal. A continuación, dependiendo de la forma de entrada, pueden concatenarse otros símbolos multi-carácter, que indican el aspecto, modo y tiempo verbal.

En algunos casos del ejemplo 3, el análisis arroja más de un resultado e incluso otras categorías gramaticales, como *alè* +Adv[duda] y *ala'* +Sust, correspondientes con el adverbio de duda *alè* ‘tal vez, quizás’<sup>7</sup> y el sustantivo *ala'* ‘olor’. Esas formas léxicas son homónimas de las formas verbales *alè* ‘cocino’ y *ala'* ‘cocino para molestar a alguien’ (modo adversativo).

Cuando el análisis arroja más de una forma léxi-

<sup>7</sup>Constenla, Elizondo, y Pereira (1998) lo llama partícula de duda.

ca, evidentemente, se trata de casos de homonimia que existen en la lengua y que el desarrollo actual de la herramienta no nos permite desambiguar porque aún no hemos trabajado en el nivel sintáctico; en otras palabras, aunque podemos analizar morfológicamente frases y enunciados, la herramienta no tiene en cuenta el contexto de la frase u oración.

Por cierto, con excepción de los símbolos multicarácter definidos como tales, los diacríticos bribris son tratados por el transductor como símbolos individuales, lo cual permite distinguir y analizar de manera correcta pares bribris que solo se distinguen por el tono o la nasalidad. Por ejemplo, la distinción dentro de los verbos transitivos, entre el imperfecto primero y el perfecto prospectivo, solo se marca por el tono alto y el tono descendente, respectivamente, i.e. *kékará ye' árrós alè* 'yo cocino arroz siempre' y *ye' árrós alé chki* 'yo cociné arroz ayer', esta distinción se encuentra soportada en la herramienta.

Por último, conviene recordar que la escritura bribri no posee un sistema estandarizado; además coexisten al menos tres variantes dialectales, el bribri de Amubri, Coroma y Salitre. Como es común en muchas lenguas indígenas de América, los hablantes de cada variante defienden un sistema de escritura propio que refleje fielmente sus particularidades fonéticas. Así, el verbo *stsók* 'cantar', documentado con -s inicial en Constenla *et al.* (1998) y Margery Peña (1982) –ambos investigadores documentan principalmente los dialectos de Amubri y Salitre– pierde esta s en el dialecto de Coroma, hecho que se refleja en la escritura de esa forma en la variedad de Coroma y posiblemente en la variedad de Amubri, *tsók* 'cantar'.

También el verbo *ijtsók* 'sentir' sufre aféresis de -ij en los tres dialectos, hecho que se refleja en la forma escrita utilizada mayoritariamente, *tsók* 'sentir'; por consiguiente, esta forma es homónima de *tsók* 'cantar'.

En todos estos casos, hemos decidido recoger en el lexicón todas las formas documentadas y utilizadas. Empleamos el mecanismo *upper:lower* del lenguaje *Lexc* para reflejar estos cambios, como se observa en el ejemplo 4:

- (4) sts -ë ; ! cantar intr.  
sts:ts -ë ; ! cantar intr.

ijts -ë ; ! sentir, oír tr.  
ijts:ts -ë ; ! sentir, oír tr

Tras compilar la gramática, el análisis de *stsók* devuelve una única forma léxica:

- (5) apply up > stsók  
stsë + V + Inf

Mientras que el análisis de *tsók* devuelve dos formas léxicas:

- (6) apply up > tsók  
ijtsë + V + Inf  
stsë + V + Inf

Como puede comprobarse en el ejemplo 6, las dos formas léxicas son válidas para nuestro sistema, y también para el sistema de escritura bribri actual. Asimismo, cuando realizamos el proceso contrario, la forma léxica *stsë + V + Inf* genera (*apply down*) los siguientes resultados:

- (7) apply down > stsë + V + Inf  
stsók  
tsók

Es decir, dos formas superficiales válidas (ejemplo 7).

#### 4 Conclusiones

El presente trabajo se ha focalizado en modelar el sistema verbal bribri. Lo más destacable a este respecto ha consistido en la elaboración de los paradigmas verbales. La descripción formal que hemos realizado de la morfología verbal bribri nos ha permitido mejorar la comprensión del funcionamiento del sistema verbal y realizar satisfactoriamente su implementación en *Lexc*.

Una de los principales limitaciones de trabajar computacionalmente con lenguas no mayoritarias es la falta de corpora y recursos lingüísticos. En consecuencia, la evaluación del analizador se ha llevado a cabo con el único corpus (Flores Solórzano, 2017) elaborado hasta la fecha, compuesto por textos provenientes del habla cotidiana que fueron grabados y transcritos entre los años 2014 y 2016. La cobertura de este corpus es del 100 %, con excepción de algunos préstamos castellanos. Somos conscientes de la necesidad de aumentar su tamaño y también de la conveniencia de utilizar otros corpus. Parte del trabajo futuro será precisamente aumentar el corpus con más textos provenientes del habla cotidiana, y al mismo tiempo reunir las publicaciones existentes, solicitar los respectivos permisos y derechos de autor, para contar con un corpora sólido que nos permita no solo mejorar la cobertura del analizador, sino desarrollar otras herramientas útiles para contribuir a la documentación y estudio de esta lengua.

#### Agradecimientos

Esta investigación ha sido posible gracias al financiamiento otorgado por la Vicerrectoría de Investigación de la Universidad de Costa Rica mediante el proyecto B5233.

#### Simbología

(->) Sustitución optativa.

.#. Marcador de límite de palabra.

- \_ Marca la posición contextual del elemento que se sustituye en una regla de sustitución.
- | Operador que indica la unión de dos conjuntos.
- || Marca el contexto en el que se va a aplicar la regla.

+**Adv** Adverbio.

+**Anterior** Forma anterior.

**Cons** Consonante.

+**Dir** Direccional.

+**Imp1** Imperfecto primero (medio).

+**Imp1Tran** Imperfecto primero transitivo.

+**Imp2** Imperfecto segundo.

+**Imp2Neg** Imperfecto segundo negativo.

+**ImpFut** Imperfecto futuro.

+**ImpFutNeg** Imperfecto futuro negativo.

+**ImpPot** Imperfecto potencial.

+**Inf** Infinitivo.

+**ModoAdvers** Modo adversativo.

+**ModoFin** Modo de finalidad.

+**ModoImp** Modo imperativo.

+**ModoImpNeg** Modo imperativo negativo.

+**ModoOpt** Modo optativo.

+**PerfImp** Perfecto improspectivo.

+**Prosp** Prospectivo.

+**Suf** Sufijo.

+**Sust** Sustantivo.

+**V** Verbo.

**Vn** Vocal nasal.

**Vo** Vocal oral.

+**VozMedia** Voz media.

## Bibliografía

- Assini, A. A. 2013. Natural language processing and the mohawk language: creating a finite state morphological parser of mohawk formal nouns. Master's thesis.
- Beesley, K. R. y L. Karttunen. 2003. *Finite state morphology*. CSLI (Center of the Study of Language and Information), Stanford, California.
- Bowers, D., A. Arppe, J. Lachler, S. Moshagen, y T. Trosterud. 2017. A morphological parser for odawa. En *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, páginas 1–9.
- Chandía, A. 2012. Dunggupeyem: analizador y generador morfológicos para mapudungun.
- Chomsky, N. y M. Halle. 1968. *The sound pattern of English*. Harper and Row, New York.
- Constenla, A., F. Elizondo, y F. Pereira. 1998. *Curso básico de bribri*. Editorial de la Universidad de Costa Rica, San José.
- Flores Solórzano, S. 2010. Teclado chibcha: Un software lingüístico para los sistemas de escritura de las lenguas bribri y cabécar. *Revista de Filología y Lingüística de la Universidad de Costa Rica*, 36(2):155–161.
- Flores Solórzano, S. 2017. Corpus oral pandialectal oral de la lengua bribri. Disponible en <http://bribri.net>.
- Hulden, M. 2009. Foma: a finite-state compiler and library. En *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 29–32. Association for Computational Linguistics.
- Instituto Nacional de Estadística y Censos. 2013. *X Censo Nacional de Población y VI de Vivienda: Territorios Indígenas*. INEC, San José, Costa Rica, 1 edición, abril.
- Jara, C. V. 2013. Morfología verbal de la lengua bribri. *Estudios de Lingüística Chibcha*, 32:95–152.
- Mager, M., X. Gutierrez-Vasques, G. Sierra, y I. Meza. 2018. Challenges of language technologies for the indigenous languages of the americas. En *In Proceedings of the 27th International Conference on Computational Linguistics*.
- Margery Peña, E. 1982. *Diccionario fraseológico bribri-español, español-bribri*. Editorial Universidad de Costa Rica, San José.
- Rios, A. 2010. *Applying Finite-State Techniques to a Native American Language: Quechua*. Ph.D. thesis, Institut für Computerlinguistik, Universität Zurich.
- Snoek, C., D. Thunder, K. Lõo, A. Arppe, J. Lachler, S. Moshagen, y T. Trosterud. 2014. Modeling the noun morphology of plains cree. 06.

***Tesis***



# La interfaz estructura informativa-prosodia: el rol de la tematicidad jerárquica basado en un modelo empírico

## *The Information Structure–Prosody Interface: On the Role of Hierarchical Thematicity in an Empirically-grounded Model*

Mónica Domínguez Bajo

Universitat Pompeu Fabra

C. Roc Boronat, 138

08018 Barcelona

monica.dominguez@upf.edu

**Resumen:** Tesis doctoral en tecnologías de la información y las comunicaciones realizada por Mónica Domínguez Bajo en la Universidad Pompeu Fabra bajo la dirección del Dr. Leo Wanner y la Dra. Mireia Farrús Cabecerán. El acto de defensa tuvo lugar el viernes 17 de noviembre de 2017 ante el tribunal formado por los doctores Bernd Möbius (Universidad de Saarland), Pilar Prieto Vives (Universidad Pompeu Fabra) y Catherine Lai (Universidad de Edimburgo). Obtuvo la calificación de Sobresaliente.

**Palabras clave:** Estructura informativa, estructura comunicativa, tematicidad, tema, rema, prosodia, parámetros acústicos, síntesis de voz, texto a habla, concepto a habla, etiquetado automático de prosodia

**Abstract:** PhD thesis in communication and information technologies written by Mónica Domínguez Bajo at the University Pompeu Fabra under the supervision of Dr. Leo Wanner and la Dra. Mireia Farrús Cabecerán. The author was examined on Friday, 17th November 2017 by a committee formed by the doctors Bernd Möbius (University of Saarland), Pilar Prieto Vives (University Pompeu Fabra) y Catherine Lai (University of Edinburgh). It obtained the grade of Excellent.

**Keywords:** Information structure, communicative structure, thematicity, theme, rheme, prosody, acoustic parameters, speech synthesis, text-to-speech, concept-to-speech, automatic prosody labeling

## 1 *Introducción*

Las tecnologías del habla han pasado en poco tiempo de desarrollar tareas de lectura simples, como el sistema MITalk, a mantener conversaciones con interlocutores humanos, como es el caso de aplicaciones en el ámbito de la salud<sup>1</sup>. A pesar de que estas interacciones son relativamente sencillas, los asistentes virtuales están consiguiendo un impacto considerable en nuestra sociedad. No obstante, en lo que se conoce como tecnologías de ‘texto a habla’ (TTS, por sus siglas en inglés) aún no se han llegado a integrar aspectos comunicativos que doten a la generación de habla sintética de la versatilidad que existe en el lenguaje natural.

La expresividad de las voces sintéticas a través de la modelización de prosodia en los TTS tiene en cuenta, hasta cierto punto, algunos rasgos y funciones lingüísticas, sin embargo, aún no se ha llegado a alcanzar la riqueza que la prosodia tiene en el lenguaje humano. Por este motivo las voces sintéticas se siguen percibiendo como monótonas, sobre todo en discurso monologado con frases largas y complejas. Especialmente en el contexto de las tecnologías conversacionales, se espera que los agentes virtuales sean capaces tanto de expresarse de una manera apropiada al contexto como de mantener el interés del interlocutor. Esto, en la actualidad, supone un reto que requiere de una actualización en la agenda de investigación para incluir aspectos comunicativos que no se están teniendo en cuenta en tecnologías del habla.

<sup>1</sup>Entre otros, el avatar conversacional KRISTINA: <http://kristina-project.eu/en/>

Existe una amplia bibliografía en lingüística teórica que viene enfatizando: (i) que la prosodia desempeña un papel clave a la hora de expresar la intención comunicativa del hablante; (ii) que dicha intención comunicativa se articula en términos de la estructura informativa y (iii) que la estructura informativa se puede generar mediante un procedimiento computacional de organización del contenido semántico y sintáctico. A principios del siglo XXI se atisbaban modestos intentos de aplicar este conocimiento lingüístico en la práctica implementando conceptos básicos de la estructura informativa, principalmente, la segmentación en términos de tematicidad. La tematicidad describe cómo se empaqueta el contenido en función de “lo que se está hablando”, el *tema*, y “lo que se dice al respecto”, el *rema*. Los intentos de implementación de la tematicidad en TTS se basaron en una correspondencia básica entre tema-rema con patrones de entonación ascendentes-descendentes. Sin embargo, se estaban subestimando dos aspectos esenciales desde el punto de vista computacional: la asignación de la tematicidad dado un texto cualquiera y la generación de un abanico de rasgos prosódicos suficiente para aportar la variabilidad y expresividad necesarias.

El objetivo principal de esta tesis es demostrar empíricamente la viabilidad de una generación de prosodia en habla sintética que tenga en cuenta la intención comunicativa del mensaje. Con este fin, se propone una metodología para avanzar en el estudio de la interfaz estructura informativa-prosodia desde un punto de vista empírico para su aplicación en implementaciones de generación de habla sintética a partir de texto.

## 2 Organización de la tesis

La tesis se estructura en siete capítulos incluyendo introducción y conclusiones. En esta sección, se presenta un breve resumen de los capítulos centrales, es decir, del capítulo 2 al 6.

El capítulo 2 incluye una descripción de los conceptos fundamentales de las distintas áreas que abarca la tesis, en concreto:

- Estructura informativa o comunicativa: se explica la teoría de Igor Mel’čuk, que introduce una representación formal para aplicaciones computacionales de generación en el ámbito del procesamien-

to del lenguaje natural. De este modo, se enmarca el objeto de estudio de la presente tesis: la tematicidad jerárquica, tal y como la define Mel’čuk. Se detalla cómo esta teoría identifica tres segmentos sobre proposiciones: tema, rema y especificador. Dichos segmentos son recursivos y por lo tanto pueden definirse a distintos niveles de tematicidad.

- Prosodia: se introduce la convención de etiquetado ToBI<sup>2</sup>, así como los parámetros acústicos que se usan en los experimentos. También se explican los procedimientos y convenciones que se usan en los sintetizadores del estado del arte para generar y modificar la prosodia.

Se describen las estrategias de los TTS del estado del arte para derivar prosodia:

- Sistemas de reglas: emplean características lingüísticas de bajo nivel, por ejemplo, posición de las palabras, tipología (si es una palabra funcional o de contenido) y puntuación.
- Árboles de decisión: parten de información morfo-sintáctica, por ejemplo, si la palabra es un sustantivo o un verbo y qué palabras dependen de ella.
- Superposición de parámetros prosódicos: se emplean etiquetas usando un lenguaje de marcado para síntesis de habla, entre los más populares está la convención Speech Synthesis Markup Language (conocida por sus siglas en inglés, SSML)<sup>3</sup>. La manipulación de la prosodia basada en una simplificación de las etiquetas ToBI también se utiliza en algunos sintetizadores.

El capítulo 3 resume los estudios relacionados con la interfaz estructura informativa-prosodia desde el punto de vista de la lingüística computacional, aunque también se hace referencia a una selección de estudios teóricos. Por otro lado, se mencionan las herramientas informáticas existentes para anotar y analizar prosodia, sus ventajas y limitaciones.

En el capítulo 4, se explica la metodología propuesta, el corpus de trabajo y los procedimientos para anotar prosodia empleados

<sup>2</sup>Siglas en inglés correspondientes a *Tones and Breaks Indices*.

<sup>3</sup><https://www.w3.org/TR/speech-synthesis11/>

mediante la convención ToBI y extrayendo parámetros acústicos automáticamente.

Los capítulos 5 y 6 detallan los experimentos realizados en el marco de estudio de la tesis. El capítulo 5 presenta los experimentos de anotación de prosodia así como el sistema desarrollado para realizarlo automáticamente. El capítulo 6 incluye los experimentos de análisis de la correspondencia estructura informativa y prosodia a través de pruebas estadísticas y aprendizaje automático. Así mismo, se presenta la implementación en el entorno de TTS para testear las conclusiones extraídas del análisis del corpus de estudio.

### 3 Contribución de la tesis

La tesis contribuye al avance del estado del arte en la integración de la interfaz estructura informativa-prosodia en tecnologías del habla desde dos ámbitos: teórico y técnico.

#### 3.1 Contribución teórica

Los experimentos de análisis basado en un corpus de habla leída en inglés americano confirman que existe una correspondencia entre tematicidad jerárquica y prosodia que se puede modelizar para enriquecer la generación de prosodia.

En primer lugar se realiza un análisis estadístico de los parámetros prosódicos relacionados con los elementos que definen la prosodia, a saber, las pausas o fraseología, frecuencia fundamental (F0), la intensidad y la velocidad de habla. A continuación, se justifica por qué la tematicidad jerárquica propuesta por Mel'čuk es más adecuada que la segmentación binaria que se venía empleando en aplicaciones computacionales. Se realizan experimentos de clasificación para comprobar el potencial de predicción de prosodia a partir de tematicidad y viceversa mediante algoritmos de aprendizaje automático usando dos representaciones de prosodia: con etiquetas ToBI y con parámetros acústicos.

Finalmente, los experimentos en habla sintética muestran que los sintetizadores del estado del arte, tanto comerciales como de código libre, no tienen en cuenta la estructura comunicativa. Se muestra que el etiquetado ToBI, que es la convención más usada en estudios de prosodia, es muy limitado para su uso en enriquecimiento de prosodia en sintetizadores debido a su poca flexibilidad. Por este motivo, se propone el uso de la convención SSML, que permite una flexibilidad

mayor no solo en modificaciones de F0, sino también en intensidad, pausas y velocidad de habla.

#### 3.2 Contribución técnica

En esta tesis se han desarrollado dos aplicaciones de código abierto, que se detallan a continuación.

##### 3.2.1 Etiquetador automático de prosodia

El etiquetador automático de prosodia, está disponible en un servicio web, Praat on the Web<sup>4</sup>. El código es abierto y se distribuye bajo Licencia GNU v3<sup>5</sup>. La herramienta es modular y utiliza una extensión de Praat para el etiquetado automático de muestras de habla con o sin alineación por palabras. Se trata de un sistema basado en reglas que anota prominencia y fraseología prosódicas teniendo en cuenta parámetros acústicos. La evaluación de esta herramienta se realiza con muestras de habla leídas y espontáneas tanto en español como en inglés.

##### 3.2.2 Módulo de enriquecimiento prosódico

El módulo de enriquecimiento prosódico basado en la tematicidad jerárquica se puede usar con aplicaciones TTS que interpreten etiquetas SSML. Esta herramienta permite testear los resultados del análisis empírico del corpus sobre la correspondencia tematicidad jerárquica-prosodia en el entorno de habla sintética. Por lo tanto, se promueve un estudio aplicado de la interfaz estructura informativa-prosodia, lo cual supone un paso de la teoría lingüística en este área a la práctica en el contexto de tecnologías del habla, que hasta ahora no se había realizado.

#### Agradecimientos

La autora ha recibido financiación durante la tesis de las siguientes entidades: la Universidad Pompeu Fabra mediante una beca predoctoral del departamento de tecnologías de la información y las comunicaciones, el Ministerio de Economía y Competitividad mediante el programa de excelencia María de Maeztu (MDM-2015-0502) y la Comisión Europea a través del proyecto KRISTINA (H2020-RIA-645012).

<sup>4</sup><http://kristina.taln.upf.edu/praatweb/>

<sup>5</sup><https://github.com/monikaUPF>

**Bibliografía**

- Domínguez, M., A. Burga, M. Farrús, y L. Wanner. 2018. On the Role of Communicative Structure in Read Aloud Applications for the Elderly. En *Proceedings of the Workshop on Intelligent Conversation Agents in Home and Geriatric Care Applications, AAMAS'18*, Stockholm, Sweden.
- Domínguez, M., M. Farrús, A. Burga, y L. Wanner. 2014. The Information Structure–Prosody Language Interface Revisited. En *Proceedings of the 7th International Conference on Speech Prosody*, páginas 539–543, Dublin, Ireland.
- Domínguez, M., M. Farrús, A. Burga, y L. Wanner. 2016. Using hierarchical information structure for prosody prediction in content-to-speech applications. En *Proceedings of the 8th International Conference on Speech Prosody*, páginas 1019–1023, Boston, USA.
- Domínguez, M., M. Farrús, y L. Wanner. 2016a. An Automatic Prosody Tagger for Spontaneous Speech. En *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, páginas 377–387, Osaka, Japan.
- Domínguez, M., M. Farrús, y L. Wanner. 2016b. Combining acoustic and linguistic features in phrase-oriented prosody prediction. En *Proceedings of the 8th International Conference on Speech Prosody*, páginas 796–800, Boston, USA.
- Domínguez, M., M. Farrús, y L. Wanner. 2017. A thematicity-based prosody enrichment tool for cts. En *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017)*, páginas 3421–2, Stockholm, Sweden.
- Domínguez, M., M. Farrús, y L. Wanner. 2018a. Compilation of corpora to study the information structure–prosody interface. En *11th edition of the Language Resources and Evaluation Conference (LREC2018)*, páginas 4030–4035, Mijazaki, Japan.
- Domínguez, M., M. Farrús, y L. Wanner. 2018b. Thematicity-based Prosody Enrichment for Text-to-Speech Applications. En *Proceedings of the 9th International Conference on Speech Prosody 2018 (SP2018)*, páginas 612–616, Poznań, Poland.
- Domínguez, M., M. Farrús, y L. Wanner. 2018c. Towards Expressive Prosody Generation in TTS for Reading Aloud Applications. En *Proceedings of IberSpeech 2018: International Speech Communication Association*, páginas 40–44, Barcelona, Spain.
- Domínguez, M., I. Latorre, M. Farrús, J. Codina, y L. Wanner. 2016. Praat on the Web: An Upgrade of Praat for Semi-Automatic Speech Annotation. En *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, páginas 218–222, Osaka, Japan.
- Domínguez, M., M. Farrús, A. Burga, y L. Wanner. 2014. Towards Automatic Extraction of Prosodic Patterns for Speech Synthesis. En *Proceedings of the 7th International Conference on Speech Prosody*, páginas 1105–1109, Dublin, Ireland.

# Contribuciones a la comprensión lectora: mecanismos de atención y alineamiento entre n-gramas para similitud e inferencia interpretable \*

## *Contributions to language understanding: n-gram attention and alignments for interpretable similarity and inference*

Iñigo Lopez-Gazpio

Grupo Ixa, Universidad del País Vasco (UPV/EHU)

Manuel Lardizabal 1, 20018 Donostia

inigo.lopez@ehu.eus

**Resumen:** Tesis doctoral titulada “Hizkuntza-ulermenari Ekarpenak: N-gramen arteko Atentzio eta Lerrokatzeak Antzekotasun eta Inferentzia Interpretagarriak / Contribuciones a la Comprensión Lectora: Mecanismos de Atención y Alineamiento entre N-gramas para Similitud e Inferencia Interpretable”, defendida por Iñigo Lopez-Gazpio en la Universidad del País Vasco (UPV/EHU) y elaborada bajo la dirección de los doctores Eneko Agirre (Departamento de Lenguajes y Sistemas Informáticos) y Montse Maritxalar (Departamento de Lenguajes y Sistemas Informáticos). La defensa tuvo lugar el 30 de octubre de 2018 ante el tribunal formado por los doctores Kepa Sarasola (Presidente, Universidad del País Vasco (UPV/EHU)), Gorka Azkune (Secretario, Universidad de Deusto) y David Martínez (Vocal, IBM). La tesis obtuvo la calificación de sobresaliente Cum Laude y mención internacional.

**Palabras clave:** Similitud semántica, inferencia textual, redes neuronales, mecanismos de atención, alineación n-gramas, composicionalidad

**Abstract:** Ph. D. thesis entitled “Hizkuntza-ulermenari Ekarpenak: N-gramen arteko Atentzio eta Lerrokatzeak Antzekotasun eta Inferentzia Interpretagarriak / Contributions to Language Understanding: N-gram Attention and Alignments for Interpretable Similarity and Inference”, written by Iñigo Lopez-Gazpio at the University of Basque Country (UPV/EHU) under the supervision of Dr. Eneko Agirre (Languages and Computer Systems Department) and Dr. Montse Maritxalar (Languages and Computer Systems Department). The viva voce was held on October 30 2018 and the members of the commission were Dr. Kepa Sarasola (President, University of Basque Country (UPV/EHU)), Dr. Gorka Azkune (Secretary, University of Deusto) and Dr. David Martínez (Vocal, IBM). The thesis obtained Cum Laude excellent grade and international mention.

**Keywords:** Textual similarity, language inference, neural networks, attention mechanisms, n-gram alignments, compositionality

### 1 *Introducción de la tesis*

Esta tesis doctoral se ha realizado en el grupo Ixa de la Universidad del País Vasco (UPV/EHU) y presenta contribuciones relacionadas con la comprensión lectora de sistemas inteligentes por medio de las cuales estos sistemas incrementan su capacidad para entender el texto en entornos educativos. Principalmente trata la línea de investigación

de la composicionalidad de textos en vectores distribucionales, y la línea de investigación de la identificación e interpretación de similitudes y diferencias entre textos.

La sociedad actual muestra cierto interés por el aprendizaje continuo incluso en etapas avanzadas de la vida, y esto resulta en un creciente interés de cursos de diversas áreas de los cuales muchos se imparten *online*. Una gran ventaja de estos cursos en línea o del *e-learning* en general reside en su capacidad para favorecer la expansión y llegar a muchos es-

\*Esta tesis doctoral ha sido realizada con una beca predoctoral del Ministerio de Educación, Cultura y Deporte. Referencia: MINECO FPU13/00501.

tudiantes sin ninguna restricción geográfica. Como consecuencia de abarcar un espectro tan amplio de estudiantes potenciales es habitual que estos cursos lleguen a tener un número masivo de estudiantes. No en vano estos cursos son conocidos con el acrónimo *MOOC* del inglés *Massive Open Online Course*. El principal problema de los cursos MOOC es que los docentes de dichos cursos no son capaces de afrontar las necesidades individualizadas de los estudiantes inscritos en sus cursos, debido a su gran número. Como consecuencia de esta sobrecarga emplean evaluaciones tipo test para corregir las actividades de los estudiantes.

La motivación principal de esta tesis es desarrollar técnicas de procesamiento de lenguaje natural (*PLN*) con las que poder evaluar de forma automática a los estudiantes con respecto a una respuesta de referencia de un experto docente. Además, nuestra motivación es que los sistemas expertos de PLN sean capaces de identificar y relacionar segmentos entre el texto escrito por un estudiante y la respuesta de referencia, de forma que identifiquen explícitamente similitudes y diferencias entre ambos textos. Identificar estas relaciones es clave para poder producir retroalimentación en tiempo real con respecto a una respuesta de un estudiante que se desea evaluar.

## 2 Estructura de la tesis

La presente tesis tiene dos objetivos principales: 1) el desarrollo de sistemas inteligentes de PLN que sean capaces de evaluar respuestas de estudiantes contra respuestas de referencia de expertos docentes, y 2) que estos sistemas inteligentes sean capaces de producir retroalimentación útil para que los estudiantes puedan continuar su labor de aprendizaje.

Para organizar estos objetivos de forma secuencial hemos dividido la tesis en cinco secciones, la cual se presenta como compilación de artículos. En una primera sección introductoria presentamos la motivación, los objetos de estudio y las líneas de investigación que utilizaremos a lo largo de la tesis. Esta sección introductoria también contiene un resumen de todos los artículos relacionados. En la segunda sección nos centramos en realizar un análisis profundo del estado del arte con respecto a tecnologías del PLN, así como a analizar tareas y sistemas del ámbito educacional llevadas a cabo hasta la fecha

(Agirre et al., 2015a). Realizamos un énfasis especial en las arquitecturas basadas en redes neuronales y en las tareas de *Similitud Textual Semántica* (STS) (Cer et al., 2017) e *Inferencia Lógica* (NLI), ya que dichas arquitecturas y tareas forman la base para nuestro desarrollo de nuevos sistemas inteligentes.

En la tercera sección presentamos nuestro primer artículo (actualmente bajo revisión) que aborda el primer objetivo de la tesis: el desarrollo de sistemas inteligentes de PLN que sean capaces de evaluar un par de textos de entrada. Para el desarrollo del sistema inteligente analizamos diversas técnicas de modelado y representación de texto en vectores distribucionales, como sistemas basados en agrupaciones de palabras (*Bag-of-Words*), en redes neuronales recurrentes (*Recurrent Neural Networks*) y en redes convolucionales (*Convolutional Neural Networks*). En esta tercera sección proponemos una arquitectura novedosa en el estado del arte basada en redes neuronales capaz de modelar, representar y alinear n-gramas arbitrarios entre los textos de entrada. Si bien la alineación entre pares individuales de palabras es algo conocido y explotado con éxito en el estado del arte (Artetxe et al., 2018), la alineación entre n-gramas es una línea de investigación novedosa explorada en el marco de esta tesis.

En la cuarta sección presentamos nuestro segundo artículo (Lopez-Gazpio et al., 2017) que aborda el segundo objetivo de la tesis: explorar la capacidad de los sistemas inteligentes de PLN de forma que sean capaces de identificar las similitudes y diferencias entre un par de textos. De forma que esta capacidad adquirida en los sistemas permita generar retroalimentación útil a los estudiantes. Para implementar esta capacidad desarrollamos una nueva capa un nivel por encima de la Similitud Textual Semántica y la Inferencia Lógica, de forma que este nuevo nivel de anotación permite identificar y relacionar pares de agrupaciones de palabras. Llamamos Similitud Textual Semántica Interpretable (*interpretable STS* o *iSTS*) a esta nueva capa, y con ella es posible entrenar sistemas inteligentes de PLN para que sean capaces de reconocer de manera detallada las diferencias y similitudes entre textos. No sólo hemos diseñado la capa *iSTS* dentro del marco de esta tesis, sino que también hemos estado activos organizando dicha tarea en SemEval durante diversos años (Agirre et al., 2015b; Agirre et

al., 2016). Además hemos implementado distintos sistemas capaces de resolver la tarea (Agirre et al., 2015c; Lopez-Gazpio, Agirre, y Maritxalar, 2016) y finalmente sometido estos sistemas a evaluación con el objetivo de indagar si la retroalimentación es útil para los humanos en un entorno educacional.

Finalmente, en la quinta sección se presentan las contribuciones de la tesis divididas según los principales objetivos tratados y el trabajo futuro.

### 3 Contribuciones más relevantes

Para continuar presentamos en este apartado las contribuciones más relevantes divididas en dos apartados acorde con los principales objetivos enumerados en el contexto de la tesis.

#### 3.1 Modelos basados en la atención sobre n-gramas

En lo referente al desarrollo de modelos basados en la atención sobre n-gramas, hemos partido sobre la hipótesis inicial en la que considerábamos que la modelización de segmentos mayores que palabras individuales en vectores distribucionales y su respectiva alineación debería de ser superior a la modelización y alineamiento de palabras individuales.

Para llevar a cabo esta labor hemos partido de la implementación propia de un sistema de terceros descrito en el estado del arte basado en un modelo Bag-of-Words. Es decir, un sistema que en sí mismo es capaz de modelar interacción entre oraciones por medio de la interacción individual entre las palabras que conforman la oración. Tomando este sistema como *baseline* hemos realizado ciertas modificaciones para que sea capaz de modelar n-gramas en vectores distribucionales y también sea capaz de alinear dichas representaciones. También hemos diseñado otras dos variantes del modelo inicial que utilizan redes recurrentes y redes de convolución para extender la representación vectorial de las palabras, y poder evaluar nuestra propocición del modelo basado en n-gramas contra otras dos alternativas que utilizan sistemas supervisados complejos para modelar la composicionalidad.

Los resultados obtenidos en cinco conjuntos distintos de test de tareas relativas a STS (STS Benchmark y SICK-TS) y NLI (SNLI, MNLI y SICK-TE) muestran claramente la superioridad del modelo basado en n-gramas contra otras alternativas. Los resulta-

dos varían dependiendo del conjunto de test utilizado, en el que con respecto al sistema inicial hemos llegado a obtener una reducción del error relativo del 41 % en SICK-TS, del 38 % en STS Benchmark y del 29 % en SICK-TS. Con respecto a los conjuntos de test de SNLI y MNLI hemos obtenido reducciones del error relativo más limitadas en torno al 8 % y 11 %. También hemos observado que los modelos empleando redes neuronales recurrentes y redes convolucionales para extender la representación distribucional de las palabras obtienen mejores resultados que el sistema inicial, que no utiliza ningún mecanismo complejo de composicionalidad.

Con esta línea de investigación demostramos que el alineamiento entre n-gramas es útil de cara a la representación de oraciones, ya que es capaz de introducir contexto en la representación distribucional de los segmentos de la oración. Desde nuestro punto de vista modelar n-gramas es un paso intermedio entre los sistemas basados en agrupaciones de palabras (Bag-of-Words) y los sistemas basados en árboles de dependencias (Tree-RNN) que efectivamente son capaces de incorporar parte de la estructura sintáctica de las oraciones.

#### 3.2 Capacidad de interpretación entre textos

En lo referente al desarrollo de la capacidad de interpretabilidad ya hemos mencionado que nuestra principal aportación ha sido la de diseñar una capa encima de STS y NLI capaz de modelar explícitamente las similitudes y diferencias entre un par de oraciones. Para ello hemos diseñado una nueva tarea (iSTS) en la cual segmentamos las oraciones de entrada, y después realizamos alineamientos entre los segmentos identificados, estableciendo una etiqueta y un valor numérico para cada alineación. Con las etiquetas podemos especificar si un segmento es equivalente, similar, más o menos específico, contradictorio o está relacionado con otro segmento; y con el valor numérico podemos establecer la fuerza de esta etiqueta mediante un valor numérico dentro de una escala, teniendo  $valor \in [0, 5]$  Con esta anotación detallada los sistemas son capaces de aprender a identificar y diferenciar las relaciones de grano fino entre las oraciones, y producir retroalimentación útil a estudiantes. Diversos experimentos ponen de manifiesto un aumento en la correlación cuando

los humanos tenían a mano verbalizaciones producidas por sistemas expertos de PLN entrenados en iSTS.

### 3.3 Recursos

Dentro del marco de esta tesis se han creado y liberado una serie de recursos materiales y de software con el objetivo de aportar nuevas herramientas a la comunidad científica.

En relación con el primer objetivo de la tesis se han liberado sistemas basados en redes neuronales capaces de realizar las tareas de STS y NLI<sup>1</sup>.

En relación con el segundo objetivo de la tesis se han liberado diversos sistemas basados en aprendizaje automático y redes neuronales capaces de realizar la tarea de iSTS. Además, con respecto a la organización de la tarea en SemEval<sup>23</sup> también se han liberado herramientas para realizar la evaluación de sistemas, una aplicación de anotación de datos para facilitar la tarea de creación de nuevos conjuntos de datos, directrices para la anotación de datos, y un total de tres conjuntos de entrenamiento y otros tres conjuntos de test para entrenar nuevos sistemas en la tarea de iSTS. Los conjuntos de datos pertenecen al dominio de titulares de noticias, descripciones de imágenes y respuestas de estudiantes, entre los que suman más de 2500 pares de oraciones anotadas.

### Agradecimientos

Agradecemos el apoyo de la corporación NVIDIA por la donación de dos unidades de procesamiento gráfico utilizadas para esta investigación (Tesla K40 y Pascal Titan X).

### Bibliografía

Agirre, E., I. Aldabe, O. L. de Lacalle, I. Lopez-Gazpio, y M. Maritxalar. 2015a. Erantzunen kalifikazio automatikorako lehen urratsak. *EKAIA Euskal Herriko Unibertsitateko Zientzia eta Teknologia Aldizkaria*, (29).

Agirre, E., C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, I. Lopez-Gazpio, M. Maritxalar, R. Mihalcea, G. Rigau, L. Uria, y J. Wiebe. 2015b. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on

Interpretability. En *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, CO, June. Association for Computational Linguistics.

Agirre, E., A. Gonzalez-Agirre, I. Lopez-Gazpio, M. Maritxalar, G. Rigau, y L. Uria. 2015c. Ubc: Cubes for english semantic textual similarity and supervised approaches for interpretable sts. En *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, páginas 178–183, Denver, Colorado, June. Association for Computational Linguistics.

Agirre, E., A. Gonzalez-Agirre, I. Lopez-Gazpio, M. Maritxalar, G. Rigau, y L. Uria. 2016. Semeval-2016 task 2: Interpretable semantic textual similarity. En *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, páginas 512–524, San Diego, California, June. Association for Computational Linguistics.

Artetxe, M., G. Labaka, I. Lopez-Gazpio, y E. Agirre. 2018. Uncovering divergent linguistic information in word embeddings with lessons for intrinsic and extrinsic evaluation. En *Proceedings of the 22nd Conference on Computational Natural Language Learning*, páginas 282–291.

Cer, D., M. Diab, E. Agirre, I. Lopez-Gazpio, y L. Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. En *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, páginas 1–14. Association for Computational Linguistics.

Lopez-Gazpio, I., E. Agirre, y M. Maritxalar. 2016. iubc at semeval-2016 task 2: Rnns and lstms for interpretable sts. En *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, páginas 771–776, San Diego, California, June. Association for Computational Linguistics.

Lopez-Gazpio, I., M. Maritxalar, A. Gonzalez-Agirre, G. Rigau, L. Uria, y E. Agirre. 2017. Interpretable semantic textual similarity: Finding and explaining differences between sentences. *Knowledge-Based Systems*, 119:186–199.

<sup>1</sup><https://github.com/lgazpio>

<sup>2</sup><http://alt.qcri.org/semeval2015/task2/>

<sup>3</sup><http://alt.qcri.org/semeval2016/task1/>

# Similitud entre palabras: aportaciones de las técnicas basadas en bases de datos \*

## *Word similarity: contributions of knowledge-based methods*

Josu Goikoetxea Salutregi

Grupo Ixa, Universidad del País Vasco (UPV/EHU)

Manuel Lardizabal 1, 20018 Donostia

josu.goikoetxea@ehu.eus

**Resumen:** Tesis doctoral titulada “Hitzen arteko antzekotasuna: ezagutza-baseetan oinarritutako tekniken ekarpenak”, defendida por Josu Goikoetxea Salutregi en la Universidad del País Vasco (UPV/EHU) y elaborada bajo la dirección de los doctores Eneko Agirre (Departamento de Lenguajes y Sistemas Informáticos) y Aitor Soroa (Departamento de Ciencias de la Computación e Inteligencia Artificial). La defensa tuvo lugar el 13 de julio del 2018 ante el tribunal formado por los doctores Arantza Díaz de Ilarraza (Presidenta, Universidad del País Vasco (UPV/EHU)), Diego Lopez de Ipiña (Secretario, Universidad de Deusto) e Itziar Aduriz (Vocal, Universidad de Barcelona). La tesis obtuvo la calificación de sobresaliente.

**Palabras clave:** Similitud semántica, redes neuronales, embeddings, multilingüe, bases de datos, corpus de texto

**Abstract:** Ph.D. thesis entitled “Hitzen arteko antzekotasuna: ezagutza-baseetan oinarritutako tekniken ekarpenak”, written by Josu Goikoetxea Salutregi at the University of Basque Country (UPV/EHU) under the supervision of Dr. Eneko Agirre (Languages and Computer Systems Department) and Dr. Aitor Soroa (Computer Science and Artificial Intelligence). The viva voce was held on July 13 2018 and the members of the commission were Dr. Arantza Díaz de Ilarraza (President, University of Basque Country (UPV/EHU)), Dr. Diego Lopez de Ipiña (Secretary, University of Deusto) and Dr. Itziar Aduriz (Vocal, University of Barcelona). The thesis obtained excellent grade.

**Keywords:** Semantic similarity, neural networks, embeddings, cross-lingual, knowledge bases, text corpus

## 1 *Introducción de la tesis*

Esta tesis doctoral ha sido llevada a cabo en el grupo IXA de la Universidad del País Vasco (UPV/EHU) y su línea de investigación es la similitud semántica.

Desde la perspectiva de la psicología cognitiva, la similitud es una capacidad innata a los humanos, y sirve para estructurar la información proveniente de la realidad, para darla sentido. Tomando en cuenta esto último, es vital poder reproducir esa capacidad cognitiva en las máquinas si queremos que aprendan del mundo.

En concreto, la similitud se encuentra en

el núcleo de muchas tareas del procesamiento del lenguaje natural, siendo crucial su correcta reproducción. En esta tesis se ha trabajado con modelos distribucionales que se basan en el teoría de Harris y que calculan representaciones densas (en inglés *embeddings*) de significados de palabras. Siguiendo el teoría de Harris, los *embeddings* de palabras que comparten contextos parecidos van a ser similares, y, por tanto, la capacidad de reproducir la similitud es intrínseca a esas representaciones. De hecho, una de las evaluaciones más conocidas es la similitud entre palabras: cuanto más se acercan los resultados de similitud de un modelo computacional a los criterios humanos, mejor es la calidad de sus representaciones.

La motivación principal de este trabajo es

\*Esta tesis doctoral ha sido realizada con una beca predoctoral otorgada por el Vicerrectorado de euskera la Universidad Del País Vasco.

mejorar la calidad de los *embeddings*, y así, mejorar también los resultados en similitud entre palabras. Teniendo en cuenta que los *embeddings* sólo se basan en texto, y siguiendo la hipótesis de que la información semántica de métodos basados en corpus de texto y en bases de datos es complementaria, se han propuesto técnicas para combinar estas dos fuentes y crear representaciones híbridas de mayor calidad. Además, hemos extendido esos métodos al entorno bilingüe. Esta tesis consta de tres partes: 1) introducción de la tesis y precedentes de métodos basados en texto y bases de datos, 2) métodos propuestos para mejorar similitud semántica, y 3) conclusiones y trabajo futuro.

## 2 Estructura de la tesis

La primera parte se divide en los dos primeros capítulos. El primer capítulo está dedicado a definir la similitud semántica partiendo de una perspectiva cognitiva, para luego motivar el uso de la similitud en los modelos computacionales que calculan significados de palabras. Es importante mencionar la distinción que se hace entre similitud<sup>1</sup> y asociación<sup>2</sup>, ya que en el procesamiento del lenguaje según la tarea nos interesa una u otra. El segundo capítulo resume los precedentes de la tesis: primero, presenta las bases teóricas de la Semántica Distribucional y los modelos distribucionales actuales basados en texto; luego, describe las características de los métodos basados en bases de datos; a continuación, explica los métodos que aúnan información de los dos anteriores; después, resume los principales métodos para calcular representaciones multilingües; finalmente, termina explicando el método de evaluación de las representaciones de palabras, que se realizará calculando la correlación Spearman entre los resultados de similitud del modelo computacional y los patrones-oro creados por humanos.

La segunda parte consta de tres capítulos, que coinciden con las tres publicaciones principales de la tesis. El tercer capítulo describe nuestra propuesta (Goikoetxea, Soroa, y Agirre, 2015) para crear *embeddings* de lexicalizaciones de bases de datos. Este método tiene dos fases: en la primera, se crea un seudocorpus con las lexicalizaciones emitidas

<sup>1</sup>Incluye sinonimia, hiponimia e hiperonimia.

<sup>2</sup>Incluye, además de las de la similitud, meronimia, antonimia, asociaciones funcionales y otra serie de relaciones poco comunes.

en los caminos aleatorios llevados a cabo por la herramienta UKB<sup>3</sup>; en la segunda, se calculan los *embeddings* de las lexicalizaciones del seudocorpus mediante *word2vec*<sup>4</sup>. La base de datos que hemos usado es WordNet<sup>5</sup>, evaluando los *embeddings* resultantes en un patrón-oro de similitud y otro de asociación, y comparando los resultados con métodos en el estado del arte.

En el cuarto capítulo presentamos ocho combinaciones (Goikoetxea, Agirre, y Soroa, 2016) que fusionan la información complementaria de espacios semánticos separados de texto y bases de datos que dan como resultado representaciones híbridas. Las combinaciones se dividen en cuatro grupos: combinaciones de *embeddings*, de corpus, de resultados y las basadas en correlaciones. Así como en el capítulo anterior, los resultados de las combinaciones se evalúan en siete patrones-oro (tres de similitud y cuatro de asociación), y se comparan con resultados de varios métodos en el estado del arte.

En el quinto capítulo extendemos todo lo explicado anteriormente a espacios bilingües (Goikoetxea, Soroa, y Agirre, 2018), y presentamos un algoritmo para crear seudocorpus bilingües a través de caminos aleatorios bilingües, un método para crear corpus híbridos bilingües y una adaptación de la arquitectura Skip-gram de *word2vec* que introduce restricciones bilingües a la función-objetivo original. Al igual que en los dos capítulos previos, evaluamos las representaciones en patrones-oro bilingües y comparamos los resultados con las representaciones en el estado del arte.

El capítulo final resume las principales contribuciones de la tesis y el trabajo futuro. En los apéndices se describe de forma detallada la arquitectura original de Skip-gram y la extensión de Skip-gram con restricciones propuesta en el capítulo cuatro.

## 3 Contribuciones más relevantes

En los siguientes apartados describimos las contribuciones en nuestra línea de investigación, así como los recursos generados.

<sup>3</sup><http://ixa2.si.ehu.es/ukb/>

<sup>4</sup><https://code.google.com/archive/p/word2vec/>

<sup>5</sup>En concreto, la versión 3.0 con glosas. Esta versión también se ha usado en los dos siguientes capítulos.

### 3.1 Seudocorpus de WordNet

El seudocorpus de Wordnet se ha creado a partir de UKB, una colección de herramientas basadas en grafos. Este tipo de técnicas tratan las bases de datos como si fueran grafos y procesan la totalidad de la estructura del grafo. Se ha modificado UKB de forma que en su algoritmo de caminos aleatorios emita a un fichero las lexicalizaciones de los conceptos por los que pasa dicho camino. De esta manera, se codifica la información estructural de WordNet en las coocurrencias de un seudocorpus, abriendo la posibilidad de procesarlo modelo distribucional.

### 3.2 Embeddings de WordNet

Se han procesado el seudocorpus de WordNet mediante Skip-gram y CBOW y calculado luego los *embeddings* de todas las lexicalizaciones. Si comparamos esas representaciones de WordNet con las de los métodos basados en bases de datos de ese momento, nuestras representaciones son mucho más compactas y computacionalmente menos costosas.

Los mejores resultados se han obtenido calculando los *embeddings* de WordNet con Skip-gram, superando los resultados del patrón-oro de similitud SimLex999 con un Spearman de 0.52 e igualando los del patrón-oro de asociación WordSim353 con un Spearman de 0.683 a los vectores de UKB del estado del arte. Además, se ha propuesto una técnica para combinar los resultados de diferentes fuentes de información, superando del estado del arte del momento con un Spearman de 0.552 en SimLex999 mediante la combinación de *embeddings* de WordNet, texto y vectores UKB.

Unido con la aportación del seudocorpus de WordNet, hemos comprobado que efectivamente las coocurrencias de caminos aleatorios codifican la información relacional de WordNet. Debido a esto, los *embeddings* de WordNet son capaces de recoger la misma información estructural que los vectores de UKB, pero en un formato mucho más compacto y eficiente.

### 3.3 Representaciones híbridas

Basándonos en la complementariedad de la información semántica en corpus de texto y bases de datos, se han propuesto métodos eficientes para crear representaciones híbridas combinando espacios semánticos separados de los dos recursos mencionados. De los

ocho métodos propuestos, los mejores resultados son con la aplicación de PCA a la concatenación de *embeddings* de texto y de WordNet. Este última combinación introduce una ganancia absoluta media respecto a las correlaciones Spearman de *embeddings* de texto de 9.6 en los patrones-oro de similitud, 6.5 en los de asociación y 8.9 en general, superando con creces al método del estado del arte llamado *retrofitting*, que enriquece información *embeddings* de texto con información de WordNet. Además, se combinan hasta un total de seis recursos semánticos, superando con una combinación los seis recursos el estado del arte proveniente de diferentes métodos.

Por un lado, hemos demostrado que crear representaciones de significados a partir de espacios separados es más eficiente que crear representaciones que aprenden de forma conjunta desde esos espacios. Por otro, hemos comprobado que cuanto más recursos de diferente naturaleza semántica combinemos, mayor será la calidad de las representaciones.

### 3.4 Extensión bilingüe

En la última fase de la investigación se han extendido al entorno bilingüe las aportaciones descritas en los tres apartados anteriores. Hemos usado el inglés, el castellano, el euskera y el italiano, creando las siete combinaciones bilingües posibles en esos idiomas en tres tipos de corpus: texto, seudocorpus e híbrido. Además, hemos calculado todos esos *embeddings* usando tres métodos: el mapeo de espacios monolingües separados, procesamiento de corpus mediante Skip-gram original y con el Skip-gram que integra restricciones bilingües. Todos estos *embeddings* se han evaluado en tres patrones-oro bilingües creados automáticamente por nosotros. La combinación que mejores resultados ha obtenido es la de los *embeddings* calculados mediante el Skip-gram extendido procesando corpus híbridos bilingües, obteniendo resultados significativamente mejores que los *embeddings* obtenidos mediante el mapeo de espacios monolingües separados, e igualando a los vectores NASARI.

Las conclusiones obtenidas en los anteriores tres apartados se mantienen en espacios bilingües: a saber, que los seudocorpus bilingües de WordNet son capaces de codificar su información relacional, y que los *embeddings* de corpus híbridos bilingües son muy

eficientes en similitud. Además, hemos comprobado que la inserción de restricciones bilingües provenientes de WordNet mejoran todavía más los resultados en similitud.

### 3.5 Recursos

A lo largo de esta tesis se han hecho públicos para la comunidad científica los recursos descritos a continuación. En cuanto al software liberado, el algoritmo de caminos aleatorios para crear pseudocorpus (tanto monolingües como bilingües) está implementado en la versión 2.1 de UKB<sup>6</sup>, y la extensión de Skip-gram con restricciones se encuentra en el repositorio `github`<sup>7</sup>. Los *embeddings* monolingües<sup>8</sup> de WordNet y híbridos, así como todos bilingües<sup>9</sup> del apartado anterior también se han hecho públicos. En relación con esos últimos recursos bilingües, se puede acceder a todos los patrones-oro usados en el entorno bilingüe, incluyendo los dos patrones-oro monolingües de euskera creados expresamente para esa fase de la investigación.

### *Bibliografía*

- Goikoetxea, J., E. Agirre, y A. Soroa. 2016. Single or multiple? combining word representations independently learned from text and wordnet. En *AAAI*, páginas 2608–2614.
- Goikoetxea, J., A. Soroa, y E. Agirre. 2015. Random walks and neural network language models on knowledge bases. En *Proceedings of the 2015 Conference of NAACL/HLT*, páginas 1434–1439.
- Goikoetxea, J., A. Soroa, y E. Agirre. 2018. Bilingual embeddings with random walks over multilingual wordnets. *Knowledge-Based Systems*, 150:218–230.

<sup>6</sup>[http://ixa2.si.ehu.es/ukb/ukb\\_2.1.tgz](http://ixa2.si.ehu.es/ukb/ukb_2.1.tgz)

<sup>7</sup>[https://github.com/JosuGoiko/word2vec\\_constraints](https://github.com/JosuGoiko/word2vec_constraints)

<sup>8</sup><http://ixa2.si.ehu.es/ukb/>

<sup>9</sup>[http://ixa2.si.ehu.es/ukb/bilingual\\_embeddings.html](http://ixa2.si.ehu.es/ukb/bilingual_embeddings.html)

# Irony and Sarcasm Detection in Twitter: The Role of Affective Content

## *Detección de Ironía y Sarcasmo en Twitter: La Función del Contenido Afectivo*

Delia Irazú Hernández Farías

PRHLT Research Center, Universitat Politècnica de València, Spain  
Dipartimento di Informatica, University of Turin, Italy

Department of Computational Sciences  
National Institute of Astrophysics, Optics and Electronics  
Luis Enrique Erro 1, 72840, Santa María Tonantzintla, Puebla, Mexico  
dirazuherfa@inaoep.mx

**Resumen:** Tesis doctoral en Informática realizada por Delia Irazú Hernández Farías y dirigida por el Dr. Paolo Rosso (*Universitat Politècnica de València*) y la Dra. Viviana Patti (*University of Turin*) en el marco de un convenio de cotutela entre la Universitat Politècnica de València, España y la Universidad de Turin, Italia. La defensa de la tesis fue en Valencia, España el 25 de septiembre de 2017 ante un tribunal compuesto por: Horacio Saggion (*Universitat Pompeu Fabra*), Elisabetta Fersini (*Università degli Studi di Milano-Bicocca*) y Roberto Basili (*Univerità di Roma Tor Vergata*). Se obtuvo la mención internacional tras una estancia de 12 meses en la Universidad de Turin.

**Palabras clave:** Detección de ironía, detección de sarcasmo, procesamiento de lenguaje figurado, análisis de sentimientos

**Abstract:** PhD thesis in Computer Science written by Delia Irazú Hernández Farías under the supervision of Dr. Paolo Rosso (*Universitat Politècnica de València*) and Dra. Viviana Patti (*University of Turin*). This thesis was developed under a cotutelle between the Universitat Politècnica de València, Spain and the University of Turin, Italy. The thesis defense was done in Valencia, Spain on September 25, 2017. The doctoral committee was integrated by: Horacio Saggion (*Universitat Pompeu Fabra*), Elisabetta Fersini (*Università degli Studi di Milano-Bicocca*), and Roberto Basili (*Univerità di Roma Tor Vergata*). The International mention was achieved after a 12 months internship at University of Turin.

**Keywords:** Irony detection, sarcasm detection, figurative language processing, sentiment analysis

## 1 Introduction

People tend to use irony and sarcasm in social media to achieve different communication purposes. Dealing with such figurative language devices represents a big challenge for computational linguistics. The fuzzy separation between irony and sarcasm could lead to confusion. Commonly, the term “irony” is used as an umbrella term also covering sarcasm<sup>1</sup>.

Irony is closely associated with the expression of feelings, emotions, attitudes, and eva-

luations toward a particular target. It allows us to convey very subjective ideas and opinions in an indirect way, going beyond the literal meaning of the words. Therefore, since irony is considered as an affective manner of communication, taking advantage of affect-related information may help to identify ironic content in social media.

This thesis aimed at investigating the role of affect-related information in irony detection. We proposed to exploit affect-based information to characterize ironic content in Twitter. A complex and multifaceted phenomenon such as irony merits to be addressed

<sup>1</sup>In what follows we shall use “irony” in the same perspective.

not only considering broad aspects of affect (such as positive and negative sentiment) but also affective information in a finer-grained perspective (taking into account different models of emotions). This research was conducted paying special attention on three main aspects:

**I.** We analyzed the presence of different aspects of affect in ironic utterances in order to identify potential features to characterize such phenomenon in social media. We proposed a model, called *emotIDM*, which exploits an extensive set of resources covering different facets of affect ranging from sentiment to finer-grained emotions for characterizing ironic utterances. To evaluate our model, we collected a set of Twitter corpora used by scholars in previous research, to be used as benchmarks with a two-fold purpose: to compare the performance of our model against other approaches in the state of the art, and to evaluate its robustness across different aspects related to the characteristics of the corpora. Results show that *emotIDM* has a competitive performance across the experiments carried out, validating its effectiveness.

**II.** Aiming to investigate the differences between tweets labeled with *#irony*, *#sarcasm*, and *#not*, we analyzed different facets of affective information in instances containing such labels. We find data-driven arguments suggesting that the above mentioned hashtags are used to refer different figurative language devices. The results of our analysis allow us to probe distinctions and similarities between tweets labeled as *#irony* and *#sarcasm*. Our results allow us to contribute to the assumption that there is a separation between these figurative language devices. Furthermore, bearing in mind the importance of irony detection for sentiment analysis tasks, we investigated also the differences in polarity reversal terms of such tweets.

**III.** Detecting irony in user-generated content could have a broad range of applications. One of them is on sentiment analysis (SA). We analyzed the impact of irony in the performance of systems dedicated to SA, observing a drop when they deal with such figurative language devices. We proposed an irony-aware sentiment analysis system that incorporates *emotIDM* in a first stage before determining the overall polarity of a given Twitter message. The obtained results are competitive with the state of the art.

## 2 Thesis Overview

This thesis is comprised of a compendium of research articles already published (two international journal papers, an international conference paper, and a chapter in a book) as well as unpublished content. It consists of 8 chapters that are briefly introduced below.

Chapter 2 (Hernández Farías and Rosso, 2016) presents an overview of some state-of-the-art approaches to deal with irony and sarcasm detection in social media. Furthermore, an analysis of the performance of sentiment analysis systems on the presence of ironic content is presented. In Chapter 3 (Hernández Farías, Benedí, and Rosso, 2015), we described an irony detection model exploiting surface patterns as well as some features coming from sentiment analysis. The latter being the most relevant ones, thereby giving insights into the importance of such information for detecting irony in Twitter.

Chapter 4 (Hernández Farías, Patti, and Rosso, 2016) presents *emotIDM* which considers several aspects of affect ranging from sentiment to fine-grained models of emotions. The obtained results outperform the performance of the state-of-the-art approaches validating the importance of affect-related information for irony detection in Twitter. In Chapter 5 (Sulis et al., 2016), we analyzed different facets of affective information in tweets labeled with *#irony*, *#sarcasm* and *#not* with the aim of distinguishing between them. We also considered the role in terms of polarity reversal of tweets containing such hashtags.

Chapter 6 (Hernández Farías et al., 2017) describes an approach for performing sentiment analysis in tweets with figurative language. We proposed a pipeline that comprises two phases: first, we exploited *emotIDM* for identifying irony; then, by taking advantage of several affective resources we determine a polarity degree also considering the presence of ironic content. In Chapter 7, we presented the obtained results of some further experiments and analysis carried out with the aim to enhance the research work. Finally, in Chapter 8 we draw the main conclusions of the thesis, as well as the contributions and future work.

## 3 Contributions

Irony is a complex mode of communication closely related to the expression of feelings.

In this thesis, we introduced the problem of detecting irony and sarcasm in social media by considering mainly the subjective intrinsic value enclosed in ironic expressions. In this line, the following contributions were made within the development of the present research:

- *Overview of irony detection and its impact on SA.* We presented a brief description of the proposed approaches in the literature together with an analysis of shared tasks regarding SA where the participating systems were evaluated on the presence of figurative language devices. Overall, there is a drop in performance of the systems. It allows validating the assumption concerning to the importance of irony detection for determining the sentiment in a text.
- *emoIDM: an irony detection model.* We proposed a novel model for identifying irony in social media by taking advantage of different facets of affective information to capture ironic intention in Twitter. Unlike other research works, we did not build our own dataset, instead we collected a set of Twitter corpora previously used in the literature. The obtained results overall outperform those from the related work validating the importance of such kind of content for irony detection.
- *Analysis on the use of different hashtags related to ironic phenomena.* We investigated the use of different hashtags (#irony, #sarcasm, and #not), concluding that these labels are indeed used to tag different phenomena. Moreover, we explored the controversial subject to separate irony from sarcasm outperforming the state of the art. With respect to #not, it seems that it is used to represent a different figurative language device, although closer to #sarcasm than #irony. Furthermore, we investigated the behaviour of tweets labeled with such hashtags in terms of polarity reversal. It seems that in tweets labeled with #sarcasm often there is a full reversal (varying from a polarity to its opposite, almost always from positive to negative polarity), whereas in the case of those tagged with #irony there is an attenuation of the polarity (mostly from negative to neutral).
- *Development of an irony-aware sentiment analysis system.* Irony is a phenomenon

having an impact on the performance of systems dedicated to calculate the overall sentiment expressed in a given piece of text. We incorporated our irony detection model in a pipeline of sentiment analysis that relies mainly on sentiment and emotional-related resources. The obtained results were competitive and serve to validate the relevance of exploiting affective related features as well as the presence of irony for determining the sentiment in a given tweet.

### Acknowledgments

This work has been funded by the National Council for Science and Technology (CONACyT - Mexico) with the Grant No. 218109/313683. Part of the research was carried out in the framework of the SomEMBED TIN2015-71147-C2-1-P MINECO project.

### References

- Hernández Farías, D. I., C. Bosco, V. Patti, and P. Rosso. 2017. Sentiment polarity classification of figurative language: Exploring the role of irony-aware and multifaceted affect features. In *Computational Linguistics and Intelligent Text Processing - 18th International Conference, CICLing 2017, Revised Selected Papers, Part II*, pages 46–57.
- Hernández Farías, D. I., V. Patti, and P. Rosso. 2016. Irony Detection in Twitter: The Role of Affective Content. *ACM Trans. Internet Technol.*, 16(3):19:1–19:24.
- Hernández Farías, D. I. and P. Rosso. 2016. Irony, Sarcasm, and Sentiment Analysis. Chapter 7. In F. A. Pozzi, E. Fersini, E. Messina, and B. Liu, editors, *Sentiment Analysis in Social Networks*. Morgan Kaufmann, pages 113–127.
- Hernández Farías, I., J.-M. Benedí, and P. Rosso. 2015. Applying Basic Features from Sentiment Analysis for Automatic Irony Detection. In *Pattern Recognition and Image Analysis*, volume 9117 of *LNCS*, pages 337–344. Springer International Publishing.
- Sulis, E., D. I. Hernández Farías, P. Rosso, V. Patti, and G. Ruffo. 2016. Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not. *Knowledge-Based Systems*, 108:132–143.



# A Cross-domain and Cross-language Knowledge-based Representation of Text and its Meaning\*

*Una representación translingüe y transdominio del texto y su significado basada en el conocimiento*

Marc Franco-Salvador  
PRHLT Research Center  
Universitat Politècnica de València  
Camino de Vera s/n, 46022. Valencia, Spain

Symanto Research  
Pretzfelder Str. 15, 90425 Nürnberg  
marc.franco@symanto.net

**Abstract:** Ph.D. thesis (international doctorate mention) in Computer Science written by Marc Franco Salvador under the supervision of Dr. Paolo Rosso at the Universitat Politècnica de València. The author was examined in Valencia in May 2017 by a jury composed of the following doctors: Nicola Ferro (University of Padua), Bernardo Magnini (Fondazione Bruno Kessler), and Simone Paolo Ponzetto (University of Mannheim). The international doctorate mention was granted thanks to the completion of the following research internships: 1 year at the Sapienza University of Rome (Italy) under the supervision of Dr. Roberto Navigli, 2 months at the IIIT of Hyderabad and at Veooz (India) under the supervision of Dr. Vasudeva Varma and Dr. Prasad Pingali, 1 month at the INAOE (Mexico) under the supervision of Dr. Manuel Montes-y-Gómez, and 3 months at Symanto Group (Germany) under the supervision of Dr. Yassine Benajiba. The obtained grade was Excellent with *Cum Laude* distinction.

**Keywords:** Cross-language, cross-domain, knowledge graphs, plagiarism detection, information retrieval, text classification, sentiment analysis

**Resumen:** Tesis doctoral (con mención de doctorado internacional) en Informática realizada por Marc Franco Salvador bajo la supervisión del Dr. Paolo Rosso en la Universitat Politècnica de València. La lectura de la tesis fue realizada en Valencia en Mayo del 2017 por un jurado compuesto por los siguientes doctores: Nicola Ferro (University of Padua), Bernardo Magnini (Fondazione Bruno Kessler) y Simone Paolo Ponzetto (University of Mannheim). La mención de doctorado internacional fue otorgada gracias a la realización de las siguientes estancias de investigación: 1 año en la Sapienza University of Rome (Italia) bajo la supervisión del Dr. Roberto Navigli, 2 meses en el IIIT de Hyderabad y en Veooz (India) bajo la supervisión del Dr. Vasudeva Varma y el Dr. Prasad Pingali, 1 mes en el INAOE (México) bajo la supervisión del Dr. Manuel Montes-y-Gómez y 3 meses en Symanto Group (Alemania) bajo la supervisión del Dr. Yassine Benajiba. La calificación obtenida fue Sobresaliente con mención *Cum Laude*.

**Palabras clave:** Translingüe, transdominio, grafos de conocimiento, detección de plagio, recuperación de información, clasificación de texto, análisis del sentimiento

\* This research has been carried out in the framework of the European Commission project WIQ-EI IR-SES (no. 269180), and the national projects DIANA-APPLICATIONS - Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01), Destilado de opiniones desde contenidos generados por

usuarios (TIN2011-14726-E), and SomEMBED: Social Media language understanding - EMBEDding contexts (TIN2015-71147-C2-1-P).

## 1 Introduction

One of the most challenging aspects of Natural Language Processing (NLP) involves enabling computers to derive meaning from human natural language. To do so, several meaning or context representations have been proposed with competitive performance. However, these representations still have room for improvement when working in a cross-domain or cross-language scenario.

In this thesis we study the use of knowledge graphs<sup>1</sup> as a cross-domain and cross-language representation of text and its meaning. To do so, we generate the knowledge graphs with BabelNet,<sup>2</sup> the widest-coverage multilingual semantic network. This allows to have graphs that expand and relate the original concepts belonging to a set of words present in a text. This also provides with a language coverage of hundreds of languages and millions human-general and -specific concepts.

### 1.1 Motivation and Objectives

The use of the recent and popular distributed representations enables to accurately model text meaning and produced significant improvements in NLP tasks. However, there is a room of improvement in the cross-language scenario. Most of the approaches need high amounts of data in order to train representative models. In addition, the computational complexity and the amount of training data is proportional to the number of languages employed.

The use of knowledge graphs provided in the past with state-of-the-art results in the task of mono- and cross-language word sense disambiguation. The starting point of this work is the observation that, if these graphs provided with the correct disambiguations of a text, even at cross-language level, they are adequate as representation of the meaning of that text. In addition, we believe that BabelNet, the multilingual semantic network employed to generate the graphs, with a language coverage of hundreds of languages and millions concepts, makes this representation domain- and language-independent. Therefore, its complexity is also independent of

the number of languages employed. In consequence, knowledge graphs are adequate for cross-domain and cross-language NLP and Information Retrieval (IR) tasks. Moreover, knowledge graphs have several implicit characteristics (i.e., Word Sense Disambiguation (WSD), vocabulary expansion, and language independence) that have different impact on their performance in NLP similarity analysis tasks. Finally, we consider that this representation may be useful for other non-cross-language or non-cross-domain NLP tasks such as community questions answering, native language identification, and language variety identification.

Considering what aforementioned statements said, this thesis has the following objectives:

- To study the potential of knowledge graph-based features for cross-domain NLP tasks.
- To develop a cross-language similarity analysis model for NLP and IR tasks.
- To study the knowledge graph characteristics for cross-language similarity analysis tasks.
- To evaluate the performance of the developed approaches and compare them with the state-of-the-art models.
- To employ knowledge graphs for other NLP tasks.

## 2 Thesis Overview

This thesis has six chapters and is presented as a compendium of research articles which were published during the study phase of this PhD. We include two international journal articles and an international conference paper as chapters of this work. The thesis is structured as follows.

In Chapter 1 we present the introduction of this thesis. It includes the related work, motivation and objectives, our research questions, and the contributions of this research.

In Chapter 2 we present our journal article published in Franco-Salvador et al. (2015). In that work we employed knowledge graph-based features, such as WSD and vocabulary expansion-based ones (along with other traditional ones: bag of words and  $n$ -grams), for single- and cross-domain polarity classification. The evaluation includes a thorough

<sup>1</sup>A knowledge graph is a subset of a semantic network (also known as knowledge base) focused on the concepts belonging to a text, and the intermediate concepts and relations between them.

<sup>2</sup><http://babelnet.org>

analysis of the knowledge graph-based features and a comparison with the state of the art in domain adaptation.

In Chapter 3 we present our journal article published in Franco-Salvador, Rosso, and Montes y Gómez (2016). This is the reference work of our cross-language knowledge graph analysis model for cross-language similarity analysis. That method employs knowledge graphs as a cross-language representation of the text and its meaning. We also study the implicit and most relevant characteristics of the knowledge graphs at cross-language level, i.e., WSD, vocabulary expansion, and language independence. The evaluation includes the task of Spanish-English and German-English plagiarism detection and a comparison of the models in cases of plagiarism with paraphrasing.

In Chapter 4 we present our conference article published in Franco-Salvador, Rosso, and Navigli (2014). This publication presents a modified version of our cross-language knowledge graph analysis model. This also includes a vector component to cover shortcomings such as out-of-vocabulary words and verbal tenses. The evaluation in the tasks of cross-language document retrieval and categorization compares this new model with the state of the art using several language pairs.

In Chapter 5 we discuss the results that have been obtained on the previous chapters. Moreover, we complement our study with some further experiments to complete the picture at task level, and analyse the obtained results from a cross-domain and cross-language perspective. In addition, we present our experiments and results with knowledge graphs in other NLP tasks such as community questions answering, native language identification, and language variety identification.

In Chapter 6 we draw the main conclusions of the thesis, as well as its contributions and research lines for future work.

### 3 *Thesis Contributions*

The main contributions of this thesis are described below.

From the representation viewpoint, we proved that knowledge graphs can be employed as a cross-domain and cross-language representation of text and its meaning. We used several reference datasets to show diverse results and comparisons with the state of the

art and to justify the validity and potential of this representation. We supported all our conclusions with standard tests of statistical significance of results. In addition, we studied from a theoretical and practical perspective, the main characteristics that contribute to the knowledge graphs performance.

With respect to the tasks, we showed how to obtain state-of-the-art performance with knowledge graphs in several single- and cross-domain NLP and IR tasks: single- and cross-domain polarity classification (Franco-Salvador et al., 2015; Giménez-Pérez, Franco-Salvador, and Rosso, 2017), cross-language plagiarism detection (Franco-Salvador, Gupta, and Rosso, 2013; Franco-Salvador, Rosso, and Montes y Gómez, 2016; Franco-Salvador et al., 2016), document retrieval and categorization (Franco-Salvador, Rosso, and Navigli, 2014), and community questions answering. In addition, we showed the potentiality of knowledge graphs for native language identification (Franco-Salvador, Kondrak, and Rosso, 2017) and language variety identification (Rangel, Franco-Salvador, and Rosso, 2016).

From the modelling viewpoint, we employed knowledge graphs to obtain state-of-the-art performance in two different ways: (i) as a source of feature extraction for classification and regression, and (ii) as a representation, as part of the proposed cross-language similarity analysis models. With respect to these two models, we proposed one that employs knowledge graphs as representation of the text and its meaning, and we proposed another one that complements that representation with a vector-based representation to cover the graph shortcomings. In addition, we proposed a new embedding-based weighting scheme for the semantic relations between the knowledge graph concepts. This scheme proved to outperform the classical one employed in the BabelNet multilingual semantic network.

Finally, some contributions only partially related to knowledge graphs were achieved during this research. First, we proposed the continuous word alignment-based similarity analysis model that notably improved the performance of distributed representations of words in cross-language plagiarism detection. Next, we proved the relationship between the native language and the language variety identification tasks by solving both with the same approach wit-

hout any task-specific adaptation. The proposed string kernels-based approach obtained state-of-the-art performance in several datasets of the two tasks. Finally, we hypothesised that there is a relationship between knowledge graphs and distributed representations. We studied, with interesting results, how both complement each other for several NLP and IR tasks.

## References

- Franco-Salvador, M., F. L. Cruz, J. A. Troyano, and P. Rosso. 2015. Cross-domain polarity classification using a knowledge-enhanced meta-classifier. *Knowledge-Based Systems*, 86:46 – 56.
- Franco-Salvador, M., P. Gupta, and P. Rosso. 2013. Cross-language plagiarism detection using a multilingual semantic network. In *Proceedings of the 35th European Conference on Information Retrieval (ECIR'13)*, LNCS(7814), pages 710–713. Springer-Verlag.
- Franco-Salvador, M., P. Gupta, P. Rosso, and R. E. Banchs. 2016. Cross-language plagiarism detection over continuous-space and knowledge graph-based representations of language. *Knowledge-Based Systems*, 111:87–99.
- Franco-Salvador, M., G. Kondrak, and P. Rosso. 2017. Bridging the native language and the language variety identification tasks. In *Proceedings of the 21st International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES'17)*.
- Franco-Salvador, M., P. Rosso, and M. Montes y Gómez. 2016. A systematic study of knowledge graph analysis for cross-language plagiarism detection. *Information Processing & Management*, 52(4):550–570.
- Franco-Salvador, M., P. Rosso, and R. Navigli. 2014. A knowledge-based representation for cross-language document retrieval and categorization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL'14)*, pages 414–423. Association for Computational Linguistics.
- Giménez-Pérez, R. M., M. Franco-Salvador, and P. Rosso. 2017. Single and cross-domain polarity classification using string kernels. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL'17)*. Association for Computational Linguistics.
- Rangel, F., M. Franco-Salvador, and P. Rosso. 2016. A low dimensionality representation for language variety identification. In *Proceedings of the 17th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'16)*. Springer-Verlag.

# Cross-view Embeddings para la Recuperación de Información

## *Cross-view Embeddings for Information Retrieval*

Parth Gupta

Pattern Recognition Human Language Technology (PRHLT) Research Center  
 Universitat Politècnica de València  
 Camino de Vera s/n, 46022. Valencia, Spain  
 pgupta@dsic.upv.es

**Resumen:** Tesis doctoral en Informática realizada por Parth Gupta bajo la supervisión del Dr. Paolo Rosso (Universitat Politècnica de València) y el Dr. Rafael E. Banchs (Institute for Infocomm Research, Singapore). La tesis se defendió en Valencia (España) el 26 de enero de 2017. El comité de doctorado estuvo compuesto por los siguientes doctores: Eneko Agirre (Universidad del País Vasco), Julio Gonzalo (Universidad Nacional de Educación a Distancia) y Jaap Kamps (Universidad de Amsterdam). La tesis obtuvo la calificación de sobresaliente *Cum Laude*.

**Palabras clave:** Recuperación de información, multilingüe, aprendizaje profundo

**Abstract:** Ph.D. thesis in Computer Science written by Parth Gupta under the supervision of Dr. Paolo Rosso (Universitat Politècnica de València) and Dr. Rafael E. Banchs (Institute for Infocomm Research, Singapore). The thesis was defended in Valencia (Spain) on January 26, 2017. The doctoral committee comprised of the following doctors: Eneko Agirre (University of the Basque Country), Julio Gonzalo (Universidad Nacional de Educación a Distancia) and Jaap Kamps (University of Amsterdam). The thesis got the grade of outstanding *Cum Laude*.

**Keywords:** Information retrieval, cross-lingual, deep learning

## 1 Introduction

In this dissertation, we dealt with the cross-view tasks related to information retrieval using embedding methods. Paired instances of data which provide the same information about each datum in different modalities are referred to as cross-view data. For example, parallel sentences are two different views of a sentence in different languages. A word and its transliteration can be seen as two different views of the same word in different scripts. In cross-view tasks, instances of different views are not directly comparable. Under this terminology, CLIR and mixed-script information retrieval (MSIR) can be seen as cross-view retrieval tasks. Broadly, there are two approaches to cross-view tasks: (i) translation; and (ii) cross-view projection. In translation approaches, one-view is translated into the other view using a translation model and the retrieval is carried using the other view. While, in cross-view projection approaches, data in both views are projected to an abs-

tract common space using dimensionality reduction techniques, where they can be compared. Such representation is also referred to as embeddings. Though translation based approaches provide very rich representation of the data, such approaches are mainly devised for actual translation task such as machine translation (MT) of text from one language to the other. On the other hand, the projection methods provide a representation which may not be interpreted clearly, but provide more flexibility in obtaining representation pertaining to a particular task. For example, it is straight-forward to induce an objective function directly related to the task at hand in the learning mechanism *e.g.* increase cosine similarity between similar documents for a retrieval task. In this dissertation, we explore the cross-view embedding models for cross-view retrieval tasks.

We formally introduced the concept of mixed-script IR, which deals with the challenges faced by an IR system when a lan-

guage is written in different scripts because of various technological and sociological factors. Mixed-script terms are represented by a small and finite feature space comprised of character n-grams. We proposed the cross-view autoencoder (CAE) to model such terms in an abstract space and CAE provides the state-of-the-art performance.

We studied a wide variety of models for cross-language information retrieval (CLIR) and propose a model based on compositional neural networks (XCNN) which overcomes the limitations of the existing methods and achieves the best results for many CLIR tasks such as ad-hoc retrieval, parallel sentence retrieval and cross-language plagiarism detection.

We also explored an effective method to incorporate contextual similarity for lexical selection in machine translation. Concretely, we investigated a feature based on context available in source sentence calculated using deep autoencoders. The proposed feature exhibited statistically significant improvements over the strong baselines for English-to-Spanish and English-to-Hindi translation tasks.

Finally, we explored the methods to evaluate the quality of autoencoder generated representations of text data and analyse its architectural properties. For this, we proposed two metrics based on reconstruction capabilities of the autoencoders: structure preservation index (SPI) and similarity accumulation index (SAI). We also introduced a concept of critical bottleneck dimensionality (CBD) below which the structural information is lost and present analyses linking CBD and language perplexity.

## 2 Research Questions

In this dissertation, we concretely investigated the following research questions.

- RQ1 To what extent mixed-script IR is prevalent in web-search and what is the best way to model terms for it? [Chapter 5]
- RQ2 How effective is text representation obtained using external data composition neural network for cross-language IR applications? [Chapter 6]
- RQ3 How cross-view autoencoder is useful for lexical selection issue in machine translation? [Chapter 6]
- RQ4 How should the number of dimensions in the lowest-dimensional representation of a deep neural network autoencoder be chosen? [Chapter 7]

## 3 Thesis Overview

The dissertation is organised into four broad blocks: (i) we first introduce the background of the main topics of the thesis (Chapters 1, 2 & 3); (ii) we present the theoretical models proposed in this dissertation (Chapter 4); (iii) we present the evaluation results and analyses for the proposed models on cross-view tasks (Chapters 5 & 6); (iv) finally, we present analyses on structural properties for a proposed model (Chapter 7). More details about the organisation of each chapter is presented below.

Chapter 1 presents the introduction and motivation of the thesis. It also highlights the research questions investigated in the thesis along with contributions.

Chapter 2 discusses the theoretical background on information retrieval and dimensionality reduction. It also presents the main challenges and current state-of-the-art around these topics.

Chapter 3 presents necessary background on neural networks, Boltzmann machines, autoencoders and the optimisation methods to understand the technical details of the proposed models.

Chapter 4 presents the main technical contributions of the dissertation and explains the necessary details of the proposed models. We present the proposed cross-view autoencoder based framework to model mixed-script terms and the details of the external-data compositional neural network (XCNN) model.

Chapter 5 presents the details of the mixed-script information retrieval. We first formally define the problem of mixed-script information retrieval with research challenges. We further analyse the query logs of the Bing search engine to understand better the mixed-script queries and their distributions. Finally, we present extensive performance evaluation of the proposed model based on cross-view autoencoder on a standard collection along with other state-of-the-art methods and present insightful analyses.

Chapter 6 presents the evaluation results of the proposed models on cross-language information retrieval tasks such as CL ad-hoc

retrieval, parallel sentence retrieval, cross-language plagiarism detection and source context modelling for machine translation. For each application, we first give the description of the problem statement followed by the details of the existing methods. Finally, the comparative evaluation on standard benchmark collections is presented with necessary analysis.

In Chapter 7, we present two metrics, structure preservation index and similarity accumulation index. First, we define these metrics and present the underlying intuition capturing the different aspects of the autoencoder’s reconstruction capabilities. With the help of these metrics we define the notion of critical bottleneck dimensionality for the autoencoder. Finally, through the multilingual analysis on a parallel data we show that different languages have different critical bottleneck dimensionalities, which happens to be closely associated with the language grammatical complexities, measured in terms of n-gram perplexities.

Finally in Chapter 8, we draw the conclusions from the dissertation, discuss limitations and outline the future work.

## 4 Contributions

There are many facets of contributions in this dissertation. For the first time, we introduce the concept of MSIR formally. We also present the deep learning based cross-view models which provide the state-of-the-art performance for modelling mixed-script term equivalents for MSIR. The embedding based cross-view models: (i) cross-view autoencoder; and (ii) external-data compositional neural network (XCNN) provide state-of-the-art performance for many cross-view tasks such as cross-language ad-hoc IR, parallel sentence retrieval, cross-language plagiarism detection, source context features for machine translation and mixed-script IR. This dissertation also provides insightful information about the structural properties of the autoencoder architecture, which helps to analyse the training process in a more intuitive way. Here are more details on each of them.

### 4.1 Mixed-script information retrieval

Information retrieval in the mixed-script space, which can be termed as mixed-script IR, is challenging because queries written in either

the native or the Roman scripts need to be matched to the documents written in both scripts. Transliteration, especially into Roman script, is used abundantly on the web not only for documents, but also for user queries that intend to search for these documents. Since there are no standard ways of spelling a word in certain non-native scripts, transliterated content almost always features extensive spelling variations; typically a native term can be transliterated into Roman script in very many ways. For example, the word *pahala* (“first” in Hindi and many other Indian languages) can be written in Roman script as *pahalaa*, *pehla*, *pahila*, *pehlaa*, *pehala*, *pehalaa*, *pahela*, *pahlaa* and so on.

This phenomenon poses a non-trivial term matching problem for search engines to match the native-script or Roman-transliterated query with the documents in multiple scripts taking into account the spelling variations. The problem of MSIR, although prevalent in web search for users of many languages around the world, has received very little attention till date. MSIR presents challenges that the current approaches for solving mono-script spelling variation and NE transliteration in IR are unable to address adequately, especially because most of the transliterated queries (and documents) belong to the *long tail* of online search activity, and hence do not have enough click-through evidence to rely on.

### 4.2 Cross-view models

We present a principled solution to handle the mixed-script term matching and spelling variation where the terms across the scripts are modelled jointly. Although cross-view autoencoder provides a good way to model mixed-script equivalents, it has some limitations in modelling text. In contrast to the most of the existing models which rely only on the comparable/parallel data, our model (external-data compositional neural network – XCNN) takes the external relevance signals such as pseudo-relevance data to initialise the space monolingually and then, with the use of a small amount of parallel data, adjusts the parameters for different languages. There are a few approaches which go beyond the use of only parallel data. The framework also allows the use of click-through data, if available, instead of pseudo-relevance data. Our model, differently from other models, optimises an

objective function that is directly related to an evaluation metric for retrieval tasks such as cosine similarity. These two properties prove crucial for XCNN to outperform existing techniques in the cross-language IR setting. We test XCNN on different tasks of CLIR and it attains the best performance in comparison to a number of strong baselines including machine translation based models.

### 4.3 Critical bottleneck dimensionality

Although deep learning techniques are in vogue, there still exist some important open questions. In most of the studies involving the use of these techniques for dimensionality reduction, the qualitative analysis of projections is never presented. Typically, the reliability of the autoencoder is estimated based on its reconstruction capability.

The dissertation proposed a novel framework for evaluating the quality of the dimensionality reduction task based on the merits of the application under consideration: the representation of text data in low dimensional spaces. Concretely, the framework is comprised of two metrics, structure preservation index (SPI) and similarity accumulation index (SAI), which capture two different aspects of the autoencoder’s reconstruction capability. More specifically, these two metrics focus on assessing the structural distortion and the similarities among the reconstructed vectors, respectively. In this way, the framework gives better insight of the autoencoder performance allowing for conducting better error analysis and evaluation. With the help of these metrics, we also define the concept of critical bottleneck dimensionality which refers to the adequate size of the bottleneck layer of an autoencoder.

## 5 Conclusions and Future Work

This dissertation deals with cross-view projection techniques for cross-view information retrieval tasks. In the exploration, a very important and prevalent problem of mixed-script IR is formally defined and investigated. The deep learning based neural cross-view models proposed in this dissertation provide state-of-the-art performance for various cross-language and cross-script applications. The dissertation also explored the architectural properties of the autoencoders which has attained less attention and establishes the no-

tion of critical bottleneck dimensionality. Some of the most important publications from the thesis work are listed in the References below.

### References

- Barrón-Cedeño, A., P. Gupta, and P. Rosso. 2013. Methods for cross-language plagiarism detection. *Knowl.-Based Syst.*, 50:211–217.
- Franco-Salvador, M., P. Gupta, P. Rosso, and R. E. Banchs. 2016. Cross-language plagiarism detection over continuous-space and knowledge graph-based representations of language. *Knowl.-Based Syst.*, 111:87–99.
- Gupta, P., K. Bali, R. E. Banchs, M. Choudhury, and P. Rosso. 2014. Query expansion for mixed-script information retrieval. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '14*, pages 677–686, New York, NY, USA. ACM.
- Gupta, P., R. E. Banchs, and P. Rosso. 2016a. Continuous space models for CLIR. *Information Processing & Management*, 53(2):359–370.
- Gupta, P., R. E. Banchs, and P. Rosso. 2016b. Squeezing bottlenecks: Exploring the limits of autoencoder semantic representation capabilities. *Neurocomputing*, 175:1001–1008.
- Gupta, P., M. R. Costa-Jussà, P. Rosso, and R. E. Banchs. 2016. A deep source-context feature for lexical selection in statistical machine translation. *Pattern Recognition Letters*, 75:24–29.

# *Información General*



# SEPLN 2019

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

Universidad del País Vasco – Bilbao (España)  
25-27 de septiembre 2019  
<http://www.sepln.org/> y <http://hitz.eus/sepln2019/>

### **1 Presentación**

La XXXV edición del Congreso Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) se celebrará los días 25, 26 y 27 de septiembre de 2019 en el Bizkaia Aretoa.

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

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

### **2 Objetivos**

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

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

### **3 Áreas Temáticas**

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

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

- Generación de Resúmenes.
- PLN para la generación de recursos educativos.
- PLN para lenguas con recursos limitados.
- Aplicaciones industriales del PLN.

#### **4 Formato del Congreso**

La duración prevista del congreso será de tres días, con sesiones dedicadas a la presentación de artículos, pósteres, proyectos de investigación en marcha y demostraciones de aplicaciones. Además, prevemos la organización de talleres-workshops satélites para el día 24 de septiembre.

#### **5 Comité ejecutivo SEPLN 2019**

Presidente del Comité Organizador

- Inma Hernáez (Universidad del País Vasco).

Colaboradores

- Eva Navas (Universidad del País Vasco).
- Ibon Saratxaga (Universidad del País Vasco).
- Jon Sanchez (Universidad del País Vasco).
- Koldo Gojenola (Universidad del País Vasco).

#### **6 Consejo Asesor**

Miembros:

- Manuel de Buenaga Rodríguez (Universidad Europea de Madrid, España)
- Sylviane Cardey-Greenfield (Centre de recherche en linguistique et traitement automatique des langues, Lucien Tesnière. Besançon, Francia)
- Irene Castellón Masalles (Universidad de Barcelona, España)
- Arantza Díaz de Ilaraza (Universidad del País Vasco, España)
- Antonio Ferrández Rodríguez (Universidad de Alicante, España)
- Alexander Gelbukh (Instituto Politécnico Nacional, México)
- Koldo Gojenola Gallettebeitia (Universidad del País Vasco, España)
- Xavier Gómez Guinovart (Universidad de Vigo, España)
- José Miguel Goñi Menoyo (Universidad Politécnica de Madrid, España)
- Ramón López-Cózar Delgado (Universidad de Granada, España)
- Bernardo Magnini (Fondazione Bruno Kessler, Italia)

- Nuno J. Mamede (Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa, Portugal)
- M. Antònia Martí Antonín (Universidad de Barcelona, España)
- M. Teresa Martín Valdivia (Universidad de Jaén, España)
- Patricio Martínez Barco (Universidad de Alicante, España)
- Eugenio Martínez Cámara (Universidad de Granada, España)
- Paloma Martínez Fernández (Universidad Carlos III, España)
- Raquel Martínez Unanue (Universidad Nacional de Educación a Distancia, España)
- Ruslan Mitkov (University of Wolverhampton, Reino Unido)
- Manuel Montes y Gómez (Instituto Nacional de Astrofísica, Óptica y Electrónica, México)
- Lluís Padró Cirera (Universidad Politécnica de Cataluña, España)
- Manuel Palomar Sanz (Universidad de Alicante, España)
- Ferrán Pla (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 (Universidad Politécnica de Valencia, España)
- Kepa Sarasola Gabiola (Universidad del País Vasco, España)
- Encarna Segarra Soriano (Universidad Politécnica de Valencia, España)
- Tamar Solorio (University of Houston, Estados Unidos de América)
- Maite Taboada (Simon Fraser University, Canadá)
- Mariona Taulé (Universidad de Barcelona, España)
- Juan-Manuel Torres-Moreno (Laboratoire Informatique d'Avignon / Université d'Avignon, Francia)
- José Antonio Troyano Jiménez (Universidad de Sevilla, España)
- L. Alfonso Ureña López (Universidad de Jaén, España)

- Rafael Valencia García (Universidad de Murcia, España)
- René Venegas Velásques (Pontificia Universidad Católica de Valparaíso, Chile)
- M. Felisa Verdejo Maíllo (Universidad Nacional de Educación a Distancia, España)
- Manuel Vilares Ferro (Universidad de la Coruña, España)
- Luis Villaseñor-Pineda (Instituto Nacional de Astrofísica, Óptica y Electrónica, México)

### ***7 Fechas importantes***

- Fechas para la presentación y aceptación de comunicaciones:
- Fecha límite para la entrega de comunicaciones: 21 de marzo de 2019.
- Notificación de aceptación: 16 de mayo de 2019.
- Fecha límite para entrega de la versión definitiva: 31 de mayo de 2019.



# Información Adicional

## Funciones del Consejo de Redacción

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

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

El consejo de redacción está formado por los siguientes miembros

L. Alfonso Ureña López (Director)

Universidad de Jaén

laurena@ujaen.es

Patricio Martínez Barco (Secretario)

Universidad de Alicante

patricio@dlsi.ua.es

Manuel Palomar Sanz

Universidad de Alicante

mpalomar@dlsi.ua.es

Felisa Verdejo Mañllo

UNED

felisa@lsi.uned.es

## Funciones del Consejo Asesor

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

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

El Consejo Asesor está formado por los siguientes miembros:

Manuel de Buenaga

Universidad Europea de Madrid (España)

Sylviane Cardey-Greenfield

Centre de recherche en linguistique et traitement automatique des langues (Francia)

Irene Castellón

Universidad de Barcelona (España)

Arantza Díaz de Ilarraza

Universidad del País Vasco (España)

Antonio Ferrández

Universidad de Alicante (España)

Alexander Gelbukh

Instituto Politécnico Nacional (México)

Koldo Gojenola

Universidad del País Vasco (España)

Xavier Gómez Guinovart

Universidad de Vigo (España)

José Miguel Goñi

Universidad Politécnica de Madrid (España)

Ramón López-Cózar Delgado

Universidad de Granada (España)

Bernardo Magnini

Fondazione Bruno Kessler (Italia)

Nuno J. Mamede

Instituto de Engenharia de Sistemas e Computadores (Portugal)

M. Antònia Martí Antonín

Universidad de Barcelona (España)

M. Teresa Martín Valdivia

Universidad de Jaén (España)

Patricio Martínez-Barco

Universidad de Alicante (España)

Eugenio Martínez Cámara

Universidad de Granada (España)

Paloma Martínez Fernández

Universidad Carlos III (España)

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

## Cartas al director

Sociedad Española para el Procesamiento del Lenguaje Natural  
 Departamento de Informática. Universidad de Jaén  
 Campus Las Lagunillas, Edificio A3. Despacho 127. 23071 Jaén  
[secretaria.sepln@ujaen.es](mailto: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/](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/>