Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1989-7553

Artículos

Bertinho: Galician BERT Representations
David Vilares, Marcos Garcia, Carlos Gómez-Rodríguez 13
Evaluación de un modelo transformador aplicado a la tarea de generación de resúmenes en distintos dominios
Isabel Segura-Bedmar, Lucía Ruz, Sara Guerrero-Aspizua 27
NECOS: An annotated corpus to identify constructive news comments in Spanish
Pilar López-Úbeda, Flor Miriam Plaza-del-Arco, Manuel Carlos Díaz-Galiano, M.Teresa Martín-Valdivia 41
Using Dependency-Based Contextualization for transferring Passive Constructions from English to Spanish
Pablo Gamallo, Gorka Labaka 53
Categorizing Misogynistic Behaviours in Italian, English and Spanish Tweets
Silvia Lazzardi, Viviana Patti, Paolo Rosso 65
Escansión automática de poesía española sin silabación
Guillermo Marco Remón, Julio Gonzalo 77
Using Guarani Verbal Morphology on Guarani-Spanish Machine Translation Experiments
Yanina Borges, Florencia Mercant, Luis Chiruzzo 89
A morphological analyser for K'iche'
Ivy Richardson, Francis M. Tyers 99
Consumer Cynicism Identification for Spanish Reviews using a Spanish Transformer Model
Samuel González-López, Steven Bethard, Francisca Cecilia Encinas Orozco, Adrián Pastor López-Monroy 111
Grammatical error correction for Spanish health records
Salvador Lima-López, Naiara Perez, Montse Cuadros 121
Mejoras aplicadas a la extracción de relaciones semánticas para la Web en español
Juan M. Rodríguez, Hernán D. Merlino, Patricia Pesado 133
Computational Reproducibility of Named Entity Recognition methods in the biomedical domain
Ana Garcia-Serrano, Sebastian Hennig, Andreas Nürnberger 141
Classifying Spanish 'se' constructions: from bag of words to language models
Nuria Aldama García, Álvaro Barbero Jiménez 153
Generación automática de resúmenes de referencia para la evaluación del manejo de estructuras discursivas y coherencia en el alumnado
Unai Atutxa, Alejandro Molina-Villegas, Mikel Iruskieta 165
Discovering topics in Twitter about the COVID-19 outbreak in Spain
Marvin M. Agüero-Torales, David Vilares, Antonio G. López-Herrera 177

Tesis

Negation Processing in Spanish and its Application to Sentiment Analysis
Salud María Jiménez-Zafra 193
Document-Level Machine Translation – EnsuringTranslational Consistency of Non-Local Phenomena
Eva Martínez Garcia 197
Hacía el análisis de sentimientos en euskera
Jon Alkorta Agirrezabala 201
Language and Structure in Polarized Communities
Mirko Lai 205
Buscando robustez en un mundo multilingüe: de pipelines a embeddings
Yerai Doval 209

Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1989-7553

Comité Editorial

Consejo de redacción

ISSN: 1135-5948
ISSN electrónico: 1989-7553
Depósito Legal: B:3941-91
Editado en: Universidad de Jaén
Año de edición: 2021

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

Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1989-7553

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 66 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 37 trabajos para este número, de los cuales 32 eran artículos científicos y 5 resúmenes de tesis. De entre los 32 artículos recibidos, 15 han sido finalmente seleccionados para su publicación, lo cual fija una tasa de aceptación del 46,88%.

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 2021
Los editores.

Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1989-7553

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 66th 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.

Thirty-seven papers were submitted for this issue, from which thirty-two were scientific papers and five doctoral dissertation summaries. From these thirty-two papers, we selected fifteen papers (46.88%) 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 2021
Editorial board.

Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1989-7553

Artículos

Bertinho: Galician BERT Representations
David Vilares, Marcos Garcia, Carlos Gómez-Rodríguez 13
Evaluación de un modelo transformador aplicado a la tarea de generación de resúmenes en distintos dominios
Isabel Segura-Bedmar, Lucía Ruz, Sara Guerrero-Aspizua 27
NECOS: An annotated corpus to identify constructive news comments in Spanish
Pilar López-Úbeda, Flor Miriam Plaza-del-Arco, Manuel Carlos Díaz-Galiano, M.Teresa Martín-Valdivia 41
Using Dependency-Based Contextualization for transferring Passive Constructions from English to Spanish
Pablo Gamallo, Gorka Labaka 53
Categorizing Misogynistic Behaviours in Italian, English and Spanish Tweets
Silvia Lazzardi, Viviana Patti, Paolo Rosso 65
Escansión automática de poesía española sin silabación
Guillermo Marco Remón, Julio Gonzalo 77
Using Guarani Verbal Morphology on Guarani-Spanish Machine Translation Experiments
Yanina Borges, Florencia Mercant, Luis Chiruzzo 89
A morphological analyser for K'iche'
Ivy Richardson, Francis M. Tyers 99
Consumer Cynicism Identification for Spanish Reviews using a Spanish Transformer Model
Samuel González-López, Steven Bethard, Francisca Cecilia Encinas Orozco, Adrián Pastor López-Monroy 111
Grammatical error correction for Spanish health records
Salvador Lima-López, Naiara Perez, Montse Cuadros 121
Mejoras aplicadas a la extracción de relaciones semánticas para la Web en español
Juan M. Rodríguez, Hernán D. Merlino, Patricia Pesado 133
Computational Reproducibility of Named Entity Recognition methods in the biomedical domain
Ana Garcia-Serrano, Sebastian Hennig, Andreas Nürnberger 141
Classifying Spanish 'se' constructions: from bag of words to language models
Nuria Aldama García, Álvaro Barbero Jiménez 153
Generación automática de resúmenes de referencia para la evaluación del manejo de estructuras discursivas y coherencia en el alumnado
Unai Atutxa, Alejandro Molina-Villegas, Mikel Iruskieta 165
Discovering topics in Twitter about the COVID-19 outbreak in Spain
Marvin M. Agüero-Torales, David Vilares, Antonio G. López-Herrera 177

Tesis

Negation Processing in Spanish and its Application to Sentiment Analysis
Salud María Jiménez-Zafra 193
Document-Level Machine Translation – EnsuringTranslational Consistency of Non-Local Phenomena
Eva Martínez Garcia 197
Hacía el análisis de sentimientos en euskera
Jon Alkorta Agirrezabala 201
Language and Structure in Polarized Communities
Mirko Lai 205
Buscando robustez en un mundo multilingüe: de pipelines a embeddings
Yerai Doval 209

Información General

XXXVII Congreso Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural 219
Información para los autores 219
Información adicional 221

Artículos

Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 13-26 recibido 27-11-2020 revisado 20-01-2021 aceptado 30-01-2021

Bertinho: Galician BERT Representations

Bertinho: Representaciones BERT para el gallego

David Vilares,[1] **Marcos Garcia,**[2] **Carlos Gómez-Rodríguez** [1]
[1]Universidade da Coruña, CITIC, Galicia, Spain
[2]CiTIUS, Universidade de Santiago de Compostela, Galicia, Spain
david.vilares@udc.es, marcos.garcia.gonzalez@usc.gal, carlos.gomez@udc.es

Abstract: This paper presents a monolingual BERT model for Galician. We follow the recent trend that shows that it is feasible to build robust monolingual BERT models even for relatively low-resource languages, while performing better than the well-known official multilingual BERT (mBERT). More particularly, we release two monolingual Galician BERT models, built using 6 and 12 transformer layers, respectively; trained with limited resources (∼45 million tokens on a single GPU of 24GB). We then provide an exhaustive evaluation on a number of tasks such as POS-tagging, dependency parsing and named entity recognition. For this purpose, all these tasks are cast in a pure sequence labeling setup in order to run BERT without the need to include any additional layers on top of it (we only use an output classification layer to map the contextualized representations into the predicted label). The experiments show that our models, especially the 12-layer one, outperform the results of mBERT in most tasks.
Keywords: BERT, Galician, embeddings, language modeling.

Resumen: Este artículo presenta un modelo BERT monolingüe para el gallego. Nos basamos en la tendencia actual que ha demostrado que es posible crear modelos BERT monolingües robustos incluso para aquellos idiomas para los que hay una relativa escasez de recursos, funcionando éstos mejor que el modelo BERT multilingüe oficial (mBERT). Concretamente, liberamos dos modelos monolingües para el gallego, creados con 6 y 12 capas de *transformers*, respectivamente, y entrenados con una limitada cantidad de recursos (∼45 millones de palabras sobre una única GPU de 24GB.) Para evaluarlos realizamos un conjunto exhaustivo de experimentos en tareas como análisis morfosintáctico, análisis sintáctico de dependencias o reconocimiento de entidades. Para ello, abordamos estas tareas como etiquetado de secuencias, con el objetivo de ejecutar los modelos BERT sin la necesidad de incluir ninguna capa adicional (únicamente se añade la capa de salida encargada de transformar las representaciones contextualizadas en la etiqueta predicha). Los experimentos muestran que nuestros modelos, especialmente el de 12 capas, mejoran los resultados de mBERT en la mayor parte de las tareas.
Palabras clave: BERT, gallego, *embeddings*, modelado del lenguaje.

1 Introduction

Contextualized word representations (Dai and Le, 2015; Peters et al., 2018; Devlin et al., 2019) have largely improved the performance of many natural language processing (NLP) tasks, such as syntactic parsing (Kitaev and Klein, 2018), question answering (Salant and Berant, 2018) or natural language inference (Jiang and de Marneffe, 2019), among many others. Contrary to static word embeddings (Mikolov et al., 2013b; Pennington, Socher, and Manning, 2014), where a given term is always represented by the same low-dimensional vector, contextualized approaches encode each word based on its context. Such process is normally learned by a neural network that optimizes a language modeling objective.

One of the most popularized and best performing models to generate contextualized representations is BERT (Devlin et al., 2019), a bidirectional language model based on transformers (Vaswani et al., 2017). BERT was initially released as a monolingual model

ISSN 1989-7553. DOI 10.26342/2021-66-1

for English, with *large* and *base* variants, made of 24 and 12 transformer layers, respectively. In addition, a multilingual BERT version (mBERT) trained on the one hundred most popular languages on Wikipedia was also released. Although mBERT has become a very popular and easy-to-use tool to address multilingual NLP challenges (Pires, Schlinger, and Garrette, 2019; Karthikeyan et al., 2020), some authors have reported that its performance is not so robust as that of the corresponding monolingual models (Wu and Dredze, 2020; Vulić et al., 2020). In this line, previous work has showed that training a monolingual BERT is worth in terms of performance, in comparison to mBERT. Among others, this is the case for languages coming from different typologies, languages families and scripts, such as Finnish (Virtanen et al., 2019), Basque (Agerri et al., 2020), Spanish (Cañete et al., 2020), Greek (Koutsikakis et al., 2020) or Korean (Lee et al., 2020).

Taking the above into account, this paper contributes with the development of BERT models for Galician, a relatively low-resource language for which, to best of our knowledge, there is no contextualized (monolingual) model available. In this regard, we train two *Bertinho* models and test their performance on several tasks. Specifically, we assess the effect of the number of layers when using limited data (less than 45 million tokens). To do so, we train models with 6 and 12 layers using a single TESLA P40 24GB GPU and compare them against the official multilingual BERT on a number of downstream tasks, including POS-tagging, dependency parsing, and named entity recognition (NER). The experiments show that the monolingual models clearly outperform mBERT: even a small 6-layer model outperforms the official multilingual one in most scenarios, and the 12-layer one obtains the overall best results. We have submitted *Bertinho* to the HuggingFace Models Hub[1]. Contemporaneously to this work, HuggingFace has released a Galician RoBERTa (Liu et al., 2019) model[2] based on the approach presented by Ortiz Suárez, Romary, and Sagot (2020).

[1] `https://huggingface.co/dvilares/bertinho-gl-base-cased` and `https://huggingface.co/dvilares/bertinho-gl-small-cased`

[2] `https://huggingface.co/mrm8488/RoBERTinha`

Apart from this introduction, this paper is organized as follows: Section 2 introduces some related work regarding static and contextualized vector representations for NLP, discussing different approaches to train new monolingual models. Then, we describe the particularities of the Galician BERT-based models in Section 3. Finally, the experiments and results are presented and discussed in Section 4, while the conclusions are drawn in Section 5.

2 Related Work

The paradigm shift of NLP architectures produced by the rise of neural networks (Bengio et al., 2003; Collobert and Weston, 2008; Collobert et al., 2011) popularized the use of vector space models, following previous work in distributional semantics (Landauer and Dumais, 1997; McDonald and Ramscar, 2001). In this scenario, several highly efficient methods have been proposed to learn low-dimensional vector representation of words (i.e., word embeddings), such as *word2vec* (Mikolov et al., 2013a), *GloVe* (Pennington, Socher, and Manning, 2014), or *fastText* (Bojanowski et al., 2017). Since then, the use of pretrained embeddings to initialize the training of deep learning NLP models has become a standard procedure, due to the positive impact provided by the distributed representations in most downstream tasks (Schnabel et al., 2015).

One of the main drawbacks of these *static* word embeddings for NLP is that they represent all the senses of a given word in the same vector, thus making it difficult to deal with different linguistic phenomena such as polysemy or homonymy. In this regard, Peters et al. (2018) introduced ELMo, a model which obtains contextualized vector representations by means of an LSTM architecture, thus providing context-specific vectors for each token. This language model runs in a unidirectional fashion, in order not to trick itself when predicting the next word.

Following this idea, Devlin et al. (2019) presented BERT, a bidirectional language representation model based on the transformer architecture (Vaswani et al., 2017) and trained on a masked language model objective and on a next sentence prediction one, in order to consider both previous and upcoming context while still being a fair training objective. BERT has obtained state-of-

the-art results for many NLP tasks, and it is easy executable through freely available hubs, becoming a strong baseline in many recent work. In addition, the success achieved by BERT has promoted an extensive implementation of BERT-based models with several goals, such as optimizing its training method, e.g. RoBERTa, (Liu et al., 2019) or reducing the size and complexity of the model, e.g. DistilBERT (Sanh et al., 2019) or ALBERT (Lan et al., 2020)[3]. Moreover, it has attracted the interest of the research community regarding how its deep architecture encodes linguistic knowledge (Lin, Tan, and Frank, 2019; Vilares et al., 2020; Ettinger, 2020).

Besides the original English and Chinese models of BERT, the authors released a multilingual one (mBERT, with 12 layers) which produces inter-linguistic representations to some extent and gives good performance at zero-shot cross-lingual tasks (Pires, Schlinger, and Garrette, 2019). However, several studies have pointed that there are significant differences in performance among the languages covered by mBERT (Wu and Dredze, 2020), and that for some tasks even the best represented ones do not produce competitive results (Vulić et al., 2020). These findings suggest that, if possible, it is worth training monolingual models, especially for those languages which are poorly represented in the multilingual version of BERT.

In fact, monolingual versions of BERT have been trained for various languages using different methods, improving the results of mBERT. Finnish (Virtanen et al., 2019), Spanish (Cañete et al., 2020) and Greek (Koutsikakis et al., 2020) models were trained with about 3 billion tokens using the same parameters as the original BERT-Base (12 layers and 768 vector dimensions). For Portuguese, Souza, Nogueira, and Lotufo (2019) trained two models using about 2.7 billion tokens: one *large*, using the original English BERT-Large (with 24 layers and a vector size of 1024) for initialization, and one *base*, initialized from mBERT. A similar approach has been also adopted for Russian (Kuratov and Arkhipov, 2019), whose RuBERTa model used mBERT as the starting point, too. A comparison with a monolingual training with random initialization showed that starting from the pre-trained mBERT reduces training time and allows for achieving better performance in various tasks. Finally, and more similar to our setting, the monolingual Basque model (Agerri et al., 2020) obtains state-of-the-art performance in different downstream tasks using only 225 million tokens for training a BERT-base model.

With the above in mind, this paper presents the work carried out to train two BERT models for Galician (one small, with only 6 layers, and one base, with 12), and evaluate them in downstream tasks such as POS-tagging, dependency parsing, and named entity recognition. Our approach can be seen as a low-resource scenario, as we only use one GPU and a small corpus of 42 million tokens for training.

3 Bertinho models

This section briefly introduces some ideas about Galician and describes the methodology that we have followed to train and evaluate the *Bertinho* models.

3.1 Galician

Galician is a romance language spoken by about 2.5 million people in the Spanish Autonomous Region of Galicia and adjacent territories (IGE, 2018). It belongs to the Western Ibero-Romance group, being evolved from the medieval Galician-Portuguese (Teyssier, 1987). Both philological and linguistic studies have traditionally classified Galician dialects as part of the same language as Portuguese (Lindley Cintra and Cunha, 1984; Freixeiro Mato, 2003), even though Galician has been standardized as an independent language since the 1970s, mainly through the use of a Spanish-based orthography (Samartim, 2012). In this regard, both Spanish and Portuguese NLP resources and tools have been used and adapted to analyse Galician data (Malvar et al., 2010; Garcia, Gómez-Rodríguez, and Alonso, 2018).

3.2 Training data for language modeling with BERT

For the pre-training phase, where BERT will be trained for language modeling in order to learn to generate robust contextualized word representations, we rely on a small corpus, extracted from the Galician version of the Wikipedia. More particularly, we used

[3]For which there is a monolingual Catalan model: `https://github.com/codegram/calbert`

the 2020-02-01 dump[4] of the Galician version of the Wikipedia. To clean the data, we used `wikiextractor`[5], which transforms the content of Wikipedia articles into raw text.[6] We did not apply any further preprocessing steps, in order to do this training phase as self-supervised as possible. `Wikiextractor` divides the output into a number of text files of 1MB each. We selected the first 95% of these Wikipedia articles for the training set (with a total of 42 million tokens), and the remainder 5% for the dev set (2,5 million tokens), to keep track of the loss and perplexity at different training points and ensure a successful training. As an enclyclopedic resource widely used by the NLP community, the resulting Wikipedia-based corpus is a well-structured and mostly clean dataset which does not contain as much noise as other crawled corpora (e.g. incomplete sentences, lines with no clear end, etc.).

Contrary to the original BERT release and some other monolingual trainings, we simply pre-train on the masked language objective and ignore the next sentence prediction one, since some recent BERT variants have shown that this second objective adds little or no benefit when it comes to fine-tune BERT-based models for downstream tasks (Liu et al., 2019), as we will be doing in this paper (see also Section 4).

3.3 Models

We now describe the procedure that we followed to pre-train our BERT models for language modeling, specifying the differences and similarities with respect to the training of other monolingual BERT models. We also will introduce the framework that we will use to fine-tune our models for downstream tasks.

3.3.1 BERT tokenizer and sub-word vocabulary

The BERT tokenizer splits the words into the so-called sub-word pieces (essentially n-grams of characters, given a word), where the least common words are simply represented by a generic unknown token, UNK. We follow the same setup as in the original English tokenizer, and define for both models a sub-word cased vocabulary of size 30,000. Such size was also set based both on the vocabularies used for BERT models for related linguistic varieties (e.g., Portuguese (Souza, Nogueira, and Lotufo, 2019) or Spanish (Cañete et al., 2020)), and on preliminary tests which suggested that this is a good trade-off between the size and the morphological correspondence of the sub-words.

Even though the tokenizer does not explicitly learn morphological information, better sub-words tend to correspond to morphological affixes, i.e., prefixes, suffixes or stems. In this regard, it is worth noting that other authors have used a larger vocabulary size to train BERT models in agglutinative languages, such as Basque (Agerri et al., 2020) or Finnish (Virtanen et al., 2019). Exploring an optimal vocabulary size for Galician falls out of the scope of this paper, but it might be an interesting future research line. Nevertheless, we show an example sentence tokenized by mBERT and by our model in Table 1. As it can be seen, mBERT splits the word *dixéronnos* ('they told us') into several sub-words, including 'dix', 'éro', 'nno', and 's'. Except for the first one ('dix') which corresponds to the frequent irregular root of the verb *dicir* ('tell'), the rest of the original token was split without taking into account morphological boundaries. However, the third sub-word identified by our model ('nos') is a correctly split clitic pronoun (masculine plural dative, 'to us'), while the second one ('éron') contains the thematic vowel and person morphemes. In addition, mBERT split the sub-word 'iño' from *camiño* ('way' or 'path'), which in turn can be confused with the very frequent diminutive suffix 'iño' (e.g., the diminutive of *carro* –'car'– is *carriño*), thus potentially involving inadequate representations of the whole word. Finally, the masculine possessive determiner *nosos* (with both plural possessor and possessed) was also split by mBERT (and not by our model, as it is a frequent word), but in this case it could be argued that the mBERT tokenization might not hurt the model, as the second sub-word 'os' corresponds to the masculine plural suffix, while the first one ('nos') could be analyzed as the morphological root (even though it also corresponds to a personal pronoun).

[4] Note that only the newest dumps are maintained over time `https://dumps.wikimedia.org/glwiki/`, but the differences should have a small effect in practice.

[5] `https://github.com/attardi/wikiextractor`

[6] In our work, we kept both the main texts and the headers, too.

Model	Tokenization
mBERT	Os nos ##os amigos dix ##éro ##nno ##s que o cam ##iño era este .
Ours	Os nosos amigos dix ##éron ##nos que o camiño era este .

Table 1: Tokenization of the sentence *Os nosos amigos dixéronnos que o camiño era este.* ('Our friends told us that this was the way.') by the original mBERT and our model. Following the same output representation as the BERT tokenizer, we use the symbol ## to specify a sub-word that is *not the first* sub-word of a split token.

3.3.2 Pre-training for language modeling

We have trained two models, varying the number of layers: (i) a BERT with 6 transformer layers (Bertinho$_{\text{SMALL}}$), and (ii) a BERT with 12 transformer layers (Bertinho$_{\text{BASE}}$), both trained on the Wikipedia corpus and using the tokenizer presented above. Each layer produces hidden representations of size 768, as in the original BERT paper for monolingual models. We use these models to explore the performance of each architecture in several downstream applications, thus being able to analyze the trade-off between the size and complexity of the models and their quality on extrinsic tasks.

With respect to the hyper-parameter settings, we mostly follow the standard pre-training configuration used in the original BERT paper. We use a learning rate of 1×10^{-4} with a linear weight decay of 0.01. We choose Adam as the network weight optimizer (Kingma and Ba, 2015) with $\epsilon = 1 \times 10^{-8}$. For the masked language objective and given an input sentence of length n, we mask randomly a 15% of the tokens. From those, 80% of them are replaced by the wildcard symbol [MASK], 10% of them are changed to a random word from the input vocabulary, and the remaining 10% are not modified.

With respect to the training process, in the original BERT model the authors train for 1M steps on sequences of 512 tokens and a batch size of 256. This strategy significantly increases the training time, as the self-attention included at each of the transformer layers in the BERT models runs in $\mathcal{O}(n^2)$, where n is the number of input tokens. Alternatively, authors such as Agerri et al. (2020) use a two-phase procedure using also a training batch size of 256. On the first phase, they train on sequences of length 128 during 900 000 steps. On the second phase, they continue the training considering sequences of length 512 during 100 000 additional steps. However, even with this more modest training setup, the authors still had access to a few TPUs.

In our work, we stuck to a even lower computational resource setup, training the models on a single TESLA P40 GPU of 24GB. Following the standard approach, we first trained the 12-layer model, and consider a two-step training procedure (with the second phase being optional). For phase 1, we used a smaller training batch imposed by the hardware limitations. More particularly, we used training batches of size 96 considering sequences of 128 tokens. To counteract such smaller training batch, we trained instead the model during more steps, up to 2M. This phase 1 took 30 days to complete the training. Optionally, if the phase 2 was applied, we kept training the model from phase 1 using sequences of length 512. However, this required to limit our training batch size by a significant amount (more particularly, we could only fit 12 sequences in memory). We trained this second sequence for 1.4M additional steps (which took 4 extra days). Thus, for Bertinho$_{\text{BASE}}$ we have an additional model trained with the two-phase strategy, that we will be referring as Bertinho$_{\text{BASE-2PH}}$. Figure 1 shows the evaluation perplexity obtained at different steps during phase 1 (shorter sequences) and phase 2 (longer sentences). However, we observed that this second phase was not useful to improve the results over the Bertinho$_{\text{BASE}}$ model obtained after finishing phase 1 when it came to fine-tune models for downstream tasks (see also Section 4). For the 6-layers model, Bertinho$_{\text{SMALL}}$, we decided to apply only the phase 1 of the training procedure, but with batches of size 128 instead and training for 1.5M steps (taking about 22 days to be completed). Figure 2 shows the perplexity of the SMALL model during evaluation.

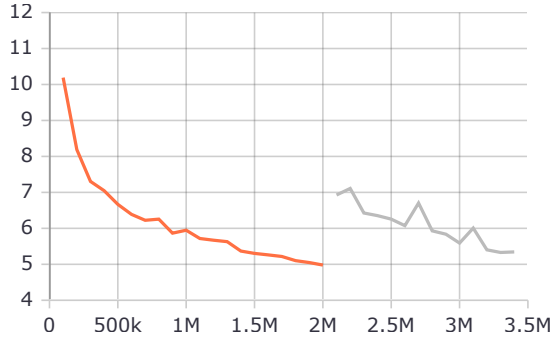

Figure 1: Eval perplexity for Bertinho$_{\text{BASE}}$ (12 layers). The eval perplexity over sequences of 128 tokens during the phase 1 of the training is showed in red (left plot). The one over sequences of 512 tokens during the second phase is showed in gray (right plot).

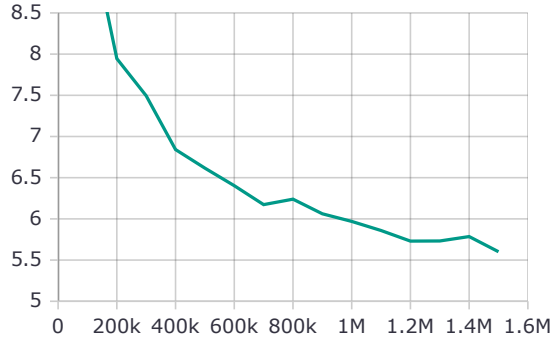

Figure 2: Eval perplexity for Bertinho$_{\text{SMALL}}$, trained only over sequence of 128 tokens.

All the models have been trained using the *transformers*[7] library provided by the Hugging Face team (Wolf et al., 2019) using, as mentioned, a single 24 GB TESLA P40 GPU.

3.3.3 Framework

Given an input sentence $\vec{w}$=$[w_1, w_2, ..., w_n]$, we propose to fine-tune the BERT pretrained models for language modeling to a number of tasks involving lexical, syntactic and semantic processing. We will cast these downstream tasks in a sequence labeling setup, i.e, we will learn a mapping $\phi : W^n \rightarrow L^n$ where L will represent the sequence of output labels. The tasks that we will study are POS-tagging, dependency parsing and named entity recognition, described in detail in Section 4.

That said, let BERT($\vec{w}$) be a pre-trained BERT model that maps a one-input vector to a sequence of contextualized representations, $\vec{h}$, we simply map each h_i to an output label y_i through a 1-layer feed-forward network using a softmax ($P(y_i = j|\vec{h}_i) = softmax(\vec{W} \cdot \vec{h}_i + \vec{b}) = \frac{e^{\vec{W}_j \cdot \vec{h}_i}}{\sum_k^K e^{\vec{W}_k \cdot \vec{h}_i}}$) to obtain a probability distribution over the set of output labels for each word, w_i. In our work, all labels are predicted in an atomic way, that is with single-task learning.[8] In all cases, the loss, $\mathcal{L}$, is optimized based on categorical cross-entropy loss ($\mathcal{L} = -\sum log(P(y_i|\vec{h}_i))$), and back-propagated through the whole network, i.e. we also fine-tune the BERT weights during the training for the downstream tasks.

4 *Experiments and results*

To evaluate the usefulness and robustness of the *Bertinho* models when it comes to fine-tuning the model for downstream tasks, we will consider three different problems: (i) part-of-speech tagging, (ii) dependency parsing, and (iii) named entity recognition. We will rely on existing datasets for Galician, which we now briefly review, indicating as well for which tasks we will use them:

- **CTAG corpus:** The *Corpus Técnico do Galego* (Galician Technical Corpus) is a linguistic corpus developed by the Seminario de Lingüística Informática of the University of Vigo, and it is composed of texts from a variety of technical domains (e.g., legal and scientific), totaling almost 18 million words[9]. A subset of the CTG corpus, the CTAG (Corpus Técnico Anotado do Galego, 'Galician Technical Annotated Corpus')[10], was manually reviewed (Guinovart and Fernández, 2009; Agerri et al., 2018), therefore producing gold-standard resources for POS-tagging and lemmatization (with more than 2 million tokens (TALG, 2016)) and NER (with about 200k tokens (TALG, 2018)).

  Tasks and observations: We will use these resources to evaluate the performance of the BERT-based models on POS-tagging and named entity recognition.

[7] https://github.com/huggingface/transformers

[8] Especially for more complex tasks, such as dependency parsing as sequence labeling, other authors (Strzyz, Vilares, and Gómez-Rodríguez, 2019) have decomposed the task into predicting partial labels using a multi-task learning setup.

[9] http://sli.uvigo.es/CTG/

[10] http://sli.uvigo.es/CTAG/

- **CTG-UD:** The CTG-UD (Guinovart, 2017) is a treebank based on a subpart of the CTG corpus which has been automatically parsed and adapted to Universal Dependencies[11] (Nivre et al., 2020).

  Tasks and observations: CTG-UD will be used to evaluate both POS-tagging and dependency parsing. For POS-tagging, we will consider both universal and language-specific part-of-speech tags (based on the fine-grained tagset of FreeLing (Padró, 2011; Garcia and Gamallo, 2010)) as separate tasks. For dependency parsing, and to cast it as a sequence labeling task, we are relying on the encodings proposed by Strzyz, Vilares, and Gómez-Rodríguez (2019), which defined different ways to map a sequence of input words of length n into a sequence of syntactic labels, also of length n, that can be decoded into full dependency trees. More particularly we will consider their bracketing encoding. Since it is outside of the scope of this work, we leave the details of the encodings for the reader, which can be found in the referrenced work, and we simply use their encoding and decoding functions as a black box.

- **Galician-TreeGal:** This treebank is a subset of the XIADA corpus (Rojo et al., 2019) that has been annotated following the guidelines of the Universal Dependencies initiative. Galician-TreeGal has around 25k tokens, and its manually revised annotation includes lemmas, POS-tags, morphological features and dependency labels (Garcia, Gómez-Rodríguez, and Alonso, 2016; Garcia, Gómez-Rodríguez, and Alonso, 2018).

  Tasks and observations: As in the case of the CTG-UD, we will use this dataset for universal and language-dependent POS-tagging (using the morphologically-rich XIADA tagset)[12], and also for UD dependency parsing. It is worth mentioning that due to the small size of the Galician TreeGal treebank, the evaluation can be considered as few-shot learning.

All the experiments have been carried out by fine-tuning the BERT-based models ($\text{Bertinho}_{\text{SMALL}}$ and $\text{Bertinho}_{\text{BASE}}$) for each specific task using the splits for training and development of each corpus[13] with the referred *transformers* library. As a baseline, we use the official mBERT released by Google (BERT-Base Multilingual Cased, with 12 layers), which includes Galician among its 104 covered languages.

4.1 Results

We first show the POS-tagging and NER results on the CTAG corpus, followed by the analyses on the Universal Dependencies treebanks (CTG-UD and Galician-TreeGal).

4.1.1 CTAG

Table 2 shows the POS-tagging and NER results on the CTAG corpus. The POS tagset contains 178 tags, and the NER labeling was approached as a BIO classification of four classes (person, location, organization, and miscellaneous) totaling 9 tags (two –B and I– for each class, and O for the tokens outside the named entities).

On POS-tagging, the best results are obtained by $\text{Bertinho}_{\text{BASE}}$ (two-phase and single-phase, respectively), followed $\text{Bertinho}_{\text{SMALL}}$, all of them surpassing mBERT in this task. Interestingly, even the 6-layer model obtained better results (2.32%) than the multilingual one, with 12 transformer layers. However, it is worth noting that mBERT outperformed all the monolingual models on named entity recognition, both in precision and recall, suggesting that for this task multilingual information may improve the performance. These and other results are discussed on Section 4.2.

4.1.2 Universal Dependencies

With respect to the Universal Dependencies treebanks, the results on the CTG-UD are shown on Table 3, including POS-tagging (on UD POS-tags and using a fine-grained tagset) and dependency parsing (both LAS and UAS values).

In this case, and also in further experiments, $\text{Bertinho}_{\text{BASE}}$ obtained the best results in both tasks and settings, followed by

[11] https://universaldependencies.org/

[12] http://corpus.cirp.es/xiada/etiquetario/taboa

[13] Since Galician-TreeGal does not contain a development set, we have splitted the train data into train/dev with a 90%/10% ratio. Similarly, we have used 150k tokens from the training set of the CTAG for development of the POS-taggers.

Model	CTAG			
	POS	NER		
	Acc	P	R	F1
mBERT	93.84	83.53	85.60	84.55
$\text{Bertinho}_{\text{SMALL}}$	96.16	78.40	82.68	80.48
$\text{Bertinho}_{\text{BASE}}$	96.40	80.49	82.74	81.60
$\text{Bertinho}_{\text{BASE-2PH}}$	96.23	80.89	84.33	82.57

Table 2: POS-tagging and NER (precision, recall, and f-score) results on the CTAG corpus. The POS tagset corresponds to the fine-grained tags of FreeLing (see Section 3.2).

Model	UPOS	FPOS	LAS	UAS
mBERT	95.41	91.91	76.48	80.80
$\text{Bertinho}_{\text{SMALL}}$	96.42	94.56	77.59	81.55
$\text{Bertinho}_{\text{BASE}}$	96.56	94.60	78.14	81.88
$\text{Bertinho}_{\text{BASE-2PH}}$	96.43	94.50	77.70	81.65

Table 3: POS-tagging accuracies and dependency parsing results on the CTG-UD treebank. POS includes universal and language-dependent tagsets, while parsing is evaluated using LAS and UAS.

$\text{Bertinho}_{\text{BASE-2PH}}$ and by $\text{Bertinho}_{\text{SMALL}}$ (almost tied) and mBERT. As expected, the performance on POS-tagging is higher when using the UPOS tagset (with 16 elements) than with the FPOS one (194 tags). On dependency parsing, all the models follow the same mentioned tendency, with $\text{Bertinho}_{\text{BASE}}$ achieving the best results: 78.14 and 81.88 (LAS and UAS, respectively).

When moving to the Galician-TreeGal treebank (Table 4), we see again that the best performance is achieved by $\text{Bertinho}_{\text{BASE}}$ (single and two-phases, respectively), followed by the SMALL variant, and finally by mBERT. It is worth recalling that this treebank is significantly smaller than the other datasets, therefore the model weights have more influence as the fine-tuning process is shorter. In this regard, the results of all models are lower than those obtained on the CTG-UD dataset, achieving POS-tagging accuracies of up to 96.61% (UPOS, with 16 tags) and 92.70% (FPOS, containing a tagset of 237 elements). On dependency parsing, the best results were of 75.26% and 80.27% on LAS and UAS, respectively.

Significance tests: We applied significance tests for POS-tagging and parsing, to determine whether the proposed monolingual models are actually different than mBERT. For POS-tagging, we applied a t-test that compares the accuracies per sentence obtained by mBERT and each of the monolingual models. All models are significantly different (with $p < 0.01$), except $\text{Bertinho}_{\text{BASE-2PH}}$ on the Galician-TreeGal (UPOS). For parsing, as in Vilares and Gómez-Rodríguez (2018), we used instead the Bikel's randomized parsing evaluation comparator, a stratified shuffling significance test. The null hypothesis is that the outputs produced by mBERT and any of the monolingual models are produced by similar models and so the scores are equally likely. To refute it, it first measures the difference obtained for a metric by the two models. Then, it shuffles scores of individual sentences between the two models and recalculates the metrics, checking if the difference is less than the original one, which would be an indicator that the outputs generated by the models are significantly different. All models are significantly different (with $p < 0.01$).

4.2 Discussion

There is a clear tendency with respect to the performance of the different models on the downstream tasks. Except for NER (discussed below), $\text{Bertinho}_{\text{BASE}}$ (12 layers) consistently obtains the best results, following by $\text{Bertinho}_{\text{SMALL}}$ (6 layers) and the official mBERT release (also with 12 layers).

Model	UPOS	FPOS	LAS	UAS
mBERT	94.27	87.67	71.54	77.67
Bertinho$_{\text{SMALL}}$	96.38	92.05	73.23	78.71
Bertinho$_{\text{BASE}}$	96.61	92.70	75.26	80.27
Bertinho$_{\text{BASE-2PH}}$	96.46	92.69	74.41	79.64

Table 4: POS-tagging accuracies and dependency parsing results on the Galician-TreeGal treebank. POS includes universal and language-dependent tagsets, while parsing is evaluated using LAS and UAS.

Apart from the models themselves and from task-specific properties, there are two parameters that seem to play an important role when fine-tuning for a particular downstream task: the size of the tagset and the amount of training data. In this regard, the gain obtained by the best models (when compared to those with lower results) is higher when using large tagsets and smaller datasets, thus suggesting that they encode better information which in turn has a stronger impact on the final performance.

We have assessed this finding by observing the differences between the results of our best model (Bertinho$_{\text{BASE}}$) with those of mBERT: With respect to the size of the tagset, we compared the results on UPOS and FPOS on both Universal Dependencies corpora. The difference between Bertinho$_{\text{BASE}}$ and mBERT on UPOS tagging (with 16 tags) was of 1.15 and 2.34 on CTG-UD and Galician-TreeGal, respectively. However, when using FPOS (with 194 and 237 different tags on the mentioned corpora), the gain achieved by Bertinho$_{\text{BASE}}$ increased to 2.69 (CTG-UD) and 5.03 (Galician-TreeGal).

Besides the divergences between the UPOS and FPOS scenarios, these results also indicate that the differences are noticeably larger on the Galician-TreeGal than on the CTG-UD dataset. As the former treebank has less than 14k tokens for training (while CTG-UD has $\approx$ 80k), the differences seem to be mainly caused by the disparity on the amount of training data. This tendency is also displayed on dependency parsing, where the LAS/UAS differences between Bertinho$_{\text{BASE}}$ and mBERT are of 1.66/1.08 on CTG-UD and of 3.72/2.60 on Galician-TreeGal. Finally, this tendency does not hold when comparing the POS-tagging results on both CTAG and CTG datasets (2.56 *versus* 2.69 on the CTAG and CTG-UD, respectively). Additionally, in this case the UD variant is 10 times smaller and has a larger dataset (194 *versus* 178), and consequently the results are lower than in the non-UD corpus (94.60 *versus* 96.40).

The mentioned tendency is not followed in the NER results, where the multilingual model achieved impressive performance (surpassing the best results published by Agerri et al. (2018)). Bertinho$_{\text{BASE}}$ beats the SMALL variant by 1.12 points, but mBERT overcomes the best monolingual model (Bertinho$_{\text{BASE-2PH}}$) by 1.98 points. At first glance, we could hypothesize that the multilingual model performs better at NER as *enamex* named entities (locations, people, organizations, and miscellaneous entities) are represented interlinguistically, so that the model takes advantage of information from various languages. Although this may affect the results in some way, a careful analysis of the output of mBERT and Bertinho$_{\text{BASE}}$ has shown that most errors of the monolingual model came from variation regarding upper and lowercase. Thus, in expressions such as "Especies máis afectadas polo *Plano de Selado*" ('Species most affected by the Sealing Plan'), "O *Código Civil* actual" ('The current Civil Code'), or "As *Illas do Sur*" ('The Southern Islands'), our model classified the expressions in italic as miscellaneous (the fist two) and location (the last one) entities, but they are not labeled in the gold standard dataset. About the identification of person entities, Bertinho$_{\text{BASE}}$ failed in some complex nouns including prepositions (e.g., "Bernardo Barreiro de Vázquez Varela" labeled as two entities –separated by 'de'– instead of one), while some organizations named with common nouns were not identified by the NER system (e.g., "Cachoeira", 'waterfall').[14] Therefore, it could

[14]Some other errors of Bertinho$_{\text{BASE}}$ were due

be interesting to analyze different techniques to deal with these issues in further work.

Finally, it is worth noting that the Bertinho$_{\text{BASE}}$ variant trained with a two-phase strategy (Bertinho$_{\text{BASE-2PH}}$) consistently obtained worse performance than the single-phase one, except in the named entity recognition scenario. Even though more investigation is needed, these results indicate that in our case, and contrary to previous related work, a two-phase training procedure (training with longer sentences during the second phase) was not beneficial, hurting the performance of the model in the downstream tasks. We hypothesize this might be partially due to our limited hardware resources, that imposed us a very small training batch during the second phase of training.

In sum, we have trained and evaluated two BERT models for Galician that obtain better results than the official multilingual model in different downstream tasks. Interestingly, we provide both a 12-layer model (with greater performance), and a 6-layer one which obtains competitive results with a computationally less expensive architecture.

5 Conclusion

We have trained two monolingual BERT models for Galician (dubbed *Bertinho*), for which we have followed a low-resource approach with less than 45 million tokens of training data in a single GPU. Both models have been evaluated on several downstream tasks, namely on dependency parsing, NER, and POS-tagging with different tagsets. Moreover, we have shown how a dedicated tokenizer improves the morphological segmentation of Galician words.

Our best model (Bertinho$_{\text{BASE}}$, with 12 layers) outperforms the official BERT multilingual model (mBERT) in most downstream tasks and settings, and even the small one (Bertinho$_{\text{SMALL}}$ with 6 layers) achieves better results than mBERT on the same tasks. However, it is worth noting that mBERT has worked better than the monolingual models on NER. Finally, our experiments have also shown that a two-phase training procedure for language modeling (with more learning steps and training with longer sequence at the end) consistently hurts the performance of the 12-layer model on most scenarios in our setup. We believe this might be due to our limited hardware resources, that forced us to use a small training batch when pre-training with very long sequences.

In further work we plan to carry out a deeper analysis of the NER results, and also to compare the different layers of the BERT models with static word embeddings such as *word2vec* or GloVe. Furthermore, we aim to extend our models using larger corpora. *Bertinho* has been trained on less data than other BERT models for related languages such as BETO (Cañete et al., 2020). We believe collecting more data for other sources such as CommonCrawl could improve the performance. However it is also fair to state that there are also studies that point in the opposite direction. For instance, the results of Raffel et al. (2020) indicate that smaller clean datasets are better than large noisy corpora, so that it could be interesting to assess to what extent our results (obtained with a small dataset) can be improved with new data crawled from the web (Agerri et al., 2018; Wenzek et al., 2020).

Finally, it is important to recall that the work performed in this study contributes to the NLP community with the release of two freely available *Bertinho* models for Galician.

Acknowledgements

This work has received funding from the European Research Council (ERC), which has funded this research under the European Union's Horizon 2020 research and innovation programme (FASTPARSE, grant agreement No 714150), from MINECO (ANSWER-ASAP, TIN2017-85160-C2-1-R), from Xunta de Galicia (ED431C 2020/11), from Centro de Investigación de Galicia 'CITIC', funded by Xunta de Galicia and the European Union (European Regional Development Fund- Galicia 2014-2020 Program), by grant ED431G 2019/01, and by Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), ERDF 2014-2020: Call ED431G 2019/04. DV is supported by a 2020 Leonardo Grant for Researchers and Cultural Creators from the BBVA Foundation. MG is supported by a Ramón y Cajal grant (RYC2019-028473-I).

to mislabelings in the gold-standard, as *Kiko* in "Camiño Neocatecumenal de *Kiko Argüelles*" (*sic*), annotated as I-PER instead of B-PER.

References

Agerri, R., X. Gómez Guinovart, G. Rigau, and M. A. Solla Portela. 2018. Developing new linguistic resources and tools for the Galician language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Agerri, R., I. San Vicente, J. A. Campos, A. Barrena, X. Saralegi, A. Soroa, and E. Agirre. 2020. Give your text representation models some love: the case for Basque. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4781–4788, Marseille, France, May. European Language Resources Association.

Bengio, Y., R. Ducharme, P. Vincent, and C. Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Cañete, J., G. Chaperon, R. Fuentes, and J. Pérez. 2020. Spanish pre-trained BERT model and evaluation data. In *Practical ML for Developing Countries Workshop (PML4DC) at ICLR*. Learning under limited/low resource scenarios.

Collobert, R. and J. Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167.

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. 2011. Natural Language Processing (Almost) From Scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Dai, A. M. and Q. V. Le. 2015. Semi-supervised sequence learning. In *Advances in neural information processing systems*, pages 3079–3087.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Ettinger, A. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Freixeiro Mato, X. R. 2003. *Gramática da Lingua Galega IV. Gramática do texto*. A Nosa Terra, Vigo.

Garcia, M. and P. Gamallo. 2010. Análise Morfossintáctica para Português Europeu e Galego: Problemas, Soluções e Avaliação. *Linguamática*, 2(2):59–67.

Garcia, M., C. Gómez-Rodríguez, and M. A. Alonso. 2016. Creación de un treebank de dependencias universales mediante recursos existentes para lenguas próximas: el caso del gallego. *Procesamiento del Lenguaje Natural*, 57:33–40.

Garcia, M., C. Gómez-Rodríguez, and M. A. Alonso. 2018. New treebank or repurposed? on the feasibility of cross-lingual parsing of romance languages with universal dependencies. *Natural Language Engineering*, 24(1):91–122.

Guinovart, X. G. and S. L. Fernández. 2009. Anotación morfosintáctica do Corpus Técnico do Galego. *Linguamática*, 1(1):61–70.

Guinovart, X. 2017. Recursos integrados da lingua galega para a investigación lingüística. *Gallaecia. Estudos de lingüística portuguesa e galega*, pages 1045–1056.

IGE. 2018. Coñecemento e uso do galego. Instituto Galego de Estatística, `http://www.ige.eu/web/mostrar_actividade_estatistica.jsp?idioma=gl&codigo=0206004`.

Jiang, N. and M.-C. de Marneffe. 2019. Evaluating BERT for natural language inference: A case study on the CommitmentBank. In *Proceedings of the 2019 Con-*

ference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6086–6091, Hong Kong, China, November. Association for Computational Linguistics.

Karthikeyan, K., Z. Wang, S. Mayhew, and D. Roth. 2020. Cross-Lingual Ability of Multilingual BERT: An Empirical Study. In *International Conference on Learning Representations (ICRL 2020)*.

Kingma, D. P. and J. Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR 2015)*. arXiv preprint arXiv:1412.6980.

Kitaev, N. and D. Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia, July. Association for Computational Linguistics.

Koutsikakis, J., I. Chalkidis, P. Malakasiotis, and I. Androutsopoulos. 2020. GREEK-BERT: The Greeks visiting Sesame Street. In *11th Hellenic Conference on Artificial Intelligence*. ACM.

Kuratov, Y. and M. Arkhipov. 2019. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. *Computational Linguistics and Intellectual Technologies*, 18:333–339.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *Proceedings of the International Conference of Learning Representations (ICLR 2020)*.

Landauer, T. K. and S. T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Lee, S., H. Jang, Y. Baik, S. Park, and H. Shin. 2020. KR-BERT: A Small-Scale Korean-Specific Language Model. *arXiv preprint arXiv:2008.03979*.

Lin, Y., Y. C. Tan, and R. Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy, August. Association for Computational Linguistics.

Lindley Cintra, L. F. and C. Cunha. 1984. *Nova Gramática do Português Contemporâneo*. Livraria Sá da Costa, Lisbon.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

Malvar, P., J. R. Pichel, Ó. Senra, P. Gamallo, and A. García. 2010. Vencendo a escassez de recursos computacionais. carvalho: Tradutor automático estatístico inglês-galego a partir do corpus paralelo europarl inglês-português. *Linguamática*, 2(2):31–38.

McDonald, S. and M. Ramscar. 2001. Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 23.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013a. Efficient estimation of word representations in vector space. In *Workshop Proceedings of the International Conference on Learning Representations (ICLR) 2013*. arXiv preprint arXiv:1301.3781.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Nivre, J., M.-C. de Marneffe, F. Ginter, J. Hajič, C. D. Manning, S. Pyysalo, S. Schuster, F. Tyers, and D. Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France, May. European Language Resources Association.

Ortiz Suárez, P. J., L. Romary, and B. Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online, July. Association for Computational Linguistics.

Padró, L. 2011. Analizadores Multilingües en FreeLing. *Linguamatica*, 3(2):13–20.

Pennington, J., R. Socher, and C. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Peters, M., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Pires, T., E. Schlinger, and D. Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July. Association for Computational Linguistics.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Rojo, G., M. López Martínez, E. Domínguez Noya, and F. Barcala. 2019. Corpus de adestramento do Etiquetador/Lematizador do Galego Actual (XIADA), versión 2.7. *Centro Ramón Piñeiro para a investigación en humanidades*.

Salant, S. and J. Berant. 2018. Contextualized word representations for reading comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 554–559, New Orleans, Louisiana, June. Association for Computational Linguistics.

Samartim, R. 2012. Língua somos: A construção da ideia de língua e da identidade coletiva na galiza (pré-) constitucional. In *Novas achegas ao estudo da cultura galega II: enfoques socio-históricos e lingüísticoliterarios*, pages 27–36.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Proceedings of The 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing* at NeurIPS2019. arXiv preprint arXiv:1910.01108.

Schnabel, T., I. Labutov, D. Mimno, and T. Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal, September. Association for Computational Linguistics.

Souza, F., R. Nogueira, and R. Lotufo. 2019. Portuguese named entity recognition using BERT-CRF. arXiv preprint arXiv:1909.10649.

Strzyz, M., D. Vilares, and C. Gómez-Rodríguez. 2019. Viable dependency parsing as sequence labeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 717–723, Minneapolis, Minnesota, June. Association for Computational Linguistics.

TALG. 2016. CTG Corpus (Galician Technical Corpus). TALG Research Group. SLI resources, 1.0, ISLRN 437-045-879-366-6.

TALG. 2018. SLI NERC Galician Gold CoNLL. TALG Research Group. SLI resources, 1.0, ISLRN 435-026-256-395-4.

Teyssier, P. 1987. *História da Língua Portuguesa*. Livraria Sá da Costa, Lisbon, 3 edition.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention Is All You Need. arXiv preprint arXiv:1706.03762.

Vilares, D. and C. Gómez-Rodríguez. 2018. Transition-based parsing with lighter feed-forward networks. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 162–172, Brussels, Belgium, November. Association for Computational Linguistics.

Vilares, D., M. Strzyz, A. Søgaard, and C. Gómez-Rodríguez. 2020. Parsing as pretraining. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9114–9121. AAAI Press.

Virtanen, A., J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter, and S. Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. arXiv preprint arXiv:1912.07076.

Vulić, I., E. M. Ponti, R. Litschko, G. Glavaš, and A. Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online, November. Association for Computational Linguistics.

Wenzek, G., M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and E. Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France, May. European Language Resources Association.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv preprint arXiv:1910.03771.

Wu, S. and M. Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online, July. Association for Computational Linguistics.

Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 27-39 recibido 25-11-2020 revisado 28-01-2021 aceptado 02-02-2021

Evaluación de un modelo transformador aplicado a la tarea de generación de resúmenes en distintos dominios

Evaluation of a transformer model applied to the task of text summarization in different domains

Isabel Segura-Bedmar, Lucía Ruz, Sara Guerrero-Aspizua
Universidad Carlos III de Madrid, Leganés, Spain
isegura@inf.uc3m.es, lruz@pa.uc3m.es, sguerrer@ing.uc3m.es

Resumen: En los últimos años, las técnicas de deep learning han supuesto un gran impulso tecnológico en muchas de las tareas de Procesamiento de Lenguaje Natural (PLN). La tarea de generación de resúmenes también se ha beneficiado de estas técnicas, y en los últimos años se han implementado distintos modelos, logrando superar los resultados del estado de la cuestión. La mayoría de estos trabajos han sido evaluados en colecciones de textos periodísticos. Este artículo presenta un trabajo preliminar donde aplicamos un modelo transformador, BART, para la tarea de generación de resúmenes y lo evaluamos en varios datasets, uno de ellos formado por textos del dominio biomédico.
Palabras clave: Generación de resúmenes, Transformadores.

Abstract: In recent years, deep learning techniques have provided a significant technological advance in many Natural Language Processing (NLP) tasks. Text summarization has also benefited from these techniques. Recently, several deep learning approaches have been implemented, surpassing the previous state of the art performances. Most of these works have been evaluated on collections of journalistic texts. This article presents a preliminary work where we apply a transforming model, BART, for text summarization. The model is evaluated on several datasets, one of them consisting of texts from the biomedical domain.
Keywords: Text summarization, Transformers.

1 Introducción

En la era del Big Data, existe una gran cantidad de información en formato digital, la mayor parte de ella en texto no estructurado. Este gran volumen de datos nos aporta conocimiento, pero necesariamente primero nos enfrenta al reto de poder procesar y comprender toda esta información. Un resumen nos proporciona las ideas principales de un texto, por tanto, la generación de resúmenes puede ayudarnos a la hora de acceder a la información esencial de un texto o colección de textos. Dado que la generación de resúmenes de manera manual es una tarea compleja que conlleva mucho tiempo y recursos, las técnicas de Procesamiento de Lenguaje Natural (PLN) pueden ayudar a reducir esta carga de trabajo.

En la generación automática de resúmenes, se distinguen dos enfoques: extractivo y abstractivo. El enfoque extractivo consiste en la selección de las oraciones más relevantes del texto para formar el resumen. Por otro lado, el método abstractivo genera nuevas oraciones para la creación del resumen. En los últimos años, las técnicas de deep learning han supuesto un gran impulso tecnológico en muchas de las tareas de PLN (Beloki et al., 2020; Miranda-Escalada y Segura-Bedmar, 2020; Colón-Ruiz, Segura-Bedmar, y Martínez, 2019; Poncelas et al., 2019), en muchos casos, obteniendo mejores resultados que los algoritmos clásicos de aprendizaje automático. La generación automática de resúmenes también se ha beneficiado de esta tecnología, y en los últimos años, muchos de los sistemas han utilizado técnicas de deep learning como las redes convolucionales (Zhang et al., 2019) o las redes recurrentes (Nallapati, Zhai, y Zhou, 2017). En 2017, se propone un nuevo modelo de deep learning, Transformers(Vaswani et al.,

ISSN 1989-7553. DOI 10.26342/2021-66-2

2017), cuya principal ventaja es que no necesita procesar los textos de forma secuencial, como hacen las redes recurrentes. Esto facilita la paralelización de la tarea y reduce los tiempos de entrenamiento. Estos modelos también han sido aplicados a la tareas de generación de resúmenes (Liu y Lapata, 2019; Zhang, Wei, y Zhou, 2019), con buenos resultados. Sin embargo, la mayoría de estos trabajos han sido evaluados sobre las colecciones de artículos periodísticos, como por ejemplo el corpus CNN/Daily Mail.

En este trabajo, abordamos la tarea de generación de resúmenes utilizando el modelo BART (Lewis et al., 2020), basado en la arquitectura Sequential-to-Sequential. Además, el modelo será evaluado sobre conjuntos de datos distintos. Este artículo presenta un trabajo preliminar, cuyo principal objetivo será comprobar si el modelo BART es capaz de generar resúmenes automáticos para textos de otros estilos distintos al periodístico, como son los textos biomédicos. Además, comprobaremos si BART es capaz de generar resúmenes de calidad a partir de conjuntos de datos, cuyas características en relación al número de instancias y al tamaño medio de sus textos y resúmenes son distintas a las del corpus CNN/DailyMail, que fue utilizado para preentrenar el modelo BART.

El artículo está organizado como sigue: la sección 2 revisa los modelos más recientes aplicados a la tarea de generación de resúmenes. En la sección 3, presentamos el modelo aplicado en este trabajo. La sección 4 describe los datasets utilizados en la evaluación del modelo. Además, presenta y discute los resultados para cada uno de los datasets. Finalmente, las principales conclusiones y líneas de trabajo futuro serán descritas al detalle en la sección 5.

2 Estado de la cuestión

Como se ha indicado en la introducción, el método transformador es un modelo de aprendizaje profundo que fue creado en el año 2017 (Vaswani et al., 2017). Es un modelo diseñado para el procesamiento de datos secuenciales, y, por tanto adecuado para abordar tareas del PLN, como la clasificación de textos, la generación de resúmenes, entre otras muchas. A diferencia del modelo de redes neuronales recurrentes (RNN), esta arquitectura procesa los datos de manera paralela, sin necesidad de ser procesados en orden. Gracias a este tipo de procesamiento se disminuyen notablemente los tiempos de entrenamiento. Este menor coste computacional, sumado a los excelentes resultados obtenidos por los sistemas que utilizan el modelo, ha revolucionado el campo del PLN, reemplazando a las RNN y a otros modelos de aprendizaje automático profundo.

La investigación en generación de resúmenes no ha permanecido ajena al auge de los modelos transformadores. Desde el 2015, distintas arquitecturas de redes profundas, como por ejemplo, las RNN han sido aplicadas para la generación automática de resúmenes (Nallapati, Zhai, y Zhou, 2017; Nallapati et al., 2016), estableciendo el estado de la cuestión en torno al 30 % de la media de las métricas Rouge-1, Rouge-2 y Rouge-L, sobre el corpus CNN/Daily Mail. A continuación, revisamos los trabajos más recientes sobre generación de resúmenes basados en modelos transformadores.

Zhong et al. (2019) presentan un primer estudio para la generación de resúmenes extractivos mediante redes de neuronas profundas, entre ellas la red recurrente Long short-term memory (LSTM) (Hochreiter y Schmidhuber, 1997), y BERT (Devlin et al., 2019), un modelo transformador. Los experimentos fueron realizados sobre dos datasets: CNN/Daily Mail (Hermann et al., 2015), descrito en el apartado 4.1, y Newsroom (Grusky, Naaman, y Artzi, 2018) formado por 1.3 millones de artículos y resúmenes extraídos de 38 fuentes de noticias a lo largo de los últimos 20 años. En este trabajo, los autores se centraron únicamente en las siete fuentes que proporcionaban un mayor número de ejemplos: New York Times (NYT), Washington Post, Fox News, The Guardian, New York Times Daily News, The Wall Street Journal (WSJ) y USA Today. En los experimentos, cada modelo es evaluado sobre cada fuente de forma separada.

Con respecto a los resultados obtenidos para el dataset CNN-Daily Mail, LSTM proporcionaba mejores resultados para las métricas Rouge-1 (41.56 %) y Rouge-L (37.83 %), mientras que el modelo transformador obtenía los valores más altos para Rouge-2 (18.85 %). Respecto a la evaluación con las distintas fuentes del dataset Newsroom, LSTM obtiene mejores resultados para New York Times (25.27 % de media), Washington Post (18.89 % de media) y Fox News (56.17 %

de media). El modelo transformador supera al modelo LSTM en el resto de fuentes. Para todos los datasets, no hay una diferencia significativa entre los resultados de ambos modelos.

El sistema descrito en (Liu y Lapata, 2019) utiliza el modelo pre-entrenado BERT para la generación de resúmenes extractivos y abstractivos. La principal contribución del trabajo es que la codificación de los textos se hace a nivel de documento, en lugar de a nivel de oración, como habían hecho trabajos anteriores. La evaluación es realizada sobre los datasets CNN-Daily Mail (Rouge-1=42.13 %, Rouge-2=19.60 %, Rouge-L=39.18 %), NYT (Rouge-1=49.02 %, Rouge-2=31.02 %, Rouge-L=45.55 %) y XSum (Narayan, Cohen, y Lapata, 2018) (Rouge-1=38.81 %, Rouge-2=16.50 %, Rouge-L=31.27 %).

El sistema también fue evaluado por un conjunto de usuarios. Para la evaluación humana se les aporta a los participantes el resumen generado por el sistema y por otros sistemas de generación de resúmenes (Narayan, Cohen, y Lapata, 2018; See, Liu, y Manning, 2017; Gehrmann, Deng, y Rush, 2018). A continuación, se les realiza una serie de preguntas para comprobar si el resumen contiene toda la información necesaria. A mayor número de preguntas contestadas correctamente, mejor evaluación. La evaluación basada en el usuario también muestra que el sistema proporciona mejores resúmenes que el resto de modelos comparados en el trabajo.

Zhang, Wei, y Zhou (2019) proponen un sistema para la generación de resúmenes extractivos basado en el uso del modelo transformador jerárquico (HIBERT). En lugar de utilizar alguno de los modelos de lenguajes ya proporcionados por BERT (Devlin et al., 2019), los autores entrenan su propio modelo de lenguaje. Con este objetivo crean un dataset, GIGA-CM, compuesto por 6,626,842 documentos, la mayor parte de ellos obtenidos del dataset GigaWord (Graff et al., 2003) y una pequeña parte extraídos del conjunto de entrenamiento del dataset CNN/Daily Mail. En total, el dataset contiene 2,854 millones de palabras. Los autores utilizan una técnica de enmascaramiento similar a la propuesta por los creadores de BERT (Devlin et al., 2019), pero en este caso se enmascaran todas las palabras de las oraciones seleccionadas. Una vez entrenado su propio modelo de lenguaje, en una segunda fase dedicada a adaptar el modelo para la tarea de generación de resúmenes, los autores utilizan los datasets CNN/Daily Mail y New York Times (NYT) (Xu y Durrett, 2019; Durrett, Berg-Kirkpatrick, y Klein, 2016).

A la hora de realizar la evaluación, comparan su sistema con otros enfoques propuestos por otros autores, que también han sido evaluados sobre los mismos datasets. Los modelos extractivos (Gehrmann, Deng, y Rush, 2018; Dong et al., 2018; Cheng y Lapata, 2016) con los que se comparan están basados en el modelo transformador, pero en lugar de entrenar su propio modelo de lenguaje, han reutilizado alguno de los modelos de lenguajes proporcionados por BERT. Respecto a los modelos abstractivos (See, Liu, y Manning, 2017; Celikyilmaz et al., 2018) con los que se comparan, están basados en modelos secuenciales con refuerzo. Los experimentos mostraron que HIBERT es computacionalmente más eficiente que aquellos sistemas que utilizaban modelos de lenguajes pre-entrenados con BERT. Respecto a los resultados, HIBERT es capaz de superar a todos los modelos estudiados en ambos datasets. Las métricas para el dataset CNN/Daily Mail son Rouge-1 de 43.19 %, Rouge-2 de 20.46 % y Rouge-L de 39.72 %.

En el dataset NYT, los resultados son aún mejores (Rouge-1=49.47 %, Rouge-2=30.11 %, Rouge-L=41.63 %). Los autores también realizaron una evaluación basada en usuarios sobre una muestra de 20 documentos elegidos aleatoriamente, junto con los resúmenes generados por el modelo HIBERT y por los modelos comparados en este estudio. Los usuarios debían clasificar de peor a mejor los resúmenes generados, teniendo en cuenta si el resumen captaba la información esencial y si era correcto gramaticalmente. El resumen generado por el modelo HIBERT es seleccionado como el mejor en el 30 % de los casos.

En un trabajo posterior, Zhang et al. (2019) proponen, además de utilizar BERT en el codificador, utilizar el modelo transformador en el decodificador. Su decodificador está diseñado en dos fases. En la primera fase, genera un resumen borrador usando el algoritmo *left-context-only-decoder*, que únicamente tiene en cuenta el contexto izquierdo de cada palabra. La segunda fase tiene co-

mo objetivo refinar el resumen borrador generado en la fase anterior. Para ello, se aplica una máscara a cada palabra, conservando ahora sus respectivos contextos al completo. Además, los autores aplican el aprendizaje por refuerzo. Este tipo de aprendizaje pretende premiar a aquellas acciones óptimas y penalizar a las menos válidas. En este artículo, proponen como objetivo obtener la mayor evaluación posible mediante Rouge, es decir, a mayor puntuación mayor refuerzo positivo. El sistema es evaluado sobre el dataset CNN/Daily Mail y comparado con otros trabajos anteriores (See, Liu, y Manning, 2017; Shi et al., 2019; Chen y Bansal, 2018; Hsu et al., 2018). En el estudio, el sistema es capaz de obtener mejores resultados que dichos sistemas (Rouge-1=41.71 %, Rouge-2=19.49 % y Rouge-L=38.79 %).

Al igual que en el trabajo (Zhang et al., 2019), (Bae et al., 2019) también aplica aprendizaje por refuerzo para la selección de las oraciones más relevantes de un texto. Después de esta selección, los autores proponen una arquitectura que utiliza BERT en el codificador y una capa LSTM en el decodificador. En una segunda fase, las oraciones seleccionadas son parafraseadas mediante un modelo Sequence-to-Sequence con atención. El sistema fue evaluado sobre el dataset CNN/Daily Mail (Rouge-1=41.90 %, Rouge-2=19.08 %, Rouge-L=39.64 %), sin superar al trabajo descrito en (Zhang et al., 2019).

Los trabajos revisados muestran que los modelos transformadores han sido capaces de establecer un nuevo estado de la cuestión en la tarea de generación de resúmenes, superando a los resultados obtenidos por trabajos anteriores basados en otras arquitecturas de redes profundas (Nallapati, Zhai, y Zhou, 2017; Nallapati et al., 2016). En concreto, HIBERT (Zhang, Wei, y Zhou, 2019) es el sistema que ha obtenido mejores resultados sobre el dataset CNN/Daily Mail, con una media de Rouge de 34.5 %- El corpus CNN/Daily Mail es el dataset utilizado por la mayoría de los sistemas para su evaluación, aunque algunos trabajos también han utilizado otros datasets tales como New York Times o XSum. Todos los datasets tienen en común que son colecciones de artículos de prensa.

3 Enfoque

El objetivo de nuestro trabajo es explorar el uso del modelo BART (Lewis et al., 2020) para la tarea de generación automática de resúmenes.

BART es un modelo basado en la arquitectura Sequence-to-Sequence (Seq2Seq), cuyo principal objetivo es la transformación de una secuencia de entrada en otra secuencia de salida. La arquitectura Seq2Seq es especialmente útil en muchas tareas de PLN, como la traducción automática o la generación de resúmenes. En ambas tareas, la entrada es una secuencia de palabras. En el caso de la traducción automática, la secuencia de salida será la traducción de la secuencia de entrada a otro idioma. En el caso de la generación de resúmenes, la salida será una secuencia de palabras que constituya el resumen de la secuencia de entrada.

Los dos principales componentes de una arquitectura Seq2Seq son el codificador (encoder) y el decodificador (decoder). El primero toma la secuencia de entrada y la transforma en un vector. El decodificador transforma dicho vector en la secuencia de salida. Tradicionalmente, estos componentes han sido implementados con redes recurrentes, y en particular, con LSTM. Estas redes procesan la entrada de forma secuencial. Así, para acceder a la célula que representa la última palabra, es necesario recorrer las anteriores. Cuando la secuencia es muy larga, puede ocurrir que el modelo olvide la información contenida en las primeras células. Además, el procesamiento secuencial implica un alto coste computacional y no permite paralelizar el aprendizaje del modelo.

El modelo transformador, presentado por Vaswani et al. (2017), es una alternativa a las redes recurrentes para implementar el codificador y el decodificador de una arquitectura Seq2Seq. Un transformador está basado en la técnica de atención (attention mechanism), que permite establecer las dependencias entre las secuencias de entrada y de salida. El codificador se encarga de representar cada posición y aplicar el mecanismo de atención para conectar palabras que no son consecutivas. El mecanismo de atención asigna una puntuación a cada palabra y compara todas las puntuaciones en la secuencia. Esto permite determinar la contribución de cada una de las palabras. Este mecanismo puede ser paralelizado, acelerando así el aprendizaje. Por tanto, esta técnica es capaz de decidir qué partes de la secuencia de entrada son las más importantes en cada paso. Mientras que

el módulo LSTM lee la entrada de forma secuencial, la principal ventaja del mecanismo de atención es que es capaz de procesar de una sola vez el contexto de cada palabra y utilizarlo para asignar más peso a las partes de la secuencia de entrada más importantes. De esta forma, el decodificador sabe identificar qué partes de la secuencia son más importantes.

Como se ha dicho anteriormente, BART sigue una arquitectura Seq2Seq basada en transformadores. BART combina dos arquitecturas como BERT y GPT (Generative Pre-trained Transformer) (Radford et al., 2018). El codificador de BART implementa un modelo BERT, mientras que el decodificador implementa un modelo GPT. En la fase de codificación, BART primero introduce ruido en la secuencia de entrada. Para ello utiliza distintas técnicas como el enmascaramiento de palabras y de secuencias de palabras, el borrado de palabras o la permutación de oraciones, entre otras. Una vez que la secuencia de entrada ha sido transformada en una secuencia con ruido, BART trata de entrenar un modelo capaz de reconstruir la secuencia de entrada. En la fase de decodificación, BART utiliza un modelo GPT que genera los tokens de izquierda a derecha.

El modelo BART está implementado en la librería Hugging Face.[1] Este modelo ha sido preentrenado sobre el dataset CNN/Daily Mail. Con el objetivo de facilitar la replicabilidad de nuestra experimentación, nuestra implementación está disponible en el siguiente repositorio de https://github.com/isegura/sepln2021-textsummarization.

4 Evaluación

4.1 Datasets

Como se ha visto en el estado de la cuestión, la mayoría de los sistemas de generación de resúmenes han sido evaluados sobre el corpus CNN/Daily Mail. Uno de los objetivos del trabajo actual es extender la evaluación a otros datasets menos utilizados, e incluir además un nuevo dataset, como es BioMRC(Pappas et al., 2020a), cuyos textos no son artículos de periódicos, sino documentos científicos. Esto nos permitirá determinar si el modelo, que ha sido pre-entrenado con otro modelo de lenguaje generado con textos periodísticos, es capaz de obtener resultados similares sobre textos de otros dominios.

Por tanto, para evaluar nuestro modelo, los experimentos han sido realizados sobre cuatro conjuntos de datos: CNN/Daily Mail (See, Liu, y Manning, 2017) (versión no anonimizada), GigaWord (Rush, Chopra, y Weston, 2015), XSum (Narayan, Cohen, y Lapata, 2018) y BioMRC (Pappas et al., 2020b). Mientras que los tres primeros datasets constan de artículos periodísticos online, el último, BioMRC, es una colección de textos del dominio biomédico. Los cuatro conjuntos de datos están formados por textos escritos en inglés. Todos los resúmenes son abstractivos, es decir, cada resumen se ha creado a partir de oraciones nuevas y no extraídas del texto original.

Los cuatro datasets son distribuidos con la librería Hugging Face, que también usaremos para implementar nuestro enfoque. Para el dataset BioMRC, Hugging Face proporciona tres versiones dependiendo del tamaño: *large*, *small* y *tiny*. Dentro de cada tamaño, se distinguen dos tipos: A y B. La principal diferencia es que los textos de B se han limpiado para eliminar ruido. En nuestro trabajo, hemos seleccionado la versión *large_B*.

La tabla 1 muestra el número de textos de cada conjunto de datos y el tamaño medio de sus textos de entrada y sus resúmenes, que se define por el número de tokens de un texto. GigaWord, una colección de artículos periodísticos, es el dataset con mayor número de instancias. Sin embargo, su tamaño medio del texto de entrada y del resumen es el segundo más pequeño (poco más de 8 palabras por resumen). La relación entre el tamaño del texto y su resumen es aproximadamente de 3.7. El segundo dataset con mayor número de instancias es BioMRC, formado por textos del dominio biomédico. Aunque es el segundo corpus con un tamaño medio de los textos de entrada, el tamaño medio de sus resúmenes es el más pequeño, con menos de 7 palabras por resumen. En este caso, el ratio entre el tamaño medio de sus textos y sus resúmenes es de 37.7.

El tercer dataset con mayor número de instancias es CNN/Daily Mail, que como se dijo anteriormente es uno de los datasets más utilizados en la evaluación de generación de resúmenes. Sus textos y resúmenes son los que tienen un mayor tamaño, con una ratio de 13.9. La principal ventaja de este data-

[1] https://huggingface.co/

set respecto al resto, es que sus textos han sido utilizados para entrenar el modelo del lenguaje utilizado en la implementación del modelo BART, que usamos en la experimentación, y que es proporcionado por Huggging Face. XSum es el dataset con menor número de instancias. Sin embargo, también es el dataset con el segundo tamaño medio de longitud de textos de entrada. El ratio entre el tamaño medio de sus textos y resúmenes es de 35.9.

Los cuatro datasets se dividen en conjuntos de entrenamiento, validación y test.

En nuestra experimentación, hemos utilizado la configuración que por defecto proporciona la librería Hugging Face, por tanto, no ha sido necesario utilizar el conjunto de validación para ajustar los parámetros de nuestro modelo. Por este motivo, los textos del conjunto de validación también han sido incluidos en el conjunto de entrenamiento de cada dataset para ajustar el modelo a las tareas de generación de resúmenes.

4.2 Resultados

El método propuesto ha sido evaluado en diversas variantes de la métrica Rouge (Lin, 2004) y con la métrica Bleu (Papineni et al., 2002) para todos los datasets descritos en el apartado 4.1.

La métrica Rouge es la más utilizada para evaluar la generación automática de resúmenes. Podemos encontrar 5 variantes: Rouge-N, Rouge-L, Rouge-W, Rouge-S y Rouge-Su. En este trabajo, sólo utilizaremos las métricas Rouge-N (en particular, para N=1 y N=2) y Rouge-L, porque son las métricas reportadas por la mayoría de los artículos.

Rouge-N mide la superposición de N-gramas entre el resumen generado y el resumen del dataset. Así para N=1, la métrica se refiere a la superposición de palabras, mientras que para N=2, nos referimos a la superposición de bigramas. Rouge-L es una métrica que mide la subsecuencia común más larga (LCS) entre el resumen original y el resumen generado por el sistema. Una descripción detallada de estas métricas y del resto de variantes de Rouge puede encontrarse en el artículo propuesto por Lin (2004).

Como se ha indicado anteriormente, también utilizaremos la métrica Bleu (Papineni et al., 2002). Bleu mide el número de n-gramas del resumen automático que están presentes en el resumen de referencia, mientras que Rouge contabilizael número de n-gramas del resumen de referencia que están presentes en el resumen automático. Bleu similar a Rouge, pero su principal diferencia es que introduce un nuevo parámetro, denominado penalización por brevedad (PB), encargado de sancionar aquellas predicciones cuya longitud sea menor que la del resumen original. La penalización por brevedad se calcula a través de la siguiente función:

$$PB = \begin{cases} 1 & si \quad c > r \\ e^{(1-r/c)} & si \quad c \leq r \end{cases} \tag{1}$$

En la fórmula anterior, c es la longitud de la predicción y r la longitud del resumen de referencia. De esta manera, la métrica Bleu se calcula de la siguiente manera:

$$BLEU = PB \cdot exp\left(\sum_{n=1}^{N} w_n log P_n\right) \tag{2}$$

En este caso, $w_n = 1/N$, correspondiente al peso de cada n-grama de longitud N. P es la precisión de los n-gramas, es decir, es el ratio entre el número de n-gramas comunes y el número n-gramas presentes del resumen generado.

La tabla 2 muestra los resultados de nuestro modelo respecto a las métricas anteriormente descritas y para cada uno de los datasets propuestos para nuestro estudio.

Como se ha dicho anteriormente, nuestro enfoque está basado en el uso del modelo BART, proporcionado por la librería Hugging Face. En este modelo, el modelo del lenguaje ha sido pre-entrenado con los textos del dataset CNN/Daily Mail. Posteriormente, en la segunda fase (fine tuning), el modelo es ajustado para la tarea de generación de resúmenes, utilizando para ello los conjuntos de entrenamiento de cada dataset.

Comenzamos discutiendo los resultados para la métrica Rouge-1, que mide la superposición de las palabras del resumen generado y su respectivo resumen de referencia. Aunque el modelo de lenguaje ha sido construido a partir de los textos de CNN/Daily Mail, los mejores resultados para Rouge-1 se consiguen sobre el corpus BioMRC, siendo este el único dataset compuesto por textos no periodísticos, sino de dominio biomédico. Por tanto, el estilo narrativo y el léxico de los textos no parecen afectar al modelo en

Dataset	Conjunto de entrenamiento	Conjunto de validación	Conjunto de Test	Tamaño textos	Tamaño resúmenes
CNN/Daily	287.113	13.368	11.490	781	56
GigaWord	3.803.957	189.651	1.951	31,4	8,3
XSum	204.017	11.327	11.333	431	12
BioMRC	700.000	50.000	62.707	254,01	6,72

Tabla 1: Descripción de los datasets utilizados en este trabajo.

Datasets	Rouge-1	Rouge-2	Rouge-L	BLEU
CNN/Daily	38.48	**17.74**	**37.52**	31.35
GigaWord	24.93	9.88	24.02	37.32
XSum	32.2	9.7	27.11	**66.50**
BioMRC	**39.58**	13.08	34.28	57.73

Tabla 2: Resultados.

lo que se refiere a la superposición de palabras. El segundo mejor resultado, con una diferencia únicamente de 1.1, se obtiene para el dataset CNN/Daily Mail, con una Rouge-1 de 39.58 %, más de 3.5 por debajo del estado de la cuestión en dicho dataset (Rouge-1=43.19 % en el trabajo (Zhang, Wei, y Zhou, 2019)).

Para los otros dos datasets, GigaWord y XSum, los resultados de Rouge-1 son significativamente más bajos que los obtenidos para BioMRC. Es de esperar que los resultados obtenidos con el dataset CNN/Daily Mail sean superiores a la de estos datasets, debido a que el modelo de lenguaje fue entrenado utilizando el dataset CNN/Daily Mail. Sin embargo, llama la atención que el modelo obtenga la mejor Rouge-1 para la colección de textos biomédicos, y no de estilo periodístico. Un tamaño pequeño en los resúmenes de referencia (con sólo 6.5 palabras en el caso de BioMRC) parece tener un efecto positivo sobre la métrica Rouge-1. Otra posible razón de los buenos resultados obtenidos sobre BioMRC es que, como se explicó anteriormente, el dataset BioMRC ha sido limpiado para reducir ruido, y eso podría redundar en una mejor calidad de los textos y resúmenes de referencia.

Respecto a Rouge-2, que se refiere al número de bigramas que comparten el resumen generado con el resumen de referencia, como era previsible, obtiene los mejores para el dataset CNN/Daily Mail, por lo ya explicado anteriormente. Respecto al estado de la cuestión, el modelo consigue una puntuación menor que el sistema HIBERT, con una Rouge-2 de 20.46 %. La diferencia de Rouge-2 respecto a los otros tres datasets es muy significativa, llegando alcanzar una diferencia de hasta 8 puntos en el caso del dataset XSum. El modelo pre-entrenado con textos de estilo periodístico debería proporcionar mejores resultados para los corpus GigaWord y XSum, que para el corpus BioMRC. Sin embargo, la diferencia de Rouge-2 con la obtenida sobre BioMRC es sólo de 4 puntos, muy por debajo que las diferencias respecto a GigaWord y XSum.

Respecto a los resultados de XSum, con la Rouge-2 más baja de los 4, en un principio podríamos suponer que se debe a que es el conjunto de datos más pequeño de entrenamiento. Sin embargo, ese no parece ser el motivo real porque GigaWord, el dataset con un mayor número de ejemplos, muestra una Rouge-2 muy similar a la obtenida sobre XSum.

De manera general, en Rouge-N podemos apreciar que, a medida que aumenta el número de n-gramas, el resultado de la métrica disminuye a menos de la mitad para todos los datasets. Es previsible esperar que, al aumentar el tamaño de los n-gramas, el número de coincidencias entre el resumen generado y el de referencia sea menor. Como se muestra en la tabla 2, los resultados de Rouge-N sobre el corpus CNN/Daily Mail son considerablemente superiores que los obtenidos sobre los datasets GigaWords y XSum. Esto podría deberse a dos causas distintas: 1) el modelo del lenguaje usado en el enfoque fue preentrenado sobre CNN/Daily Mail, y 2) el tamaño medio de los textos y los resúmenes en CNN/Daily Mail es considerablemente mayor al de los otros datasets, lo que podría beneficiar a la hora de encontrar co-ocurrencias entre los resúmenes generados y los de refe-

rencia.

Con la métrica Rouge-L, que tiene en cuenta la subsecuencia común más larga (LCS) entre el resumen original y el resumen generado por el sistema, el modelo vuelve a obtener los mejores resultados sobre el dataset CNN/Daily Mail. Este dataset contiene el mayor tamaño medio del resumen referencia (56), y por tanto es más probable que las subsecuencias comunes sean más largas. También, como se puede comprobar, el tamaño medio de los textos es más del doble que en el resto de datasets. Como consecuencia, la extensión de los resúmenes propuestos será mayor. Al ser ambos resúmenes más extensos que en el resto de datasets, es de esperar que la subsecuencia común sea más larga. El modelo obtiene una Rouge-L de 37.52 %, 2.2 puntos por debajo del estado de la cuestión reportado en el artículo de Zhang, Wei, y Zhou (2019), con una Rouge-2 de 39.72 %. De nuevo, el modelo obtiene los mejores resultados para el dataset BioMRC, a pesar de pertenecer a un dominio distinto.

Como se explicó anteriormente, la métrica BLEU introduce un parámetro de penalización para aquellas predicciones más cortas que el resumen de referencia. El modelo consigue la mejor puntuación sobre el dataset XSum (Bleu=66.50), seguido de la puntuación obtenida en BioMRC (Bleu=57.73 %). Una posible causa podría estar relacionada con el ratio entre el tamaño medio de los textos y el tamaño medio de los resúmenes. Ambos datasets muestran un ratio considerablemente alto (por encima de 35), mientras que los otros datasets presentan ratios mucho menores. Así la relación entre el tamaño resumen del texto y del resumen es sólo de 3.9 en en GigaWord y 13.9 en CNN/Daily Mail.

En nuestra evaluación, también hemos querido estudiar el tiempo de entrenamiento para cada uno de los datasets, que son mostrados en la tabla 3. Como era de esperar, el tiempo de entrenamiento está directamente relacionado con el tamaño del corpus. Así, el dataset GigaWord tiene un tiempo de entrenamiento mucho mayor que el resto de datasets, porque su tamaño también es mucho mayor. Para una colección de casi 4 millones de instancias de entrenamiento, que su tiempo de entrenamiento sea de poco más de una hora parece bastante razonable y demuestra que los modelos transformers son más eficientes en relación a otras arquitecturas profundas como las RNN. Por la misma razón, XSum, el dataset más pequeño, requiere sólo 8 minutos de entrenamiento. En todos los datasets, el tiempo medio para generar un nuevo resumen es aproximadamente de 15 segundos.

	Tiempo entrenamiento
CNN/Daily	14 minutos
GigaWord	1 hora y 13 minutos
XSum	8 minutos
BioMRC	23 minutos

Tabla 3: Tiempo de entrenamiento para cada conjunto de datos.

La tabla 4 muestra un ejemplo de resumen automático del corpus BioMRC, junto con su correspondiente texto original y resumen de referencia. Este corpus contiene un gran número de conceptos del dominio biomédico, que probablemente no están representados en el corpus CNN/Daily Mail, utilizado para pre-entrenar el modelo BART. A pesar de eso, podemos ver que el resumen generado es bastante similar al resumen de referencia, tanto en contenido semántico como en longitud. Por tanto, podemos concluir que el uso del modelo BART es capaz de generar resúmenes para textos del dominio biomédico.

La tabla 5 muestra un ejemplo tomado del corpus GigaWord. Como se muestra en la tabla 1, este es el corpus con el tamaño más pequeño de textos (un tamaño medio de 31.4 palabras) y de resúmenes, con sólo 8.3 palabras de media. Hemos observado que las predicciones generadas por el modelo automático suelen ser demasiado largas en comparación a los resúmenes de referencia, incluso a veces llegando a superar el tamaño del texto original. En el ejemplo mostrado, consideramos que el resumen generado automáticamente ofrece más información que el resumen original, aunque contiene algunas incoherencias y errores gramaticales (President's invitation of China's President Hu Jintao's).

En el caso del corpus XSum (tabla 6), el resumen automático es completamente correcto desde el punto de vista sintáctico, y también puede ser considerado correcto en cuanto a su significado, aunque difiera del contenido del resumen de referencia. Mientras que el resumen automático pone el foco en que la visita fue realizada por los duques de Cambridge, el resumen de referencia des-

Texto original
The case is reported of a sixty-four-year-old @entity0 with DeBakey type I aortic dissection in whom postoperative extensive intra-aortic balloon pumping was applied. Surgical repair involved replacing the ascending aorta with a Medtronic Hall valved conduit. After surgery severe @entity2 occurred. Despite the use of high-dose inotropic drugs the @entity0 could not be hemodynamically stabilized. An intra-aortic balloon pump was finally applied as a therapeutical last resort. Within three days, under counterpulsation, the @entity0 reached a stable hemodynamic condition. After twenty-one days in the intensive care unit, he could be transferred to a normal ward. The @entity0 was discharged on the fifty-fourth postoperative day. During counterpulsation there were no balloon- or catheter-induced complications. Follow-up at five months showed the @entity0 in good general health: echocardiography did not identify any lesions of the thoracic aorta which could be linked to counter-pulsation. It is concluded that the postoperative use of intra-aortic balloon pump in the event of DeBakey type I dissecting @entity1 of the aorta, and adversely affected @entity0 hemodynamics, is a justifiable therapeutical alternative.
Resumen original
The use of intra-aortic balloon pump after surgical treatment of DeBakey type I dissecting XXXX of the aorta.
Resumen automático
Postoperative use of intra-aortic balloon pumping in the event of DeBakey type I dissecting XXXX of the aorta.

Tabla 4: Ejemplo tomado del corpus BioMRC.

Texto original
US President George W. Gush arrived here saturday evening for a three-day visit to china at the invitation of chinese President Hu Jintao
Resumen original
US President arrives in Beijing
Resumen automático
Bush arrives in China for three-day visit at President's invitation of China's President Hu Jintao's

Tabla 5: Ejemplo tomado del corpus GigaWord.

Texto original
The White Garden, at Kensington Palace, was planted to mark 20 years since Princess Diana died in a car crash.The Duchess of Cambridge joined the princes on the garden tour... Members of the public have been leaving tributes and flowers at the gates of the palace to mark the anniversary of Diana's death. ...It is the fourth London memorial created in tribute to Diana - the others are the Diana Memorial Playground at Kensington Palace, the Diana Memorial Fountain in Hyde Park, and the Diana Memorial Walk at St James's Palace.
Resumen original
Prince William and Prince Harry have visited a London memorial garden for their mother on the eve of the 20th anniversary of her death.
Resumen automático
The Duke and Duchess of Cambridge have visited the White Garden at Kensington Palace to mark the 20th anniversary of Princess Diana's death.

Tabla 6: Ejemplo tomado del corpus XSum.

taca que la visita fue realizada por los príncipes, William y Harry. Ambos resúmenes son correctos.

5 Conclusiones

Este artículo presenta un trabajo preliminar donde exploramos el uso de un modelo trans-

formador para abordar la tarea de generación de resúmenes. Dicho modelo está basado en la arquitectura BART y su implementación es proporcionada por la librería Hugging Face. Mientras que la mayoría de los trabajos anteriores se han centrado en evaluar sus enfoques sobre el dataset CNN/Daily Mail, en nuestro estudio ampliamos la evaluación a cuatro datasets con distintas características. Uno de los datasets está formado por textos biomédicos, mientras que los otros tres son colecciones de artículos periodísticos. Además, los cuatro datasets presentan características distintas en relación al número de instancias y al tamaño medio de sus textos y resúmenes. El modelo ha sido evaluado por las métricas Rouge y Bleu.

Como era de esperar la mejor media de Rouge (31.2 %) sobre el dataset CNN/Daily Mail. La principal razón es porque el modelo de lenguaje usado ha sido pre-entrenado sobre esta misma colección de textos. Esta media es similar a la obtenida por otros trabajos previos basados también en transformadores(Bae et al., 2019; Zhang et al., 2019), pero no llega a superar el estado de la cuestión actual conseguido por el trabajo descrito en (Zhang, Wei, y Zhou, 2019).

Debemos destacar que la segunda mejor media (28.98) en Rouge es obtenida sobre el dataset BioMRC, compuesto por textos biomédicos. Esta media es significativamente superior a las obtenidas en los otros dos datasets, XSum y GigaWord, que sí están formados por artículos periodísticos como el dataset CNN/Daily Mail. Esto podría decirnos que el estilo narrativo y el vocabulario de los textos no parece asegurar unos mejores resultados. También nos permite concluir que el modelo BART puede ofrecer buenos resultados en la generación de resúmenes automáticos en dominios como el biomédico.

Un mayor número de instancias en el corpus de entrenamiento, como es el caso de GigaWord, no parece garantizar la obtención de mejores resultados. De hecho, el modelo obtiene la peor media sobre dicho dataset, a pesar de ser el dataset que tiene mayor número de textos. Por el contrario, la relación entre el tamaño de los textos y el de sus resúmenes podría estar afectando de alguna manera a los resultados, ya que se ha observado que a mayor ratio entre estos tamaños, también es mayor la puntuación obtenida en Rouge-L.

Como trabajo futuro, nos planteamos explorar otros enfoques recientes de transfer learning, tales como T5 (Raffel et al., 2020) y estudiar su adaptación a la tarea de generación de resúmenes. Además, debido a que encontrar las causas de por qué un modelo funciona mejor sobre un determinado dataset no siempre es una tarea trivial, como hemos visto en nuestra experimentación preliminar, también nos gustaría realizar una experimentación con usuarios. Este tipo de evaluación nos podría arrojar más luces sobre las ventajas y las desventajas de cada modelo, y la posible relación entre las características de un conjunto de datos y los resultados obtenidos por un modelo.

La investigación descrita en este artículo forma parte del proyecto NLP4Rare (NLP4Rare-CM-UC3M). El proyecto está financiado por la Comunidad de Madrid y la Universidad Carlos III de Madrid, que pretenden estimular la investigación de naturaleza interdisciplinar de jóvenes doctores. El proyecto NLP4Rare tiene como principal objetivo aumentar el conocimiento sobre las enfermedades raras mediante el uso de técnicas de PLN. Para ello, entre los objetivos específicos persigue la generación automática de resúmenes, tanto para médicos como pacientes, que faciliten la comprensión de toda la información publicada al respecto a una enfermedad o conjunto de enfermedades raras. Debido a que en la actualidad no existe ningún dataset para este dominio, otro de nuestros principales retos será la creación de una dataset para un dominio distinto al periodístico, como es el de las enfermedades raras. Nuestro objetivo es que incluya no sólo textos en inglés, sino también de otras lenguas como el español. Esto nos permitirá evaluar nuestros enfoques de generación de resúmenes para otros idiomas distintos al inglés.

Agradecimientos

Este trabajo ha sido financiado por el Programa de Investigación del Ministerio de Economía y Competitividad del Gobierno de España, (Proyecto DeepEMR TIN2017-87548-C2-1-R) y por el Programa para proyectos interdisciplinares para jóvenes doctores en la Universidad Carlos III de Madrid financiado por la Comunidad de Madrid (Proyecto NLP4Rare-CM-UC3M).

Bibliografía

Bae, S., T. Kim, J. Kim, y S.-g. Lee. 2019. Summary level training of sentence rewriting for abstractive summarization. En *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, páginas 10–20, Hong Kong, China, Noviembre. Association for Computational Linguistics.

Beloki, Z., X. Saralegi, K. Ceberio, y A. Corral. 2020. Grammatical error correction for basque through a seq2seq neural architecture and synthetic examples. *Procesamiento del Lenguaje Natural*, 65:13–20.

Celikyilmaz, A., A. Bosselut, X. He, y Y. Choi. 2018. Deep communicating agents for abstractive summarization. En *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, páginas 1662–1675, New Orleans, Louisiana, Junio. Association for Computational Linguistics.

Chen, Y.-C. y M. Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. En *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 675–686.

Cheng, J. y M. Lapata. 2016. Neural summarization by extracting sentences and words. En *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume1: Long Papers)*, páginas 484–494.

Colón-Ruiz, C., I. Segura-Bedmar, y P. Martínez. 2019. Análisis de sentimiento en el dominio salud: analizando comentarios sobre fármacos. *Proces. del Leng. Natural*, 63:15–22.

Devlin, J., M.-W. Chang, K. Lee, y K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. En *NAACL-HLT (1)*.

Dong, Y., Y. Shen, E. Crawford, H. van Hoof, y J. C. K. Cheung. 2018. BanditSum: Extractive summarization as a contextual bandit. En *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, páginas 3739–3748, Brussels, Belgium, Octubre-Noviembre. Association for Computational Linguistics.

Durrett, G., T. Berg-Kirkpatrick, y D. Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. En *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 1998–2008.

Gehrmann, S., Y. Deng, y A. Rush. 2018. Bottom-up abstractive summarization. En *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, páginas 4098–4109, Brussels, Belgium, Octubre-Noviembre. Association for Computational Linguistics.

Graff, D., J. Kong, K. Chen, y K. Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

Grusky, M., M. Naaman, y Y. Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. En *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, páginas 708–719.

Hermann, K. M., T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, y P. Blunsom. 2015. Teaching machines to read and comprehend. En *Advances in neural information processing systems*, páginas 1693–1701.

Hochreiter, S. y J. Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hsu, W.-T., C.-K. Lin, M.-Y. Lee, K. Min, J. Tang, y M. Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. En *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 132–141.

Lewis, M., Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, y L. Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. En *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, páginas 7871–7880, Online, Julio. Association for Computational Linguistics.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. En *Text Summarization Branches Out*, páginas 74–81, Barcelona, Spain, Julio. Association for Computational Linguistics.

Liu, Y. y M. Lapata. 2019. Text summarization with pretrained encoders. En *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, páginas 3730–3740, Hong Kong, China, Noviembre. Association for Computational Linguistics.

Miranda-Escalada, A. y I. Segura-Bedmar. 2020. One stage versus two stages deep learning approaches for the extraction of drug-drug interactions from texts. *Proces. del Leng. Natural*, 64:69–76.

Nallapati, R., F. Zhai, y B. Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. En S. P. Singh y S. Markovitch, editores, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, páginas 3075–3081. AAAI Press.

Nallapati, R., B. Zhou, C. dos Santos, Ç. GuÌ‡lçehre, y B. Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. En *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, páginas 280–290, Berlin, Germany, Agosto. Association for Computational Linguistics.

Narayan, S., S. B. Cohen, y M. Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. En *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, páginas 1797–1807, Brussels, Belgium, Octubre-Noviembre. Association for Computational Linguistics.

Papineni, K., S. Roukos, T. Ward, y W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. En *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, páginas 311–318.

Pappas, D., P. Stavropoulos, I. Androutsopoulos, y R. McDonald. 2020a. Biomrc: A dataset for biomedical machine reading comprehension. En *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, páginas 140–149.

Pappas, D., P. Stavropoulos, I. Androutsopoulos, y R. McDonald. 2020b. BioMRC: A dataset for biomedical machine reading comprehension. En *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, páginas 140–149, Online, Julio. Association for Computational Linguistics.

Poncelas, A., K. Sarasola, M. Dowling, A. Way, G. Labaka, y I. Alegria. 2019. Adapting NMT to caption translation inwikimedia commons for low-resource languages. *Proces. del Leng. Natural*, 63:33–40.

Radford, A., K. Narasimhan, T. Salimans, y I. Sutskever. 2018. Improving language understanding by generative pre-training.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, y P. J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Rush, A. M., S. Chopra, y J. Weston. 2015. A neural attention model for abstractive sentence summarization. En *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, páginas 379–389, Lisbon, Portugal, Septiembre. Association for Computational Linguistics.

See, A., P. J. Liu, y C. D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. En *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 1073–1083, Vancouver, Canada, Julio. Association for Computational Linguistics.

Shi, J., C. Liang, L. Hou, J. Li, Z. Liu, y H. Zhang. 2019. Deepchannel: Salience estimation by contrastive learning for extractive document summarization. En *Proceedings of the AAAI Conference on Artificial Intelligence*, volumen 33, páginas 6999–7006.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, y I. Polosukhin. 2017. Attention is all you need. En *Advances in neural information processing systems*, páginas 5998–6008.

Xu, J. y G. Durrett. 2019. Neural extractive text summarization with syntactic compression. En *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, páginas 3283–3294.

Zhang, H., J. Cai, J. Xu, y J. Wang. 2019. Pretraining-based natural language generation for text summarization. En *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, páginas 789–797, Hong Kong, China, Noviembre. Association for Computational Linguistics.

Zhang, X., F. Wei, y M. Zhou. 2019. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. En *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, páginas 5059–5069, Florence, Italy, Julio. Association for Computational Linguistics.

Zhang, Y., D. Li, Y. Wang, Y. Fang, y W. Xiao. 2019. Abstract text summarization with a convolutional seq2seq model. *Applied Sciences*, 9(8):1665.

Zhong, M., P. Liu, D. Wang, X. Qiu, y X. Huang. 2019. Searching for effective neural extractive summarization: What works and what's next. En *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, páginas 1049–1058, Florence, Italy, Julio. Association for Computational Linguistics.

Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 41-51 recibido 23-11-2020 revisado 24-01-2021 aceptado 01-02-2021

NECOS: An annotated corpus to identify constructive news comments in Spanish

NECOS: Un corpus anotado para identificar comentarios constructivos de noticias en español

Pilar López-Úbeda, Flor Miriam Plaza-del-Arco, Manuel Carlos Díaz-Galiano, M. Teresa Martín-Valdivia
Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{plubeda, fmplaza, mcdiaz, maite}@ujaen.es

Abstract: In this paper, we present the NEws and COmments in Spanish (NECOS) corpus, a collection of Spanish comments posted in response to newspaper articles. Following a robust annotation scheme, three annotators labeled the comments as constructive and non-constructive. The articles were published in the newspaper *El Mundo* between April 3rd and April 30th, 2018. The corpus is composed of a total of 10 news articles and 1,419 comments. Three annotators manually labeled NECOS with an average Cohen's kappa of 78.97. Our current focus is the study of constructiveness and the evaluation of the Spanish NECOS corpus. In order to address this goal, we propose a benchmark testing different machine learning systems based on Natural Language Processing: a traditional system and the novel Transformer-based models. Specifically, we compare multilingual models with a monolingual model trained on Spanish in order to highlight the need to create resources trained on a specific language. The monolingual model fine-tuning on NECOS obtain the best result by achieving a macro-average F_1 score of 77.24%.
Keywords: Corpora, constructiveness, Natural Language Processing, Transformer-based models.

Resumen: En este artículo presentamos un corpus de noticias y comentarios en español (NECOS). Estas noticias están publicadas en el periódico El Mundo en un período comprendido entre el 3 de abril y el 30 de abril de 2018. El corpus contiene un total de 10 noticias y 1.419 comentarios. Siguiendo un esquema de anotación, tres anotadores etiquetaron manualmente los comentarios como constructivos y no constructivos obteniendo un promedio de 78,97 usando el coeficiente de kappa de Cohen. En este estudio nos centramos en estudiar la constructividad y hacer la evaluación del corpus NECOS. Para abordar este objetivo, proponemos la experimentación con diferentes sistemas basados en Procesamiento del Lenguaje Natural usando aprendizaje automático: un clasificador tradicional y métodos recientes basados en *Transformers*. Concretamente, comparamos modelos multilingües con un modelo monolingüe entrenado para el español. Con ello, pretendemos demostrar la importancia de crear recursos entrenados para un idioma en particular. El modelo monolingüe evaluado en NECOS obtiene el mejor resutlados alcanzando un 77,24% de *macro-average* F_1.
Palabras clave: Corpora, constructividad, Procesamiento del Lenguage Natural, modelos basados en *Transformer*.

ISSN 1989-7553. DOI 10.26342/2021-66-3

1 Introduction

Worldwide, millions of users have the possibility to easily access the web, publishing and sharing content on a variety of topics. For instance, popular newspapers, with the aim of keeping readers up to date, publish daily news from different areas such as politics, technology and socioeconomics, often receiving in a short period of time a large number of comments from users. Sometimes, this leads to inappropriate online content, as it can involve offensive, hateful, fake or non-constructive comments. Since 2007, The New York Times employed a team of full-time moderators to review all comments submitted on its website and to remove inappropriate content (Etim, 2017). However, reviewing these types of comments manually is a very time-consuming process for moderators because of the large number of comments posted. In order to deal with this issue, in the last years a great interest is growing in the Natural Language Processing (NLP) community in the development of systems capable of analyzing comments automatically(Kolhatkar et al., 2019).

In the NLP literature, most of the studies based on the constructiveness of comments identify which ones provide insight, contribute to the dialogue and encourage a healthy discussion (Kolhatkar and Taboada, 2017b; Kolhatkar and Taboada, 2017a). A number of approaches have been proposed for the identification of constructiveness, from traditional supervised systems applying text-based features (Kolhatkar and Taboada, 2017a; Fujita, Kobayashi, and Okumura, 2019) to the use of latest neural network models that are proving to be successful (Kolhatkar et al., 2020). To train these type of systems, linguistic resources such as corpora are essential. The systems use the corpus to generate a language model that allows the prediction of new data. However, there is a great shortage of datasets annotated for constructiveness of individual comments and most of them are in English. To the best of our knowledge, there is no such resource for Spanish which is the second most spoken language in the world and the third most used language on the Web (Instituto Cervantes, 2018).

In this paper, we help bridge this gap by presenting the first Spanish corpus from newspaper comments in order to identify constructive and non-constructive comments. In addition, we propose a benchmark by applying the most advanced approaches in the scope of NLP and machine learning.

In particular, the contributions of this paper are summarized below:

- We release the first Spanish dataset of constructive and non-constructive comments by collecting a number of comments from the newspaper *El Mundo*[1].
- We analyze the constructive and non-constructive comments of the corpus introducing some linguistic statistics.
- We establish a benchmark for our corpus applying Transformer-based language models.
- We conduct an error analysis in order to understand the capabilities and the drawbacks of our system.

The rest of the paper is structured as follows: In Section 2 we present some previous studies related to constructiveness in comments on newspaper articles. Linguistic statistics related to the NECOS corpus are described in Section 3 presenting some statistics. The different machine learning approaches, the results obtained and the error analysis of the best system are shown in Section 4. Finally, conclusions and future work are presented in Section 5.

2 Related work

Constructive comments were initially defined by Napoles et al. (2017) in terms of ERICs (Engaging, Respectful, and/or Informative Conversations). The dataset used was the Yahoo News Annotated Comments Corpus (YNACC) which contains nearly 140k threads posted on Yahoo News articles in April 2016 (Napoles, Pappu, and Tetreault, 2017). In this study, they compared four approaches to classifying ERICs: a pipeline (Conditional Random Fields and binary classification), a linear classifier with linguistic and social features, an augmented pipeline that incorporates features from the linear model, and a Convolutional Neural Network (CNN). The best result obtained was using the augmented pipeline reaching 0.73 in average F1-score.

[1] `https://www.elmundo.es/`

Kolhatkar and Taboada (2017a) provided a new annotated corpus to the scientific community interested in analyzing constructive comments. They crawled 1,121 comments in English from 10 articles of the Globe and Mail news website covering a variety of subjects. In this study they proposed different training sets and their new corpus as a testing set. The two corpora taken into account in the training set were YNACC and the Argument Extraction Corpus (AEC) (Swanson, Ecker, and Walker, 2015). Finally, they carried out the experimentation using Bidirectional Long short-term memory (BiLSTM) architecture achieving 72.59% accuracy.

Following up on their previous research, Kolhatkar and Taboada (2017b) created a new training set in order to identify constructive comments in the news. In this case, the positive examples included in the training set were obtained from the New York Times (NYT) Picks using the NYT API[2], and for the negative examples they used the negative comments from the YNACC. Support Vector Machines (SVM) with features and a BiLSTM were employed for the experimental results. The combination of features included in SVM obtained the best results reaching a 0.84 of average F1-score.

More recently, Kolhatkar et al. (2020) presented a new corpus named Constructive Comments Corpus (C3) including constructive and non-constructive labels in English. They continue to combine training and test sets from different datasets in order to achieve the desired results. The approaches evaluated on this study were more recent including deep learning models (BiLSTM and CNN) and models based on Transformers (BERT) (Vaswani et al., 2017; Devlin et al., 2019). BERT obtained the best result (0.93 of average F1-score) using partitions for training and a test set using C3. In addition, they provided interesting results according to the features that characterize constructive comments.

Regarding languages other than English and following the previous study, Fujita, Kobayashi, and Okumura (2019) created a new dataset including 100K+ Japanese comments with constructiveness scores in collaboration with Yahoo! News. These scores were based on the number of annotators who labeled a comment as constructive. Since the dataset was composed by numerical ranking scores, they used Normalized Discounted Cumulative Gain (NDCG) as the primary evaluation measure. In order to perform a comparison of results they used different methods such as SVM, rankSVM and SVR.

As previous studies show, most of the corpora annotated with constructiveness are in English. However, for Spanish, as far as we know there are no corpora labeled with this phenomenon. For this reason, we contribute to the scientific NLP community by introducing and releasing the first dataset for constructiveness in Spanish named NECOS[3] (News and Comments in Spanish). In addition, we use our dataset to demonstrate the feasibility of implementing systems to automatically detect constructive comments in newspapers.

3 NECOS corpus

In this section, we present NECOS, a corpus of online news comments enriched with constructiveness annotations.

3.1 Data collection

In order to study constructiveness in news comments, we crawled 1,419 comments from 10 articles from the *El Mundo* newspaper. The news articles that were downloaded are dated between April 3rd and April 30th, 2018. These news articles encourage debate and controversy among users and are related to various topics such as politics, assassinations, current affairs, among others.

We selected 10 news items from the newspaper and downloaded up to 150 random comments from these items, although some selected news articles contained less than 150 comments.

3.2 Annotating constructiveness

Based on previous studies based on constructive comments, we followed the same annotation guide to annotate the NECOS corpus (Kolhatkar et al., 2020). Bellow, we point out some characteristics and attributes of constructive and non-constructive comments:

- Constructive:

[2] `https://developer.nytimes.com/`

[3] NECOS corpus: `https://github.com/plubeda/NECOS`

  - Comments which add something substantial to the conversation and encourage dialogue.
  - Comments which are supported by appropriate evidence.
  - Comments which offer solutions, new perspectives and insights.
  - Comments which provide a personal story or experience.

- Non-constructive:
  - Comments which do not have much content or are unsubstantial.
  - Comments which are not civilized.
  - Comments which do not respect the views and beliefs of others.
  - Comments which express opinions without providing evidence.
  - Sarcastic comments.
  - Provocative comments.

The corpus is annotated in a binary form: label 1 has been assigned to constructive comments and label 0 for non-constructive comments. In order to understand the corpus, different comments are shown in Table 1.

As we can see in examples 1 and 2, the comments are constructive because the authors recommend possible solutions. Comment number 3 is not considered constructive because it provides a personal opinion without adding possible solutions or evidence. The author of comment 4 does not provide evidence and the comment does not contain much content.

Annotators should not evaluate comments from a personal point of view. They can agree or disagree with the comment, but this should not affect their labeling of the corpus.

3.3 Inter-annotator agreement

The annotation of the entire corpus was done by three annotators in an average time of 4-5 minutes per comment. These annotators have followed the previously mentioned annotation guidelines.

In binary classification tasks, Cohen's kappa is often used as a quality measure for data annotations because this metric expresses the level of agreement between two annotators on a classification (Cohen, 1960).

In order to measure the level of agreement among the three annotators, we measure the agreement by pairs of annotators using the kappa coefficient and finally we calculate the average to obtain the final value. Table 2 shows the agreement between each pair of annotators and the percentage of coincidence between them.

On the one hand, in this table we can see that the average kappa is 78.97, which corresponds to a moderate level of interpretation (McHugh, 2012). On the other hand, the percentage of agreement ranges from 89.92 to 91.9 and finally obtains an average of 91.03. Both metrics get acceptable values for annotation and agreement in the NECOS corpus.

It is important to emphasize that the most disagreement in the annotation of comments has been on issues of justice, economy and gender violence, where the annotators disagreed on more than 20% of the comments. In contrast, the highest level of agreement reached by annotators has been on political issues. To clarify the disagreement with further analysis, Table 3 shows the number of disagreement annotations for each news item in the NECOS corpus.

3.4 Corpus analysis

In this section we highlight some statistics regarding the NECOS corpus. These statistics refer to the number of comments, number of words and number of sentences in each annotated label.

Table 4 shows some basic linguistic statistics of the corpus. According to the number of comments of each class, we found that the corpus is unbalanced since it contains 985 non-constructive and 434 constructive comments. As the corpus contains more non-constructive comments, the number of total sentences is also higher in this class. On the other hand, although there are more non-constructive comments, the size of the vocabulary is almost the same. Finally, we want to highlight that constructive comments contain more words and sentences than non-constructive comments, which means that people in constructive comments use more words to express their opinion.

Finally, we have analyzed the same linguistic statistics according to the 10 news items downloaded for the NECOS corpus (see the 10 news items in Table 5). These statistics have been shown in Table 6. As we can see, in all the news items there are more constructive than non-constructive comments except in the item number 6. In addition, the constructive comments contain a higher average number of tokens than the

	Comment	Label
1	*Espero que no se aprueben los presupuestos, que se convoquen elecciones y gane ciudadanos y acabe con el cupo y el PP. Asco de partido apoyando a los fascistas nacionalistas vascos por nada.* I hope that the budgets will not be approved, that elections will be called and *Ciudadanos* will win and that the quota and the PP will be ended. A disgusting party that supports the Basque nationalist fascists for nothing.	1
2	*Trump cometió una equivocación porque debería haber dejado que el ejército soltara toda su fuerza para que de una vez el mundo sepa que con los EEUU no se juega.* Trump made a mistake because he should have let the army release its full force so that the world would finally know that the U.S. is not to be trifled with.	1
3	*España como criadero de incompetentes y chorizos. Menos mal que soy catalán.* Spain is a breeding ground for incompetents and crooks. Luckily I am Catalan.	0
4	*Eso es una compra de votos. Es delito. Malversación de fondos públicos.* That's vote buying. It is a crime. Embezzlement of public funds.	0

Table 1: Examples of comments tagged in the NECOS corpus, along with English translations.

	Cohen's kappa	Agreement (%)
Annotator 1 and 2	80.90	91.90
Annotator 2 and 3	76.49	89.92
Annotator 1 and 3	79.53	91.26
Average	78.97	91.03

Table 2: Inter-annotation agreement in the NECOS corpus.

non-constructive ones, which means an increase in the size of the vocabulary in some cases.

4 Experiments and results

In this section we introduce a benchmark for the constructive corpus. In particular, we propose two different approaches: transformer-based methods and a baseline system based on traditional machine learning.

4.1 Constructive classification

First of all, the pre-processing is a fundamental step in NLP. In this process, we prepare and clean up the text before including it in the classification systems. This step is one of the most important because it can help to improve the performance of the classifier and speed up the classification process. The pre-processing addressed in the NECOS corpus was carried out as follows:

- Remove references: references to other comments have been removed in the comments.
- Remove URLs: the existing URLs in the comments have been replaced by the token *URL*.
- Lowercase: the comments have been converted to lowercase.

After performing the pre-processing step, we carry out the experiments on NECOS corpus. In all of our experiments, we use 10-fold cross validation to evaluate the machine learning classification systems.

The models we have chosen to test the effectiveness of NECOS are described below:

News item	# of comments	# of comments disagreeing	Disagreement (%)
1	149	9	6.04
2	148	33	22.30
3	150	13	8.67
4	149	36	24.16
5	149	10	6.71
6	142	12	8.45
7	150	8	5.33
8	128	1	0.78
9	150	26	17.33
10	104	22	21.15
Total	1,419	170	11.98%

Table 3: Analysis of disagreement comments for each news item in the NECOS corpus (see the 10 news items in Table 5).

	Constructive	Non-constructive
Number of comments	434	985
Vocabulary size	6,734	6,367
Avg. of tokens in comments	84.89	32.73
Number of sentences in comments	1,552	1,997
Avg. of sentence in comments	3.57	2.03

Table 4: Dataset analysis.

- **SVM** is a set of supervised learning methods used for classification, regression and outlier detection. (Pedregosa et al., 2011). A number of studies have reported that this classifier is one of the most accurate methods for text classification (Puri and Singh, 2019; Chatterjee, Jose, and Datta, 2019). Therefore, we use this classifier as our baseline for the constructive corpus. For text representation, we use the Term frequency-inverse document frequency (Tf-idf). For the classifier, we use the default configuration provided in the scikit-learn module from Python.
- **Multilingual BERT (aka mBERT)** (Devlin et al., 2019). mBERT is a multilingual model that follows the same model architecture and training procedure as BERT, except with data from Wikipedia in 104 languages. In mBERT, the WordPiece modeling strategy allows the model to share embeddings across languages. In particular, for this study we chose the BERT-Base, Multilingual Cased checkpoint[4].
- **XLM** is a cross-lingual language model pre-trained, which uses a pre-processing technique and a dual language training mechanism with BERT in order to learn the relations between words in several languages. In this study we use the XLM model trained on 100 languages (XLM-100) (Lample and Conneau, 2019).
- **XLM-Roberta** proposed by Conneau et al. (2019). It is based on Facebook's RoBERTa model released in 2019 (Liu et al., 2019). XLM-Roberta is a large multilingual language model, trained from CommonCraws data on 100 different languages.
- **BETO** is a BERT-based language model pre-trained specifically on Spanish data and is similar in size to a BERT model. It has 12 self-attention layers with 16 attention-heads each (Vaswani et al., 2017), using 1024 as hidden size. The model is trained from different sources including Wikipedia and all of the sources of the OPUS Project (Aulamo and Tiedemann, 2019). Specif-

[4] `https://github.com/huggingface/transformers`

	News headline
#1	*Presupuestos 2018: el Gobierno ofrece un 32% más de inversión en el País Vasco tras mejorarle el Cupo.* Budgets 2018: the Government offers 32% more investment in the Basque Country after improving its quota.
#2	*El juez desoye a la Fiscalía y deja libre al informático Falciani, detenido ayer en Madrid a petición de Suiza.* The judge disregards the Prosecutor's Office and releases Falciani, who was arrested yesterday in Madrid at the request of Switzerland.
#3	*El Rey apoya en Barcelona a los jueces como "garantía de los derechos y el respeto a la ley".* The King supports judges in Barcelona as a "guarantee of rights and respect for the law"
#4	*Cristina Cifuentes: "No he cometido ilegalidad, no he mentido y no voy a dimitir".* Cristina Cifuentes: "I have not committed any illegality, I have not lied and I am not going to resign".
#5	Hombres víctimas de la violencia machista, el otro eslabón del maltrato a la mujer. Male victims of male violence, the other link in the chain of abuse of women.
#6	*Trump planeaba un ataque de mayor envergadura pero lo limitó para evitar una confrontación con Rusia.* Trump had planned a larger attack but limited it to avoid a confrontation with Russia.
#7	*El sargento agredido en Alsasua: "Algunos jaleaban y había bastantes móviles grabando".* The sergeant assaulted in Alsasua: "Some people were cheering and there were a lot of mobile phones recording".
#8	*Apagón y silencio absoluto en Podemos ante la gravedad de la crisis interna con Bescansa y Errejón.* Blackout and absolute silence in Podemos in the face of the seriousness of the internal crisis with Bescansa and Errejón.
#9	*La Policía requisa camisetas amarillas a la entrada del Wanda Metropolitano.* Police seize yellow T-shirts at the entrance to the Wanda Metropolitano.
#10	*Detenido un hombre por matar a su ex pareja de una paliza en Burgos.* Man arrested for beating ex-partner to death in Burgos.

Table 5: News headlines from the NECOS corpus.

	Constructive					**Non-constructive**				
	# Comments	Vocab. size	Avg. tokens	# Sents	Avg. sent	# Comments	Vocab size	Avg. tokens	# Sents	Avg. sent
#1	33	1005	87.73	146	4.42	116	1239	35.73	274	2.36
#2	46	1167	76.54	143	3.11	102	958	28.39	196	1.92
#3	33	929	81.03	118	3.58	117	1144	28.22	224	1.91
#4	38	1125	92.61	128	3.37	111	1206	31.15	207	1.86
#5	80	1861	88.31	289	3.61	69	922	35.09	129	1.87
#6	67	2055	103.21	294	4.39	75	1141	39.17	165	2.2
#7	42	1111	77.4	124	2.95	108	1183	31.52	225	2.08
#8	7	276	78.71	28	4	121	1403	36.94	242	2
#9	48	1111	73.08	155	3.23	102	1047	30.14	210	2.06
#10	40	953	73.58	127	3.18	64	731	33.25	125	1.95

Table 6: Dataset analysis by news topic.

ically, in this study we use the BETO cased checkpoint[5].

For all the pre-trained language models we use the same hyperparameters. Specifically, the models are fine-tuning using 2 epochs, a batch size of 16 and a learning rate of 0.0001. We use 256 words as max length.

4.2 Results

In this section we report and discuss the performance of the tested systems on the Spanish constructive classification task. In order to evaluate and compare the results obtained by our systems, we use the usual metrics in

[5] https://github.com/dccuchile/beto

text classification, called precision (P), recall (R), F-score (F_1) and macro-average. The metrics have been computed as follows:

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

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

where c is equal to the class (0, 1), TP = True Positive, FP = False Positive and FN = False Negative.

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

Table 7 shows the results achieved after experimenting with all the systems. In first place, it can be seen the performance of our baseline model using the SVM classifier. It obtains a macro-avg F1 score of 71.09% which is more than acceptable for a binary classification task. If we focus on the performance of each class, it is worth mentioning that the system is able to detect non-constructive comments more accurately than constructive ones.

Related to the pre-trained language models, XLM-Roberta and XLM, they do not perform as well as SVM. However, mBERT and BETO, which are based on the BERT model, outperform our baseline by a substantial margin. Specifically, BETO is the most accurate system outperforming the rest of the models by achieving a macro-avg F_1 score of 77.24%. We presume that BETO performs significantly better because it was trained on a Spanish corpus. For this reason, we point out the importance of generating resources not only for English but also for other languages such as Spanish.

If we observe the F_1 score of each system, it should be noted that all the systems achieve a better performance in the non-constructive class. On the contrary, there is a significant difference compared to the F_1 score of the constructive class. This could be because there is a higher proportion of non-constructive comments and therefore the system learns better to identify that class.

4.3 Error analysis

The goal of this section is to identify the weaknesses of our systems by conducting an error analysis of the Transformer-based models. For this purpose, we first analyze the errors conducted by the three systems and then focus in more detail on the errors produced by the best system BETO.

The results showing the numbers of wrongly assigned labels for each system are summarized in Table 8. All four models predicted the same wrong labels 91 times. As can be seen, XLM-based models predict more false negatives than BERT-based models. The common errors are highly biased towards false negatives. We have observed that the average number of tokens in the constructive comments mislabelled at the same time by the systems is much lower than the average shown in Table 4, namely 41.12 versus 84.89. Therefore, this may be one reason why systems do not correctly classify this type of comments.

In order to focus on the best system, we performed a manual analysis of some of the mislabelled comments to find the main reasons why BETO does not classify some comments correctly. Table 9 presents some examples of comments incorrectly classified by our system. In particular, there are 4 comments, two false negatives and two false positives. On the one hand, if we look at the false negatives (examples 1 and 2), the reason they are predicted as non-constructive is because BETO was not able to identify the possible solutions that the authors offer in the comments. On the other hand, if we focus on false positives, we consider that examples 3 and 4 are mislabeled by BETO because it identifies an argument, but in this case the author does not add new perspectives or substantial justifications.

Considering the number of false positives and false negatives, and given that we performed 10-fold cross validation, we decided to perform an analysis to determine the number of errors in each partition. Figure 1 shows the percentage of false positives (blue color), false negatives (red color) for each fold of the cross validation. In most of the partitions, we can observe that there are more false negatives than false positives. This problem may be due to the fact that the number of comments that exist in the corpus labeled as constructive is greater than the number of non-constructive. These results are also reflected in Section 4.2 showing a lower value of F_1 in the constructive category.

System	Non-constructive			Constructive			Macro-avg		
	P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
SVM	80.13	90.05	84.80	68.59	49.31	57.37	74.36	69.68	71.09
XLM-Roberta	75.91	93.58	82.62	34.63	28.04	28.76	55.27	60.81	55.69
XLM	76.62	**93.98**	83.65	43.36	33.85	37.06	59.99	63.62	60.36
mBERT	85.55	83.85	84.27	65.60	66.94	64.98	75.58	75.40	74.61
BETO	**86.70**	85.28	**85.59**	**68.48**	**71.80**	**68.90**	**77.59**	**78.54**	**77.24**

Table 7: Results obtained by different systems on the constructive corpus (10-fold cross validation).

System	Errors	Predicted 1	Predicted 0
XLM-Roberta	378	64	314
XLM	351	57	294
mBERT	294	156	138
BETO	268	143	125
All (in common)	91	6	85

Table 8: Number of instances mislabeled by each system, broken down by wrongly assigned label.

5 Conclusion

We have described the News and Comments in Spanish (NECOS) corpus, a resource for exploring constructiveness in news comments from the newspaper *El Mundo*. Our raw corpus comprises 10 news articles and 1,419 comment threads in response to these articles in April 2018.

Since our resource is manually annotated, we have noticed the difficulty of labeling constructive comments regardless of the annotator personal opinion. Despite this challenge, we have achieved a strong agreement among the annotators, reaching a Cohen's kappa of 78.97. While annotating, on the one hand, we learned that most of the constructive comments are based on offering solutions, new perspectives and insights, and they also provide a personal experience. On the other hand, we noticed that a great amount of non-constructive comments express opinions without providing evidence.

In order to check the effectiveness of our corpus and established a benchmark, we carry out several experiments based on machine learning approaches. Specifically, we use a traditional system (SVM) selecting it as a baseline to compare its results with the latest approaches in NLP including Transformer-based models. Given our results, we conclude that the application of text mining classification systems is a valuable tool for detecting constructiveness in comments from newspapers. Specifically, the best results are obtained using a monolingual Transformer-based model called BETO. With this model we achieved a macro-F_1 score of 77.24 which proves the importance of developing resources for a specific language, in this case Spanish.

A number of research avenues are planned for this corpus. We are interested in studying the relation between constructiveness and toxic language. We believe that this feature could be useful to help the model to easily detect constructive and non-constructive comments. In addition, sentiment analysis could be applied to study the influence of emotions in constructive comments. Finally, we plan to integrate these features into our best system with the purpose of further improving the results.

Acknowledgements

This work has been partially supported by a grant from European Regional Development Fund (ERDF), LIVING-LANG project [RTI2018-094653-B-C21], and the Ministry of Science, Innovation and Universities (scholarship [FPI-PRE2019-089310]) from the Spanish Government.

	Comment	Label	BETO
#1	*La violencia debe ser igualmente reprobable siempre, sea cual sea el sexo del agresor y del agredido. El trato preferente presente es tan aberrante como la discriminación pasada.* Violence must be equally reprehensible at all times, regardless of the sex of the aggressor and of the victim. The present preferential treatment is as bad as past discrimination.	1	0
#2	*Ánimo, pide ayuda médica y verás como dentro de un tiempo verás el suicidio como un disparate, pero lo importante es que cuando uno mismo no es capaz de salir del pozo, pida ayuda y la acepte. Todo es cuestión de paciencia y tiempo.* Come on, ask for medical help and you will see that in a while you will see suicide as nonsense, but the important thing is that when you are not able to get out of the well yourself, ask for help and accept it. It's all a matter of patience and time.	1	0
#3	*Estos sucesos son abominables y su erradicación muy complicada, no vamos a evitar que nazcan criminales en potencia, si al menos este salvaje tuviera el castigo que merece...* These events are abominable and their eradication very complicated, we are not going to prevent potential criminals from being born, if only this savage had the punishment he deserves...	0	1
#4	*En humoristas, titiriteros, raperos, actores, presentadores, activistas, que estén imputados o en la cárcel por expresar sus ideas. Y no se confunda, yo no apoyo lo que dicen, defiendo que puedan decirlo, (casi todo).* In comedians, puppeteers, rappers, actors, presenters, activists, who are accused or in prison for expressing their ideas. And don't be confused, I don't support what they say, I defend their right to say it, (almost everything).	0	1

Table 9: Examples of mislabeled comments using BETO model, along with English translations.

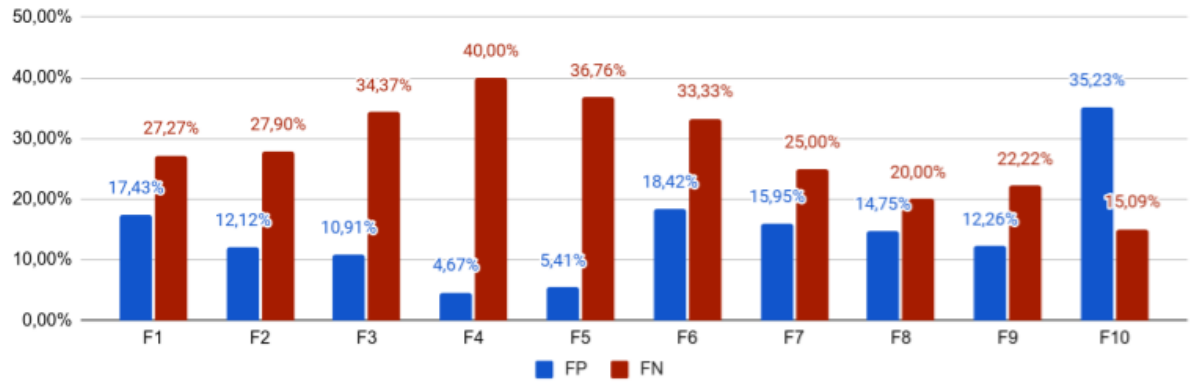

Figure 1: Percentage of false positives and false negatives for each 10-fold cross validation. F: fold.

References

Aulamo, M. and J. Tiedemann. 2019. The OPUS resource repository: An open package for creating parallel corpora and machine translation services. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 389–394, Turku, Finland, September–October. Linköping University Electronic Press.

Chatterjee, S., P. G. Jose, and D. Datta.

2019. Text classification using svm enhanced by multithreading and cuda. *International Journal of Modern Education & Computer Science*, 11(1).

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Etim, B. 2017. The times sharply increases articles open for comments, using google's technology. *The New York Times*, 13.

Fujita, S., H. Kobayashi, and M. Okumura. 2019. Dataset creation for ranking constructive news comments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2619–2626.

Instituto Cervantes. 2018. El español: una lengua viva. `https://cvc.cervantes.es/lengua/espanol_lengua_viva/pdf/espanol_lengua_viva_2018.pdf`.

Kolhatkar, V. and M. Taboada. 2017a. Constructive language in news comments. In *Proceedings of the First Workshop on Abusive Language Online*, pages 11–17.

Kolhatkar, V. and M. Taboada. 2017b. Using new york times picks to identify constructive comments. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 100–105.

Kolhatkar, V., N. Thain, J. Sorensen, L. Dixon, and M. Taboada. 2020. Classifying constructive comments. *arXiv preprint arXiv:2004.05476*.

Kolhatkar, V., H. Wu, L. Cavasso, E. Francis, K. Shukla, and M. Taboada. 2019. The sfu opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, pages 1–36.

Lample, G. and A. Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

McHugh, M. L. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282.

Napoles, C., A. Pappu, and J. Tetreault. 2017. Automatically identifying good conversations online (yes, they do exist!). In *Eleventh International AAAI Conference on Web and Social Media*.

Napoles, C., J. Tetreault, A. Pappu, E. Rosato, and B. Provenzale. 2017. Finding good conversations online: The yahoo news annotated comments corpus. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 13–23.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Puri, S. and S. P. Singh. 2019. An efficient hindi text classification model using svm. In *Computing and Network Sustainability*. Springer, pages 227–237.

Swanson, R., B. Ecker, and M. Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue*, pages 217–226.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 53-64 recibido 25-11-2020 revisado 25-01-2021 aceptado 03-02-2021

Using Dependency-Based Contextualization for transferring Passive Constructions from English to Spanish

Contextualización basada en dependencias sintácticas para transferir construcciones pasivas de inglés a español

Pablo Gamallo,[1] **Gorka Labaka**[2]
[1]Centro de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
[2]IXA, Universidad del País Vasco (UPV/EHU)
pablo.gamallo@usc.gal; gorka.labaka@ehu.eus

Abstract: We hypothesize that parallel corpora as well as machine translation outputs contain many literal translations that are the result of transferring the constructions of the source language to the target language. When translating passive expressions from English to Spanish, there are several constructions available, however, both automatic and human (if of low quality) translations tend to select the periphrastic structure, which is the literal construction. The objective of this article is to make use of strategies trained on monolingual corpora to translate English passive expressions into Spanish so as to verify whether unsupervised translation with monolingual corpora benefits syntactic diversity. Special attention will be given to the monolingual-based strategy relying on dependency-based contextualization. The results of the experiments carried out show that the methods relying on monolingual corpora tend to offer more non-literal translations (middle-voice) than those trained on parallel corpora.
Keywords: Semantic Contextualization, Similarity, Unsupervised Machine Translation, Passive Voice.

Resumen: Formulamos la hipótesis de que los corpus paralelos así como los resultados de la traducción automática contienen muchas traducciones literales que son el resultado de la transferencia de las construcciones del idioma de origen al idioma de destino. Cuando se traducen expresiones pasivas del inglés al español, hay varias construcciones disponibles, sin embargo, tanto las traducciones automáticas como las humanas (si son de baja calidad) tienden a seleccionar la estructura perifásica, que es la construcción literal. El objetivo de este artículo es hacer uso de estrategias entrenadas a partir de corpus monolingües para traducir las expresiones pasivas del inglés al español, a fin de verificar si la traducción no supervisada con corpus monolingües beneficia la diversidad sintáctica. Se prestará especial atención al método semántico que se apoya en el proceso de contextualización en el marco de la sintaxes de dependencias. Los resultados obtenidos en los experimentos muestran que los métodos basados en corpus monolingües tienden a generar más traducciones no literales (voz media) que los entrenados con corpus paralelos.
Palabras clave: Contextualización semántica, Similaridad, Traducción automática no supervisada, Voz pasiva.

1 Introduction

Passive voice is a type of linguistic construction shared by most world's languages (Keenan, 1985). It is the result of a detransitivizing process that reduces the verbal valence promoting the object (or a similar function) to the subject position, which becomes the topic of an depersonalized sentence. Languages may differ with regard to formal codification of passive constructions, such as word order, case assignment, or verb morphology, but two elements are fundamental from a cross-linguistic point of view: the existence of an active counterpart and the prominence of a non-agent participant in the syntactic

ISSN 1989-7553. DOI 10.26342/2021-66-4

encoding of the passive clause (Siewierska, 1984).

There are remarkable differences in the passivization process between English and Spanish. Whereas in English passive constructions are mainly encoded by means of the periphrastic passive (*be+pp*), in Spanish there are several ways of encoding passive constructions, including the periphrastic strategy but also middle voice constructions containing the non-anaphoric pronoun *se*. Linguistic studies claim that periphrastic passive is less frequent in Spanish than in English as Spanish tends to use middle-voice constructions with *se* (Rodríguez-Vergara, 2017). In addition to that, it is important to take into account literal translation, which is a translation strategy that follows closely the surface form of the source language. By producing a structurally close translation of the source text, the translation process reaches the final result in a very efficient and fast way, and with the minimal processing effort. For this reason, literal translation is by far the most frequent method in all translation types (Scarpa, 2020). The well-known translation model defined by Vinay and Darbelnet (1995) considers that literal translation is the most suitable strategy if it does not modify the meaning of the source text or is not possible because it changes the structure of the source language. According to these ideas on literal translation, we assume that the periphrastic Spanish passive would be the most used translation of the corresponding English construction, leaving out the other non-literal alternative, i.e., middle-voice, even though this construction could be even more appropriate than the periphrastic one in many situations. Considering these facts, we propose the following hypothesis:

> **Periphrastic Passive Bias in English-Spanish Translation:** The Spanish periphrastic passive is over-represented in English-Spanish parallel corpora as it corresponds to the literal translation, which leads to the fact that English-Spanish automatic translations based on parallel corpora (supervised machine translation) tend to massively generate this structure in the Spanish texts; by contrast, the alternative passive constructions (middle-voice), which are very common structures in all Spanish domains and genres, are scarce in parallel corpora and then underused by supervised machine translation.

To check if the hypothesis is true, we propose using unsupervised methods trained on monolingual corpora (non-parallel texts) in both English and Spanish. As middle-voice constructions are more natural and frequent in untranslated texts, they are supposed to emerge more frequently with strategies relying on monolingual corpora than with supervised methods based on parallel corpora. So far, the main reason for abandoning dependence on parallel corpora is that they are scarce especially in the case of low-resource languages. However, we now propose a linguistic reason based on the above-mentioned linguistic bias. If it is verified, we can admit that there is a general bias towards translations that are too literal (especially if the translator is not professional, is poorly paid or has little time), and this has repercussions on the quality and structural diversity of the results of machine translation.

This hypothesis is in accordance with recent findings (Vanmassenhove, Shterionov, and Way, 2019; Toral, 2019), which conclude that machine translation gives rise to lower lexical diversity than human translations. In addition, machine and human translations have lower lexical diversity than monolingual texts that are naturally written in the target language.

In the present work, we will make use of strategies trained on monolingual corpora to translate English passive expressions into Spanish. In addition to unsupervised machine translation techniques (Artetxe et al., 2018; Artetxe, Labaka, and Agirre, 2019), special attention will be paid to a dependency-based approach to perform contextualized translations from monolingual corpora (Gamallo et al., 2019). We will configure a system that follows this approach in order to select the most appropriate passive construction in Spanish given an English periphrastic passive expression. The specific objective is to permit the dependency-based strategy to transfer passive constructions from one language to another.

The rest of the article is organized as fol-

lows. In the next section (Sec. 2), we deal with passive constructions in English and, above all, in Spanish from a linguistic point of view. Section 3 introduces two unsupervised approaches to translation from monolingual corpora, paying special attention to the strategy of syntactic contextualization (Subsection 3.2). Section 4 applies the latter strategy to the transfer of passive constructions from English to Spanish. Experiments are reported in Section 5 and final conclusions are addressed in Section 6.

2 Passivization in English and Spanish

Semantically, the passive construction implies, on the one hand, defocalization of the agent, which is encoded in oblique case or even suppressed from the sentence and, on the other hand, topicalization of other participant, which can be the patient, experiencer, theme, beneficiary, etc. So, passivization can be seen as the shift in focus from the agent (or similar role) to a non-agentive participant (usually the patient) in an event. Siewierska (1984) points out that the necessity for the use of the passive varies from language to language, and the differences are not only in the way the constructions are syntactically encoded, but also in how they express semantic-discursive functions related to topicalization, impersonalization and detransitivisation.

2.1 Types of Passive Constructions in English and Spanish

From the syntactic point of view, English and Spanish share the periphrastic passive which consist of transforming the active verb into the periphrasis *be/ser + past participle*. Semantically, the periphrastic passive construction in both languages allows the conceptualization of the agent, although it remains a marginal conceptualization as it can only be encoded with an optional oblique case (*by/por + agent*) (Fernández, 2007). Table 1 shows some examples of expressions encoded with the periphrastic passive in English and Spanish. All English periphrastic expressions can be translated into Spanish using the same construction. Even though the agent can be expressed in all of them, only the first example contains an explicit agent coded with the oblique case: *by the Jesuits/por los Jesuitas*. In the other three periphrastic expressions, it has not been expressed as it is optional.

As Table 1 also shows (see the two columns on the right), there are two other types of passive constructions in Spanish that do not exist in English: both reflexive and impersonal passives, which are usually called *middle-voice* in opposition to (periphrastic) passive voice. Syntactically, they are constructed with the active verb along with the insertion of the non-anaphoric pronoun *se*. They differ in the way of encoding the non-agentive topicalized participant. In the reflexive construction the topicalized participant is the subject of the clause (there is agreement with the verb), whereas in the impersonal one it is encoded as direct object preceded by the preposition *a*. The two middle-voice constructions are semantically very similar but they are in complementary distribution in some cases: when the topicalized participant is inanimate, the preferred encoding is the reflexive passive. By contrast, if the topicalized participant is animate or human (and then a potential agent), the preferred encoding is the impersonal construction because in those cases reflexive passives are formally very close to active reflexive/reciprocal clauses. For instance, the translation of *The workers were threatened* into a middle-voice construction must be the impersonal passive *Se amenazó a los trabajadores* since the reflexive passive (*Se amenazaron los trabajadores*) may be confused with its corresponding active clause with a reflexive/reciprocal meaning: *the workers threatened each other*. As the non-agentive participant, *the workers/los trabajadores*, are human beings, they can be interpreted as agents and patients at the same time giving rise to the active and reciprocal construction. Thus, to prevent the confusion with active reflexive/reciprocal clauses, the reflexive passive is not allowed with this type of agentive participants. In fact, in those cases it is very common in Spanish to use a hybrid (but ungrammatical) structure that mixes both reflexive and impersonal passives: (**Se amenazaron a los trabajadores*) (Sánchez-López, 2002).

Semantically, unlike the periphrastic passive, the two middle-voice constructions prevent the conceptualization of the agent of the active clause from which it derives. In the first example of Table 1, the agent *by the Jesuits/por los Jesuitas* cannot be syn-

	periphrastic passive	reflexive passive	impersonal passive
en	*The church was founded in 1850 by the Jesuits*	-	-
spa	*La iglesia fue fundada por los Jesuitas*	**La iglesia se fundó por los Jesuitas*	**Se fundó a la iglesia por los Jesuitas*
en	*The church was founded in 1850*	-	-
spa	*La iglesia fue fundada en 1850*	*La iglesia se fundó en 1850*	**Se fundó a la iglesia en 1850*
en	*The treaty was signed in Lisbon*	-	-
spa	*El tratado fue firmado en Lisboa*	*El tratado se firmó en Lisboa*	**Se firmó al tratado en Lisboa*
en	*The workers were threatened*	-	-
spa	*Los trabajadores fueron amenazados*	*?Se amenazaron los trabajadores*	*Se amenazó a los trabajadores*

Table 1: Passive English sentences and their Spanish translations with different passive constructions: periphrastic, reflexive and impersonal.

tactically encoded in the reflexive passive as it is not semantically conceptualized within the depersonalized scene designed by the verb (Fernández, 2007).

Periphrastic and middle-voice constructions in Spanish are not in complementary distribution. In many cases, both options are allowed to translate the same passive in English (if this one does not contain an explicit oblique agent). The second, third and fourth examples in Table 1 are encoded in Spanish in at least two constructions: periphrastic and either reflexive or impersonal passive. However, there are aspectual and lexical restrictions that tend to favor one construction or another. Periphrastic passive tends to be used with verbs expressing singular events with an external object and agent. By contrast, the use of verbs with a habitual, repetitive (iterative) or generic lexical aspect favors middle-voice passives (de Miguel, 1999). It was also found that the middle-voice is used more with material and relational events (77% of cases) than with mental, existential and behavioural processes (Lourdes Díaz Blanca, 2008).

There are serious problems to carry out quantitative studies to compare the use of periphrastic and middle-voice constructions by using automatic approaches. It is not possible to automatically identify passivizations with the non-anaphoric pronoun *se* as this pronoun also co-occurs with verbs in the active form to build many other syntactic constructions as it is reported in García-Miguel (1985). Alarcos (1978) counted up to 9 types of uses of *se*.

2.2 Frequency of Use of Different Passivization Types in Spanish and English

The experiments reported in Jisa et al. (2002) show that the periphrastic passive constructions are used significantly more in Dutch, English, and French than in Hebrew or Spanish. In fact, Spanish shows very little reliance on this type of construction across narrative and expository texts. However, this does not mean that passivization (either with peripheral or middle-voice constructions) is not as common in Spanish as in English.

A recent study analyzed and counted the number of different constructions found in a parallel English-Spanish text (Rodríguez-Vergara, 2017). The parallel text consists of an scientific article on the medical field written in English and its translation into Spanish. The authors found 52 periphrastic passive constructions in English and 48 passive translations in Spanish, 15 being periphrastic and 33 being reflexive/impersonal (i.e., middle-voice). Despite the small size of the corpus, this study shows two trends: (i) most of the English passive constructions are translated into passivized structures in Spanish (rather than in active constructions), and (ii) middle-voice constructions are more common in Spanish than periphrastic ones. This trend should be much more pronounced in other genres and domains: literary texts, informal language, etc.

To the best of our knowledge, there are no NLP-based studies on the use of passivization in English and Spanish.

3 Machine Translation from Monolingual Corpora

In the Introduction, we have claimed that there is a periphrastic passive bias in English-Spanish translation. This bias consists of the fact that Spanish periphrastic passive is over-represented in English-Spanish parallel corpora at the expense of the other passivizations because of the influence of the source language (English) in the translation process. This also leads to MT systems based on supervised training (i.e. parallel corpora) producing a bias in favour of the periphrastic structure in their results.

To check whether this claim tends to be true, we will carry out experiments with translation strategies trained on monolingual corpora, which are not biased in favor of one type of construction, but represent natural text without the influence of the source language.

We distinguish between two types of monolingual-based strategies: unsupervised machine translation and dependency-based contextualized translation. In this section, we will briefly explain these two strategies. In the next one, we will propose an improvement for the second that will allow us to apply it to transfer passive constructions from one language to another.

3.1 Unsupervised Machine Translation

Unsupervised MT was born to minimize the dependency on parallel data by training machine translation systems using only monolingual corpora. These strategies started with neural sequence-to-sequence systems which consist of training a dual model combining back-translation and denoising autoencoding (Artetxe et al., 2018; Lample et al., 2018a). The training process is initialized with cross-lingual embeddings, which can be also generated using an entirely unsupervised method by automatically learning the mapping between two vector spaces without the support of an external bilingual dictionary (Artetxe, Labaka, and Agirre, 2018a).

These neural-based systems were recently overtaken by a more traditional phrase-based statistical MT approach also provided with an unsupervised strategy (Lample et al., 2018b; Artetxe, Labaka, and Agirre, 2018b). The new approach leverages the modular architecture of statistical MT: a phrase table is induced through cross-lingual embedding mapped from monolingual corpora, this table is combined with a n-gram language model, and the system is improved through iterative back-translation.

More recently, Artetxe et al. (2019) described a hybrid approach with state-of-the-art results for unsupervised MT. It is based on the above-mentioned statistical MT approach reported in Artetxe (2018b), which is used to initialize an unsupervised neural system improved through on-the-fly back-translation.

3.2 Dependency-Based Contextualized Translation

In a recent work, Gamallo et al. (2019) describe a compositional distributional method to generate contextualized senses of words in a cross-lingual space that is aimed at selecting the most appropriate translations in the target language using monolingual corpora. It is a dependency-based strategy inspired on previous work reported in Erk and Padò (2008; 2010) and applied to a cross-lingual space. The dependency-based translation strategy consists of the following steps:

(1) Generation of candidates: The source expression is syntactically analyzed using syntactic dependencies and the resulting construction is associated with a set of candidate translations in the target language with an equivalent syntactic construction. An example will help us illustrate this. The English phrase *catch a ball* is syntactically analyzed as a direct object dependency: $< obj, catch, ball >$. Then, by making use of a bilingual dictionary, each English constituent term is expanded with their Spanish translations giving rise to a list of candidate expressions encoded with the same construction. For instance, if the Spanish translations of *catch* found in the dictionary are the verbs *coger* and *contraer*, and those of *ball* are the nouns *pelota* and *baile*, then, we can generate four possible Spanish combinations (see Table 2).

Even though four candidates were generated, only the first one is the correct Spanish translation of the source expression. The other cases are weird combinations produced by the polysemy of the words constituting the English expression. We must point out that this is a toy example as the number of equivalent translations per word has been reduced

$< obj, coger, pelota >$	$(catch\ a\ ball)$
$< obj, coger, baile >$	$(*catch\ a\ dance)$
$< obj, contraer, pelota >$	$(*contract\ a\ ball)$
$< obj, contraer, baile >$	$(*contract\ a\ dance)$

Table 2: Spanish translation candidates for the dependency $< obj, catch, ball >$.

to facilitate explanation. Using real bilingual dictionaries, *catch* may have up to ten different Spanish translations and *ball* five, giving rise to 50 (10x5) Spanish candidates. In a related paper by the same authors (Gamallo and Garcia, 2019), they used a totally unsupervised strategy. Instead of external bilingual dictionaries, they made use of unsupervised learned cross-lingual embeddings to generate the target language candidates. The next two steps are designed to automatically select the correct translation(s) from the generated candidates.

(2) Contextualized senses: Once the candidates in the target language have been generated, the next step is to build the distributional meaning representation of both the source expression and translation candidates. The distributional meaning of each expression stands for the contextualized senses of its constituent words (Gamallo, 2019). Let us continue with the previous example. The meaning of *catch a ball* consists of two contextualized senses: the sense of the verb *catch* given the noun *ball* in the direct object position, and the sense of *ball* as direct object of *catch*. Each contextualized sense is built in two sequential processes. In the case of *catch*, the first process is to build the selectional preferences of *ball*. Intuitively, they correspond to the most relevant verbs that can be combined with the noun *ball* in the direct object position. Formally, they are defined by the vector resulting of adding the vectors of those relevant verbs. The second process consists of combining the vector of *catch* with the resulting vector representing the selectional restrictions imposed by *ball*. This combination, implemented by means of vector multiplication, represents the contextualized sense of *catch*. Similar processes are carried out with the other word *ball*. The final meaning of the expression are thus two contextualized vectors, noted $\mathbf{catch}_{obj\uparrow}$ and $\mathbf{ball}_{obj\downarrow}$, where $_{obj\uparrow}$ and $_{obj\downarrow}$ represent the *head* and *dependent* roles of *catch* and *ball*, respectively, in the direct object relation. Sense elaboration is thus the result of bi-directional operations: the head word restricts the sense of the dependent one in the same way as the latter restricts the head. The same contextualization process is applied to all the Spanish candidates so as to create their corresponding contextualized vectors. We should note here that there is a great conceptual parallelism between this contextual strategy and the recent Transformers models based on bi-directional contextualized word embeddings (Devlin et al., 2019). However, while the latter are mainly based on syntagmatic relationships (word co-occurrences in context), the former mainly relies on paradigmatic relationships established between optional words potentially occurring in the same syntactic functions.

(3) Selection by Similarity: Finally, the distributional meanings (defined as contextualized senses) of the generated candidates are compared pairwise by means of cosine similarity with the English sentence. The generated Spanish sentence associated with the most similar meaning is selected as the best Spanish translation of the English sentence. More precisely, given a specific dependency $< obj, catch, ball >$ in the source language, its contextualized translation, CT, in the target language is computed as follows:

$$CT(< obj, catch, ball >) = \quad (1)$$
$$\underset{<obj, w_1, w_2> \in \phi}{\arg\max} \frac{1}{2} S(\mathbf{catch}_{obj\uparrow}, \mathbf{w_1}_{obj\uparrow}) + S(\mathbf{ball}_{obj\downarrow}, \mathbf{w_2}_{obj\downarrow})$$

where (obj, w_1, w_2) is any target dependency belonging to the set of translation candidates, ϕ (see an example of this set in Table 2). The first S computes the similarity between the two contextualized vectors associated to the head words in the source and target languages. The second S computes the similarity between the vectors derived from the dependent words. So, the overall similarity between two composite expressions is the mean of the similarity scores obtained by comparing their head-based and dependent-based contextualized vectors. The resulting translation is, thus, the expression belonging to ϕ with the highest CT score.

4 *Applying Dependency-Based Contextualized Translation to Passivization*

Given the syntactic nature of the phenomenon of passivization, we think that the dependency-based strategy defined in Subsection 3.2 is perfectly suited to tackle the complexity of the phenomenon. For this purpose, two new requirements are needed: to define specific dependencies for the different passive constructions in English and Spanish, and to expand the set of Spanish candidates with those syntactic constructions by making use of syntactic translation equivalents.

4.1 Passive Dependencies

In order to build the contextualized senses of passive constructions, it is necessary to identify them with the appropriate syntactic analysis. Although periphrastic passives are relatively easy to identify in both English and Spanish, to the best of our knowledge, there is no syntactic parser capable of analyzing the middle-voice constructions in Spanish. In Table 3, we show the passive expression in English and Spanish (first column) along with the syntactic dependency we are looking for (second column).

Notice that we need specific dependencies that are not even defined in the Universal Dependencies (UD) project (Nivre and others, 2017). Neither $nsubj\_PP$ (nominal subject of a periphrastic passive), $nsubj\_RP$ (nominal subject of a reflexive passive) nor obj_IP (direct object of an impersonal passive) are functions defined in UD. In the case of $nsubj\_PP$, it is relatively simple to derive this dependency from the analysis: given the verb to *be* followed by a verb in past participle, its $nsubj$ must be of type PP. However, the identification of $nsubj\_RP$ and $nsubj\_IP$ is much harder. Expressions with similar surface form (*se+verb+np* and *se+verb+pp/a*) represent very different constructions. Table 4 shows Spanish expressions with similar surface form (first column), their functional analysis (second column), and the type of construction (third column). Only two of the six expressions are passive constructions: the first one (RP) and the fourth (IP).

In order to identify passive dependencies, we defined a set of syntactic rules provided with lexical restrictions on verbs that were implemented with DepPattern formalism.[1] Restrictions on verbs were learned by taking into account the syntactic-semantic classes compiled in the ADESSE database (García-Miguel, Vaamonde, and Domínguez, 2010). The resulting grammars are the basis for rule-based parsing dependencies adapted to the analysis of passive constructions both in English and, especially, in Spanish.

4.2 Syntactic Translation Equivalents between Languages

The dependency-based approach described in Subsection 3.2 is based on the generation of candidates by making use of bilingual lexical information provided by an external dictionary or a cross-lingual lexical model. In the example reported above (*catch the ball*), only one type of construction (direct object) have been used in both English and Spanish. The Spanish candidates have been generated using the same construction as in English, by combining the lexical translation equivalents of the constituent words (head and dependent) of the source expression. However, in addition to the lexical translation equivalents, we also need to consider syntactic translation equivalents.

In order to generate the set of candidates in ϕ, we combine both the set of lexical translation equivalents of the two source words with the set of syntactic translation equivalents of the source dependency by means of the Cartesian product of three sets as follows:

$$\begin{aligned}&\phi = ST(r_s) \times LT(w_{s1}) \times LT(w_{s2}) = \\ &\{< r_t, w_{t1}, w_{t2} >: r_t \in ST(r_s), \\ &w_{t1} \in LT(w_{s1}), w_{t2} \in LT(w_{s2})\}\end{aligned} \quad (2)$$

where $ST(r_s)$ is the set of syntactic translation equivalents of the source dependency r_s; $LT(w_{s1})$ is the set of lexical translation equivalents of the source head word w_{s1}; and $LT(w_{s2})$ is the set of lexical translation equivalents of the source dependent word w_{s2}. So, each $< r_t, w_{t1}, w_{t2} >$ is an ordered triple belonging to ϕ. Notice that Equation 1 defining CT above must be generalized by considering the more generic set of candidates ϕ defined in Equation 2, which includes now syntactic translation equivalents.

To deal with passivization in English-Spanish translation, we propose the set

[1] https://github.com/citiususc/DepPattern

passive expressions	dependencies
The house was sold	$< nsubj\_PP, sell, house >$
La casa fue vendida (*The house was sold*)	$< nsubj\_PP, vender, casa >$
La casa se vendió (*The house was sold*)	$< nsubj\_RP, vender, casa >$
Se despidió a los trabajadores (*The workers were fired*)	$< obj\_IP, despedir, trabajador >$

Table 3: Passive expressions and their dependency-based analysis. PP means periphrastic passive, RP reflexive passive, and IP impersonal passive.

Spanish expressions	functions	constructions
Se vendió la casa (*The house was sold*)	PRED-NSUBJ	reflexive passive (RP)
Se comió la manzana (*She/he ate the apple*)	PRED-OBJ	transitive active
Se cayó el lápiz (*The pencil fell down*)	PRED-NSUBJ	intransitive active
Se despidió a los trabajadores (*The workers were fires*)	PRED-OBJ	impersonal passive (IP)
Se comió a los niños (*(The monster) ate the children*)	PRED-OBJ	transitive active
Se arrodilló a tu lado (*She/he knelt beside you*)	PRED-OBL	intransitive active

Table 4: Spanish expressions with similar surface form to RP (*se+verb+np*) and IP (*se+verb+pp/a*) constructions.

$ST(nsubj\_PP)$ to be defined by the following elements: $\{nsub\_PP, nsub\_RP, obj\_IP\}$.

5 Experiments

5.1 Test Dataset and Evaluated Systems

In order to check to what extent the hypothesis set out in the *Periphrastic Passive Bias in English-Spanish Translation* stated in the Introduction is correct or not, we created a test dataset with 240 English passives (PP constructions). With different degrees of exigency, all the expressions of the dataset could be transferred to middle-voice constructions in Spanish (RP or IP constructions), even though most of them can also be transferred to periphrastic passives (PP), keeping so the same construction of the source language.

Table 5 quantifies the distribution of the types of constructions used by different systems for translating into Spanish the 240 English passive expressions. In this evaluation, we do not focus on the quality of the translation concerning the lexical choices, but just on the ability of the system to diversify the transfer of different passive constructions into Spanish. On the top of the table, we show four state-of-the-art commercial machine translators (supervised strategies), namely Bing,[2] DeepL,[3] Google Translator,[4] and Yandex,[5] which mainly use parallel corpora for training (all consulted in January 2020).

The test dataset was also processed with three unsupervised systems: A phrase-based SMT system (Artetxe, Labaka, and Agirre, 2018b), consisting of a log-linear combination of several statistical models learned from monolingual corpora; a hybrid SMT+NMT system (Artetxe, Labaka, and Agirre, 2019), consisting of the improved SMT system that initializes an unsupervised NMT model, which is further fine-tuned on the basis of on-the-fly back-translation; and ContextTrans, which is the enhanced version of the system based on the CT measure described above in Section 4. These three systems were trained using the same monolingual corpora, namely English and Spanish wikipedias (dump files of December 2018), with 21 and 5 billion words, respectively.

For ContextTrans, all texts were syntactically analyzed with LinguaKit (Gamallo et al., 2018), a multilingual suite which also in-

[2] https://www.bing.com/translator
[3] https://www.deepl.com/translator
[4] https://translate.google.com/
[5] https://translate.yandex.com/

supervised systems	***PP***	***RP***	***IP***	**% middle**
Yandex	219	21	-	8.75
GoogleTrans	218	22	-	9.16
DeepL	186	52	2	22.25
Bing	180	60	-	25.00
unsupervised systems	***PP***	***RP***	***IP***	**% middle**
Phrase-based_SMT	209	31	-	12.91
Hybrid_SMT+NMT	184	56	-	23.33
ContextTrans	132	94	14	**45.00**

Table 5: Distribution of the three types of passive constructions (PP, RP and IP) across the output returned by both supervised and unsupervised systems.

cludes the dependency-based parser, DepPattern (Gamallo and Garcia, 2018). The syntactically analyzed corpus was the basis for the elaboration of the salient lexico-syntactic contexts with which we constructed selectional preferences and contextualized vectors. Only lexical units occurring more than 100 times in each monolingual corpus were considered. As the lexical translation equivalents are not in the focus of the evaluation, we created a new input file for CT derived from the original test dataset. In this file, the English expressions were lemmatized and each constituent lemma was translated manually into Spanish. For instance, from the English PP expression "the aspirant was defeated", we just kept the pair of lemmas "aspirant" and "defeat", which were associated with their corresponding Spanish translations: "aspirante" and "derrotar". Each Spanish pair of lemmas represents the Cartesian product of $LT(w_{s1}) \times LT(w_{s2})$, which was combined with $ST(w_s)$ to generate all the translation candidates of each English passive expression. Notice that $LT(w_{s1})$ and $LT(w_{s2})$ are sets with cardinality 1 since we only consider variation across the set of syntactic translation candidates: $ST(w_s)$. So, after combining $ST(w_s)$, $LT(w_{s1})$ and $LT(w_{s2})$, each set of candidates, ϕ, is constituted by three dependencies, one per passive construction. Notice that as lexical units were lemmatized, ContextTrans do not consider information on aspect and tense of verbs or noun number.

5.2 Analysis of the Results

As has been said before, Table 5 shows how many times the passive constructions were used by supervised and unsupervised systems. In the case of supervised translators, the diversity in the use of different constructions is poor, as their translations are mostly done with PP construction. Besides, the use of IP is practically non-existent. Only DeepL uses it twice. The rest of the translators never use it, even though there are at least 60 expressions that could be translated that way in our dataset. Among the supervised systems, the two that return more syntactic diversity are DeepL (186 PP, 52 RP and 2 IP) and Bing (180 PP and 60 RP). Google Translator and Yandex behave very similarly with very poor diversity.

Concerning the unsupervised systems, Phrase-based_SMT is the least diverse, but it stands above the two least diverse supervised systems. Hybrid_SMT+NMT is between the two most diverse supervised systems, while ContextTrans is the system with the greatest diversity by large: 132 PP, 94 RP, 14 IP. The last column of Table 5 shows the percentage of middle-voice constructions with regard to the total number of examples in the dataset. The higher the percentage value, the less literal the translation is with respect to syntactic constructions. In the case of ContextTrans, 45% of constructions are not literal (middle-voice), which is almost twice as many cases as the most diverse supervised system. The few number of IP constructions returned by unsupervised approaches suggest that naturally written texts in Spanish contain fewer expressions with this type of construction than with RP.

The results seems to confirm the *Periphrastic Passive Bias* hypothesis, as the methods relying on monolingual corpora tend to offer more non-literal translations (middle-voice) than those trained on parallel corpora. The average of middle-voice translations with the four supervised systems is 16,29%, whereas the average for the three unsupervised systems reaches 27,08.

6 Conclusions

We have carried out an experiment to compare the syntactic diversity between supervised and unsupervised approaches to translation on one dataset consisting of English passive expressions. We have tuned a dependency-based translation strategy trained on monolingual corpora and verified, on the basis of its application to the dataset, that its syntactic diversity is greater than that of commercial translators relying on supervised techniques. These results confirm the hypothesis made at the beginning of our work in which we stated that both supervised systems have a bias towards more literal translation (PP constructions in English are translated by PP constructions in Spanish), and monolingual corpora allow learning a greater diversification of passive structures. So, natural text seems to be more syntactically diverse than MT output and parallel corpus. It should be noted that the syntax-based strategy, ContextTrans, can be considered a hybrid approach that integrates symbolic-syntactic knowledge in statistical-neural learning systems.

As the methodology can be applied to other linguistic phenomena and transferred to different language pairs, in future work we will seek to extend the experimentation towards other types of syntactic constructions taking into account the linguistic studies of Construction Grammar and its application to cross-lingual construction transfer (Boas, 2010). Experiments will also be conducted with a wider variety of linguistic genres, from literary to spoken corpora.

The code for the generic version of ContextTrans (without passive tuning), called compMT, is available at GitHub (`https://github.com/gamallo/compMT`). The dataset with the English passive expressions are availabe at `https://gramatica.usc.es/pln/resources/en_sentences240.txt.zip`.

Acknowledgments

This work has received financial support from DOMINO (PGC2018-102041-B-I00, MCIU/AEI/FEDER, UE), eRisk (RTI2018-093336-B-C21), the Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 2016-2019, ED431G/08, Groups of Reference: ED431C 2020/21) and the European Regional Development Fund.

References

Alarcos Llorach, E. 1978. Valores de 'se'. In *Estudios de gramática funcional del español*. Madrid, Gredos, pages 156–165.

Artetxe, M., G. Labaka, and E. Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia, July. Association for Computational Linguistics.

Artetxe, M., G. Labaka, and E. Agirre. 2018b. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium, October-November. Association for Computational Linguistics.

Artetxe, M., G. Labaka, and E. Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy, July. Association for Computational Linguistics.

Artetxe, M., G. Labaka, E. Agirre, and K. Cho. 2018. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations (ICLR-2018)*, April.

Boas, H. 2010. *Contrastive Studies in Construction Grammar*. John Benjamins Publishing Company.

de Miguel, E. 1999. El aspecto léxico. In I. Bosque and V. Demonte, editors, *Gramática descriptiva de la lengua española, vol. 2*. Madrid: Real Academia Española; Espasa Calpe.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Erk, K. and S. Padó. 2008. A structured vector space model for word meaning in context. In *2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-2008*, pages 897–906, Honolulu, HI.

Erk, K., Sebastian, Padó, and U. Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763.

Fernández, S. S. 2007. *La voz pasiva en español: un análisis discursivo*. Frankfurt am Main: Peter Lang.

Gamallo, P., M. Garcia, C. Piñeiro, R. Martinez-Castaño, and J. C. Pichel. 2018. LinguaKit: A Big Data-Based Multilingual Tool for Linguistic Analysis and Information Extraction. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 239–244.

Gamallo, P. 2019. A dependency-based approach to word contextualization using compositional distributional semantics. *Language Modelling*, 7(1):53–92.

Gamallo, P. and M. Garcia. 2018. Dependency parsing with finite state transducers and compression rules. *Information Processing & Management*, 54(6):1244–1261.

Gamallo, P. and M. Garcia. 2019. Unsupervised compositional translation of multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 40–48, Florence, Italy, August. Association for Computational Linguistics.

Gamallo, P., S. Sotelo, J. R. Pichel, and M. Artetxe. 2019. Contextualized translations of phrasal verbs with distributional compositional semantics and monolingual corpora. *Computational Linguistics*, 45(3):395–421.

García-Miguel, J. M., G. Vaamonde, and F. G. Domínguez. 2010. ADESSE, a database with syntactic and semantic annotation of a corpus of Spanish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Garcia-Miguel, J. M. 1985. La voz media en español: Las construcciones pronominales con verbos transitivos. *Verba: Anuario galego de filoloxia*, 12:307–343.

Jisa, H., E. Baruch, J. Reilly, E. Rosado, L. Tolchinsky, L. Verhoeven, and A. Zamora. 2002. Passive voice constructions in written texts: A cross-linguistic developmental study. *Written Language and Literacy*, 5(2):163–182.

Keenan, E. L. 1985. Passive in the world's languages. In T. Shopen, editor, *Language Typology and Syntactic Description. Vol. I.* Cambridge: Cambridge University Press.

Lample, G., A. Conneau, L. Denoyer, and M. A. Renzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the Sixth International Conference on Learning Representations (ICLR-2018)*, April.

Lample, G., M. Ott, A. Conneau, L. Denoyer, and M. A. Ranzato. 2018b. Phrase-Based & neural unsupervised machine translation, April.

Lourdes Díaz Blanca, C. L. D. 2008. Los verbos en las pasivas con se: un intento de clasificación. *Letras [online]*, 50(76).

Nivre, J. et al. 2017. Universal Dependencies 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, Prague, `http://hdl.handle.net/11234/1-1983`.

Rodríguez-Vergara, D. 2017. A systemic functional approach to the passive voice in english into spanish translation: Thematic development in a medical research article. *Open Linguistics*, 3(1).

Scarpa, F., 2020. *Translating Specialised Texts*, pages 187–290. Palgrave Macmillan UK, London.

Siewierska, A. 1984. *The Passive: Comparative Linguistic Analysis*. Routledge, Croom Helm Linguistics Series, London.

Sánchez-López, C. 2002. Las construcciones con se. estado de la cuestión. In C. Sánchez López, editor, *Las construcciones con se*. Madrid: Visor, pages 18–163.

Toral, A. 2019. Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 273–281, Dublin, Ireland, August. European Association for Machine Translation.

Vanmassenhove, E., D. Shterionov, and A. Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 222–232, Dublin, Ireland, August. European Association for Machine Translation.

Vinay, J. and J. Darbelnet. 1995. *Comparative stylistics of French and English* Benjamins, Amsterdam.

Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 65-76 recibido 19-11-2020 revisado 22-01-2021 aceptado 01-02-2021

Categorizing Misogynistic Behaviours in Italian, English and Spanish Tweets

Categorización de comportamientos misóginos en tweets en italiano, inglés y español

Silvia Lazzardi,[1] **Viviana Patti,**[2] **Paolo Rosso**[3]

[1]Dipartimento di Fisica, University of Turin, Italy
[2]Dipartimento di Informatica, University of Turin, Italy
[3]PRHLT research center, Universitat Politècnica de València, Spain
silvia.lazzardi@unito.it, patti@di.unito.it, prosso@dsic.upv.es

Abstract: Misogyny is a multifaceted phenomenon and can be linguistically manifested in numerous ways. The evaluation campaigns of EVALITA and IberEval in 2018 proposed a shared task of Automatic Misogyny Identification (AMI) based on Italian, English and Spanish tweets. Since the participating teams' results were pretty low in the misogynistic behaviour categorization, the aim of this study is to investigate the possible causes. We measured the overlap and the homogeneity of the clusters by varying the number of categories. This experiment showed that the clusters overlap. Finally, we tested several machine learning models both using the original data sets and merging together some categories according to their overlap, obtaining an increase in terms of macro F1.
Keywords: automatic misogyny identification, hate speech online.

Resumen: La misoginia es un fenómeno con múltiples facetas y puede manifestarse lingüísticamente de muchas formas. Las campañas de evaluación de EVALITA e IberEval en 2018 propusieron una tarea compartida de Identificación Automática de Misoginia (AMI) basada en tweets en italiano, inglés y español. Dado que los resultados de los equipos participantes fueron bastante bajos en la categorización del comportamiento misógino, el objetivo de este estudio es investigar las posibles causas. Medimos el solape y la homogeneidad de los clústeres variando el número de categorías. Este experimento mostró que los grupos se solapan. Finalmente probamos varios modelos de aprendizaje automático utilizando los conjuntos de datos originales y fusionando algunas categorías de acuerdo con consideraciones basadas en medidas de similitud y las matrices de confusión de los modelos, obteniendo un aumento de la F1 macro.
Palabras clave: identificación automática de misoginia, mensajes de odio online.

1 Introduction

During the last years, hateful language and in particular the phenomenon of hate speech against women, are exponentially increasing in social media platforms such as Twitter and Facebook (Poland, 2016), spreading across languages and countries. It is becoming a relevant social problem to be monitored, especially considering the results of studies on the usefulness of monitoring contents published in social media in foreseeing sexual crimes, such as the one in (Fulper et al., 2014), which confirms a correlation between the yearly per capita rate of rape and the misogynistic language used in Twitter. As women have often been targets of abusive language and hate speech, they also started to react on social media to both off and online abusive behaviours, for example through the viral use of the #mencallmethings hashtag to share experiences of sexual harassment in online environments and reflections on the perception of women on their freedom of expression (Megarry, 2014). Misogyny, defined as the hate or prejudice against women, can be linguistically manifested in different and various ways, including social exclusion, discrimination, hostility, threats of violence and

ISSN 1989-7553. DOI 10.26342/2021-66-5

sexual objectification (Anzovino, Fersini, and Rosso, 2018).

Given the huge amount of social media data in many languages, the urgency of monitoring misogynistic behaviours and contents online calls the computational linguistics community for a deeper effort on developing tools to automatically identify and categorize hateful content online against women, possibly bringing a multilingual perspective to highlight different viewpoints on how misogyny is not only perceived, but also expressed in different languages and cultures. The recent proposals of the AMI and HatEval shared tasks focusing on the detection of misogyny and hate speech at EVALITA (Fersini, Nozza, and Rosso, 2018) and IberEval (Fersini, Rosso, and Anzovino, 2018), and SemEval (Basile et al., 2019) respectively, can be read in the light of this urgency. Our starting point in this work is the Automatic Misogyny Identification (AMI) shared task proposed in 2018 first at IberEval for Spanish and English (Fersini, Rosso, and Anzovino, 2018), and then at EVALITA for Italian and English (Fersini, Nozza, and Rosso, 2018), to identify misogyny in Twitter texts at different levels of granularity. The task has been articulated in two sub-tasks: a first one devoted to distinguish between messages with misogynous content from not misogynous ones, and a second one with the goal to categorize the misogynous content at a finer grained level. The teams who participated in this shared task, despite having obtained excellent results in terms of the coarse grain binary classification of messages, didn't achieve a good performance in the finer grain sub-task on categorizing misogynistic behaviour, which is important to get a deeper understanding of the multi-faceted misogynistic attitudes. The aim of this study is to shed some light on the reasons behind this.

We focus on investigating how the misogyny categories distribution in the available data sets influences models performances and the relationships among the categories themselves. In particular, we used both unsupervised and supervised machine learning algorithms to answer to the following research questions:

RQ1 Is an unsupervised clustering algorithm able to identify specific patterns and separate data into homogeneous clusters according to the labeled categories of misogynistic behaviour? How do homogeneity and distance change as the number of clusters change?

RQ2 Is it possible to extract empirical information on the similarity relationship among the categories of misogynistic behaviour, by studying a metric among such categories?

RQ3 Using the information obtained from these studies, is it possible to improve the performance of the models on the finer grain misogynistic behaviour classification task?

The paper is organized as follows. Section 2 reports on related literature. Section 3 describes the AMI shared task. Sections 4 and 5 present methods, experiments and results obtained respectively by using unsupervised clustering techniques and supervised machine learning algorithms. Section 6 provides a discussion of the results. Section **??** draws some conclusions.

2 Related Work

Misogynistic language expressed in social media is a multifaceted phenomenon with its own specificity. It has been often treated as an expression of sexist offenses. One of the first works in this area proposed a mixed data set of racism and sexism (Waseem and Hovy, 2016).

A philosophical account of misogyny and sexism has been provided by (Manne, 2017), which arguments that they are distinct. On this line, (Frenda et al., 2019) presented an approach to detect separately both misogyny and sexism analyzing collections of English tweets.

Another important contribution is due to (Farrell et al., 2019), that investigated how misogynistic ideas spread within and across Reddit communities by building lexicons of hate which describe such misogynistic behaviours. Two very recent works addressed the problem of misogyny identification in Spanish tweets comparing the performance of several models on the HatEval data set (Plaza-Del-Arco et al., 2020), and applying sentiment analysis and social computing technologies (García-Díaz et al., 2021). In the latter work, the authors have

compiled the balanced MisoCorpus-2020 corpus. Misogyny identification has been also addressed from a multimodal perspective in (Gasparini et al., 2018), where both visual and texture features are analyzed to classify sexist contents, and in a multilingual perspective in (Pamungkas, Basile, and Patti, 2020).

3 Task and Data Sets Description

The Automatic Misogyny Identification (AMI) shared task, presented for the first time at the IberEval 2018 and then at EVALITA 2018, proposes misogyny identification, misogynistic behaviour categorization, and target classification both from Spanish, English and Italian tweets.

3.1 AMI Task

The AMI shared task of EVALITA and IberEval is subdivided into two sub-tasks:

- **Sub-task A - Misogyny identification**: binary classification of misogynous contents from the non misogynistic ones.
- **Sub-task B - Misogynistic behaviour and target classification**: identification of the particular type of misogynous behaviour against women and classification of the target of the insult, that can be *active* or *passive* respectively whether it is addressed to a specific user or to a group of women.
  The different misogynous behaviours are classified into the following categories:
  - **Derailing**: to justify male abuse of women.
  - **Discredit**: slurring without any other intentions.
  - **Dominance**: to emphasize the gender superiority of men over women.
  - **Sexual harassment**: to describe actions as sexual advances, requests for sexual services.
  - **Stereotype**: to describe women using oversimplified ideas as the physical appeal and/or comparison with narrow standards.

In Table 1 are reported the different categories of misogynistic behaviour, together with an example for each of them.

Category	Example
Derailing	@yesallwomen wearing a tiny skirt is asking for it. Your teasing a (hard working, taxes paying) dog with a bone. That's cruel. #YesAllMen
Discredit	@melaniatrump stupid fuc**ing bi**ch
Dominance	Women are inferior to men..so shut up please
Sexual harassment	I wanna touch your te**ts..you can't imagine what I can do on your body.
Stereotype	Women play your role.. #gobacktothekitchen Big te**ts and no brain..this is what I want as wife!

Table 1: Misogynous categories and examples of tweets from the AMI dataset.

In this study we deal with the part of Sub-task B devoted to the misogynistic behaviour classification. Regarding the evaluation measure used to rank the submissions for this sub-task, the Macro F_1-score was used. The misogynistic behaviour categorization is obtained as the average of F_1-score of each misogynous category ($F_1\_misogyny$).

3.2 Training and Test Data Sets

The data sets used in this study are collections of tweets. We used the Italian and the English data set from EVALITA and the Spanish one from IberEval. Each data set is distinguished in a training and a test set. The EVALITA data set contains respectively 4000 and 1000 tweets for the train and the test sets, while the IberEval one includes 3302 tweets in the training set and 831 in the test set (Fersini, Nozza, and Rosso, 2018) (Fersini, Rosso, and Anzovino, 2018). Each

tweet has been provided with several fields: the *id*; the Twitter *text*; *misogynous* which defines whether the tweet is misogynous or not (1 if the tweet is misogynous, 0 otherwise); *misogyny_category*, a label to denote the misogynous behaviour (0 if it's not misogynous) and finally the target (*active*, *passive* or '0' if the tweet is not misogynous).

3.2.1 Categories Distribution

Table 2 reports the number of samples for each misogynous category while Figure 1 shows the categories distribution across the corpus. We can observe that data present many inhomogeneities both in the training and in the test sets.

In all the data sets, about half of the tweets are not misogynous (with percentages ranging from 49.1% to 55.4%). The other half contains tweets belonging to the five labeled misogyny categories according to different percentages. The less represented category is *derailing*, with percentages ranging from just 0.2% for the Italian test set to 2.3% for the English training set. It follows *dominance* which represents just the 1.8% of the Italian training set. The English and the Spanish data set contain more *dominance* tweets, with a maximum of 12.4% for the Spanish test set. About the other three categories (*discredit*, *sexual harassment* and *stereotype*), we can note that in the Italian data set tweets with *stereotype* are present in an higher percentage respect to the English and Spanish data sets (being 16.7% in the Italian train set versus 4.5% and 4.6% in the English and Spanish ones). Otherwise, in the English and Spanish data sets *discredit* is significantly more represented.

4 Clustering

In this section we describe the clustering methods used to evaluate the data sets from an unsupervised perspective and the obtained results. We extracted features from the data sets using two methods: Bag of Words (BoW) and tf-idf. Then we applied three different clustering algorithms: K-means, Spectral Clustering and Agglomerative Clustering. As a second experiment, we computed the inter-category distance matrices using the cosine similarity.

4.1 Experimental Setting

In order to answer RQ1, we carried out a a first experiment where we clustered the tweets of the data sets without considering the category label.

Our aim was to investigate: (i) how the distance among the clusters changes by varying their number; and (ii) how the misogynistic behaviour categories are distributed in the clusters. In order to address the above points, we computed two scores:

- *Silhouette score*: a measure of the distance among clusters that is computed using the mean inter-cluster distance (a) and the mean nearest-cluster distance (b) for each tweet. Given the data matrix $\mathbf{X}= \{x_1, \ldots, x_n\}$ and the label vector $\mathbf{y}$, we define it as:

$$Sil(\mathbf{X}, \mathbf{y}) = \sum_{i=1}^{n} \frac{b_i - a_i}{max(a_i, b_i)} \quad (1)$$

  The Silhouette score takes values $\in \{-1, 1\}$. Values near 0 indicate that the clusters overlap each other.

- *Homogeneity score*, a measure of how many clusters contain only tweets that belong to a misogynistic behaviour category. Its value ranges from 0 to 1, where 1 corresponds to a perfectly homogeneous labeling.

Since there are mainly two categories underrepresented in the data sets (*derailing* and *dominance*), we run each clustering algorithm by varying the number of clusters from 6 (not misogynistic plus the 5 misogynistic behaviour categories) down to 4 (not misogynistic plus 3 misogynistic behaviour categories), in order to see if there is an improvement in the scores, ignoring the two least represented categories. Given a fixed number of clusters, we computed the Silhouette and the Homogeneity scores. Moreover, we represent the distribution of the misogynistic behaviour categories within each cluster.

In order to answer RQ2, we carried out the experiment from the opposite perspective. We considered perfectly homogeneous clusters and computed distances among the misogynistic behaviour categories using the cosine similarity. We didn't use the Euclidean distance due to the high dimensionality of the feature vectors.

4.2 Results

First we calculated the Silhoutte and Homogeneity scores resulting from the analy-

	Not mis.	Derailing	Discredit	Dominance	Sexual harass.	Stereotype	**Tot**
Italian							
Train	2172	24	634	71	431	668	**4000**
Test	491	2	104	61	167	175	**1000**
English							
Train	2214	92	1014	148	351	179	**4000**
Test	540	10	141	123	43	140	**1000**
Spanish							
Train	1658	19	977	301	197	150	**3302**
Test	416	5	286	54	50	17	**831**

Table 2: Number of samples for each misogynous category for Italian, English and Spanish both for training and test data sets.

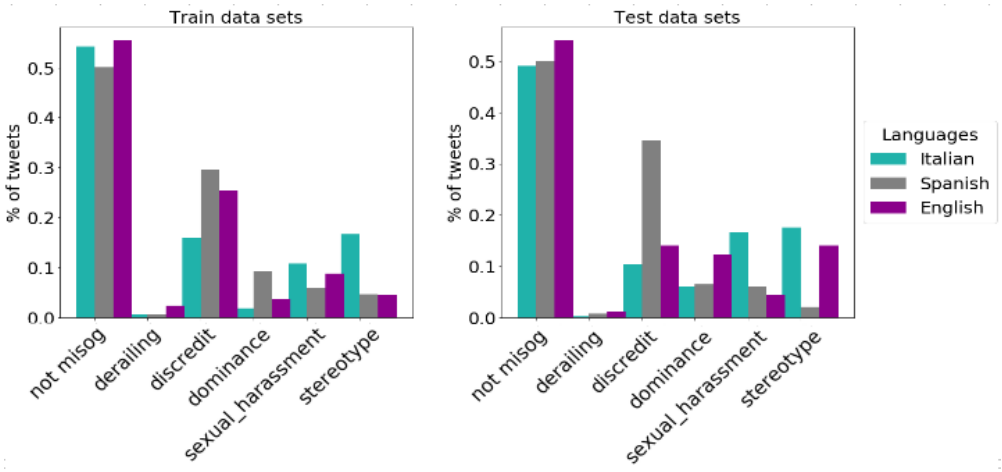

Figure 1: Categories distribution of the data sets: training (left) and test (right).

sis of the tf-idf feature vectors by decreasing the number of clusters from 6 down to 4. The Homogeneity score was always very low, which indicates that the clustering algorithms couldn't recognize a significant difference between the different categories. Then we did the same using the BoW feature vectors. We used the statistical test of Wilcoxon, to see if the difference using tf-idf and BoW was statistically significant. With a p-value of 0.66 we can say it wasn't. In Figure 2 we show the results obtained with the BoW feature vectors (similar results were obtained with the tf-idf feature vectors).

Figure 3 shows how the tweets belonging to different misogynistic behaviour categories are distributed among the clusters. Since all clusters have at least a sample for each category, it is clear why the homogeneity score was very low.

Finally, we computed the cosine similarity among the misogynistic behaviour categories for the three data sets. The results are illustrated in Figure 4. In agreement with the clustering results, we can appreciate how difficult is to separate the misogynistic behaviour categories. Interestingly *derailing* seems to be the most well defined category, followed by *dominance*. This could be due to their smaller number of tweets (i.e., less variability in their feature vectors). An interesting insight is the overlapping between *discredit* and *dominance* and between *derailing* and *sexual harassment*.

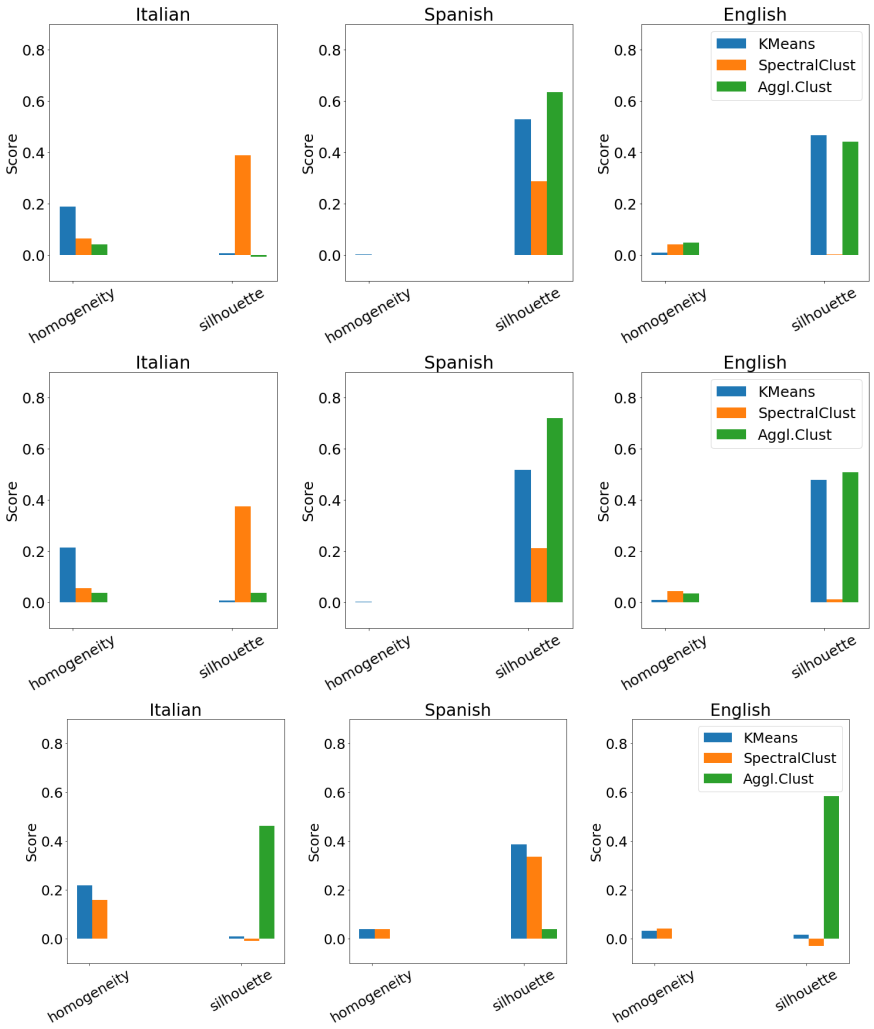

Figure 2: Silhouette and Homogeneity scores for Italian, Spanish, and English using BoW feature vectors by varying the number of clusters. From left to right: 6, 5 and 4 clusters. Train and test data sets were merged together.

5 Behaviour Classification

In this section we comment on the machine learning algorithms we used on the EVALITA and IberEval data sets, and the results we obtained.

5.1 Experimental Setting

For this experiment we tried two approaches to built feature vectors, using different kinds of features. In the first case, we extracted a list of words processing each tweet of the training data set. We did not consider those words with a frequency smaller than 0.001. In the second case, we used as list of features those taken from *Hurtlex*, a multilingual lexicon of offensive, aggressive, and hateful words (Bassignana, Basile, and Patti, 2018).

Hurtlex categorizes words according to 17 categories, from which we selected the most relevant for our purpose, as reported in Table 3.

Label	Description
PR	words related to prostitution
ASM	male genitalia
ASF	female genitalia
DDP	cognitive disabilities and diversity
DDF	physical disabilities and diversity

Table 3: S *Hurtlex*: selected categories (labels and descriptions) to build the vocabulary of features.

To convert the tweets to feature vectors we used tf-idf and BoW. To classify the tweets we trained different We run these models and compared their performance, using tf-idf and BoW, as well as the features extracted from

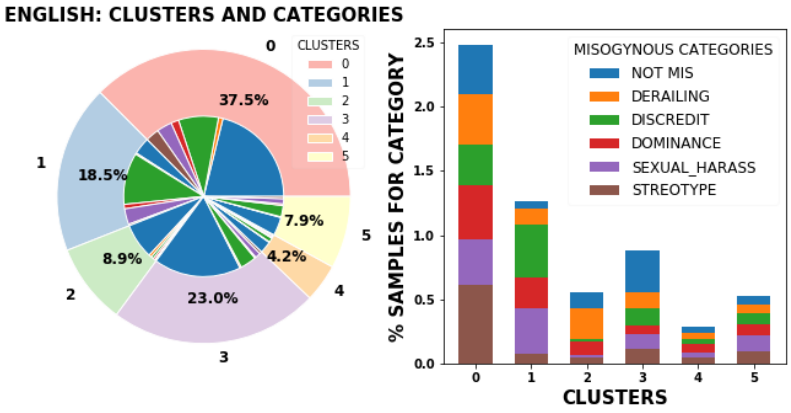

Figure 3: Example of misogynistic behaviour categories among 6 clusters using K-means on the English data set.

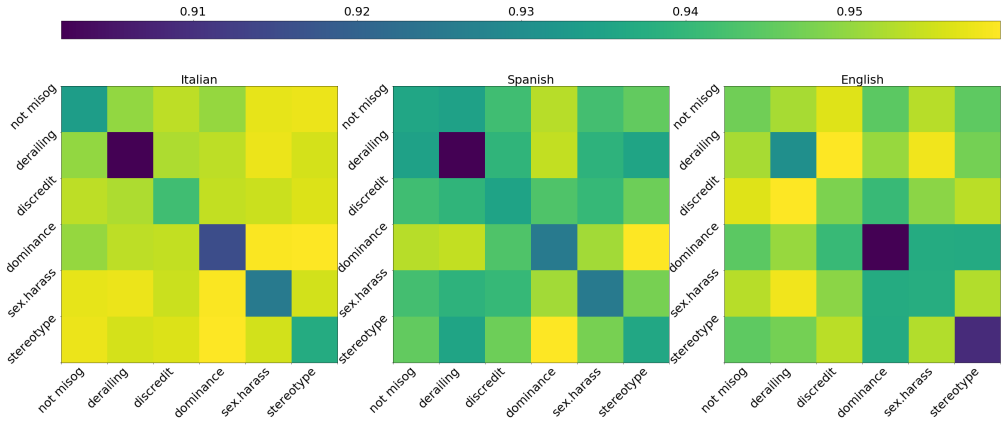

Figure 4: From left to right: cosine inter-category similarities computed respectively for Italian, Spanish and English data sets.

the tweets of the data sets considering the *Hurtex* categories of Table 3.

In order to give an answer to RQ3, we calculated the Macro F1-score considering the 5 misogynistic behaviour categories, and after merging the most similar ones: *dominance* with *discredit* (domin+discr) and *derailing* with *sexual harassment* (der+sex har). We finally show the confusion matrices of the best performing models.

5.2 Results

Table 4 shows the performance of our models on the test sets in terms of Macro F1-score. Our best results overcome the official results for the sub-task B both of EVALITA and IberEval: Macro $F_1(misogyny\_category)$ from 0.534 to 0.694 for Italian, from 0.339 to 0.347 for Spanish and from 0.361 to 0.470 for English, respectively. The results in general are not high, especially for English and Spanish. This is due to the under represented misogynistic behaviour categories in all the training sets, such as in the case of *derailing* and *dominance*, that turned out to be very difficult to detect for the models. With regard to the choice of the feature vector to use (BoW vs. tf-idf), we compared the macro F1-scores obtaining significance levels above the threshold of 5%. Therefore, it is not possible to say what technique is better. Figure 5 shows the confusion matrices of the models that obtained the best results (Linear SVM for Italian and Spanish, while for English tf-idf combined with SGD_selected, which differs from the standard SGD for the parameters settings, optimized through a python

	Italian			
Model	**BoW**	**BoW[H]**	**tf-idf**	**tf-idf[H]**
Lin SVM	0.576	0.251	**0.694**	0.254
Rbf SVM	0.57	0.249	0.222	0.245
LR	0.573	0.25	0.552	0.242
SGD	0.596	0.06	0.51	0.251
SGD_ selected	0.601	0.25	0.588	0.241
DT	0.571	0.247	0.519	0.251
	Spanish			
Model	**BoW**	**BoW[H]**	**tf-idf**	**tf-idf[H]**
Lin SVM	0.29	0.206	**0.347**	0.206
Rbf SVM	0.306	0.206	0.021	0.206
LR	0.241	0.002	0.315	0.002
SGD	0.257	0.0	0.209	0.0
SGD_ selected	0.248	0.002	0.295	0.002
DT	0.227	0.002	0.248	0.002
	English			
Model	**BoW**	**BoW[H]**	**tf-idf**	**tf-idf[H]**
Lin SVM	0.383	0.2	0.468	0.2
Rbf SVM	0.322	0.197	0.009	0.197
LR	0.353	0.0	0.452	0.0
SGD	0.286	0.0	0.305	0.0
SGD_ selected	0.339	0.0	**0.47**	0.0
DT	0.259	0.0	0.271	0.0

Table 4: F_1 for each misogynistic behaviour category and Macro $F_1(misogyny\_category)$ of the different models with the BoW and ft-idf feature vectors. We compare the results using the list of words obtained from the tweets of the data sets with the one built using [H]urtlex.

library[1]. We can observe that for the Italian data set the most problematic category is *derailing*, for which half samples are misclassified as not-misogynist. Moreover, the 11% of tweets labeled as *dominance* are classified as *derailing* and the 16% of tweets from *sexual harassment* are misclassified as *stereotype*. Regarding Spanish, the most often misclassified misogynistic behaviour category is *stereotype*, mainly confused with *discredit*, which in turn is often confused with *dominance*. We found very similar misclassification patterns in the English data set. These results are particularly interesting because in line with what we found previously. Since, especially for the Spanish and the English data sets, there is a certain similarity among *discredit* and *dominance*, *derailing* and *sexual harassment*. Therefore, decided to merge these two pairs of misogynistic behaviour categories. Table 5 shows the obtained results with tf-idf feature vectors. We can see that the F_1 scores always increase, especially for the English data set. The best Macro F1 scores in this case are respectively 0.767, 0.45 and 0.663 for the three data sets, being 0.691, 0.347 and 0.47 before merging them.

6 Discussion

The aim of this study is to understand the reasons behind the low scores obtained by the participants on the misogynistic behaviour classification at the AMI shared task, proposed at the EVALITA and IberEval evaluation campaigns in 2018.

To have a better understanding of the problem, we first studied the categories distribution for the different data sets (Figure 1). First of all, we note that there are many inhomogeneities among the misogynistic behaviour categories in the data sets. *Derailing* is the most obvious example of underrepresented category, followed by *dominance*. *Discredit* and *dominance* have a greater percent-

[1]available at `https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectFromModel.html`

	Italian				
Model	**Not Mis**	**domin+discr**	**der+sex_har**	**stereotype**	**MacroF1**
Lin SVM	0.699	0.709	0.74	0.851	**0.767**
Rbf SVM	0.055	0.988	0.041	0.069	0.366
LR	0.754	0.715	0.716	0.817	0.749
SGD	0.776	0.642	0.71	0.777	0.71
SGD_selected	0.735	0.739	0.716	0.84	0.765
DT	0.674	0.485	0.574	0.834	0.631
	Spanish				
Model	**Not Mis**	**domin+discr**	**der+sex_har**	**stereotype**	**MacroF1**
Lin SVM	0.656	0.554	0.649	0.118	0.44
Rbf SVM	0.981	0.129	0.0	0.0	0.043
LR	0.781	0.666	0.509	0.176	**0.45**
SGD	0.853	0.669	0.281	0.118	0.356
SGD_selected	0.764	0.674	0.474	0.059	0.402
DT	0.536	0.607	0.368	0.118	0.364
	English				
Model	**Not Mis**	**domin+discr**	**der+sex_har**	**stereotype**	**MacroF1**
Lin SVM	0.13	0.54	0.6	0.85	**0.663**
Rbf SVM	0.996	0.03	0.0	0.0	0.01
LR	0.254	0.672	0.564	0.721	0.652
SGD	0.537	0.706	0.4	0.179	0.428
SGD_selected	0.156	0.675	0.618	0.7	0.665
DT	0.493	0.347	0.382	0.15	0.293

Table 5: F_1 for each misogynistic behaviour category and Macro $F_1(misogyny\_category)$ for each model (using tf-idf feature vectors) on the Italian, Spanish and English data sets where *derailing* was merged with *sexual harassment* (der+sex har), and *dominance* with *discredit* (domin+discr).

Language	**BestTeam**	**OurModel**
Italian	0.555	**0.694**
Spanish	0.339	**0.347**
English	0.292	**0.470**

Table 6: Macro $F_1(misogyny\_category)$ of the best performing teams participating in the AMI shared task at EVALITA and IberEval in Italian, English and Spanish data sets vs., our best results (see Table 4).

age in the English and the Spanish data sets compared to the Italian one. The remaining categories (*derailing*, *sexual harassment* and *stereotype*) are present in a smaller percentage. We performed an unsupervised cluster analysis on the data sets, using K-means, Spectral Clustering and Agglomerative Clustering as algorithms. We measured the Silhouette and the Homogeneity scores by varying the number of clusters from 6 (not misogynistic plus the 5 misogynistic behaviour categories) down to 4 (not misogynistic, *stereotype*, *dominance + discredit*, *derailing + sexual harassment*). The obtained Silhouette scores from the cluster analysis are low, showing that the clusters are overlapping. The Homogeneity scores are near to 0, which indicate that the misogynistic behaviour categories are not easily separable in the different clusters (Figure 4), i.e., the clustering methods are not able to find hidden patterns useful to separate these categories. This is not surprising since there is a high overlap in some of them. The annotation of the tweets

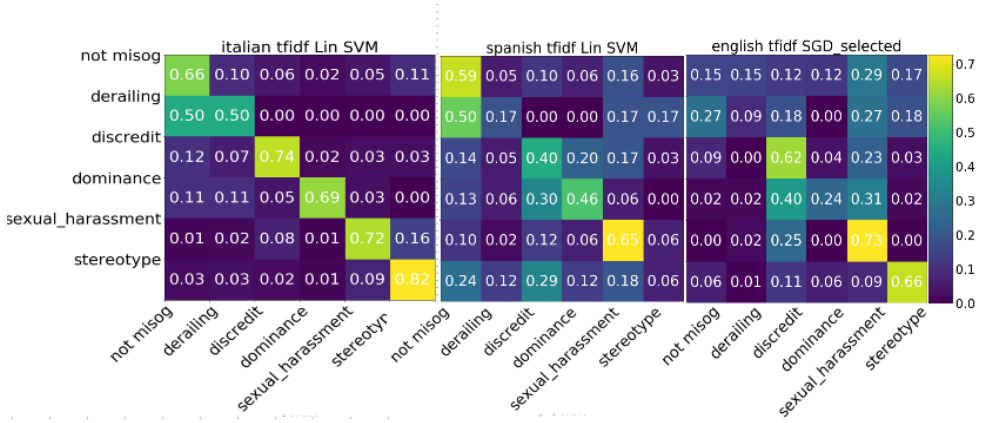

Figure 5: From left to right: confusion matrices of the models on the test sets for Italian, Spanish and English. On the x axis we have the predicted labels, while on the y axis the true ones.

may be highly subjective in some of the cases and consequently also for the models is difficult distinguish among the misogynistic behaviour categories. Doing the reverse, i.e., computing the inter-category cosine similarity among the categories, we noticed that in many cases the distance between the tweets belonging to a given category and those belonging to other misogynistic behaviour categories was pretty small. That is, in all the data sets some of the categories were closer: *derailing* to *sexual harassment*, and *discredit* to *dominance*. The best model for Italian and Spanish used tf-idf together with Lin SVM, while for English the best classification model was SGD_selected. Our highest F_1 scores are 0.694 for Italian, 0.47 for English and 0.347 for Spanish. The teams which achieved the best results respectively in the EVALITA and IberEval challenges are *bakarov* for Italian (0.555) (Bakarov, 2018), *himami* for English (0.292) (Ahluwalia et al., 2018) and *14-exlab* (0.339) (Pamungkas et al., 2018) for Spanish. Our best performance obtained better results in all the three languages (see Table 6).

Finally, what depicted in Figure 5 is in agreement with the considerations previously made on the inter-category distance matrices. Since *derailing* and *dominance* turned out to be the hardest categories to be detected, and given that the confusion matrices show that they are mostly misclassified as *sexual harassment* and *discredit*, respectively, we decided to merge them. As result, the F_1 scores improved significantly for all the data sets, increasing (for the best configurations) from 0.694 to 0.767 for Italian, from 0.47 to 0.663 for English, from 0.347 to 0.45 for Spanish.

7 *Conclusions*

In this paper we studied the multilingual Twitter data sets of misogynous texts released for the AMI shared task at EVALITA and IberEval 2018. Our main purpose was providing some insights on the fine-grained misogynistic behaviour, also in light of the difficulties encountered by the participants in identifying such categories. We found many inhomogeneities among the categories, which surely represented an important source of bias for the models, as it could be seen also from the confusion matrices. With respect to RQ1 and RQ2: from the clustering experiments on each data set, low Silhouette scores indicated the difficulty in separating the tweets of the different misogynistic behaviour categories (each cluster contained overlapping of categories). We also obtained very low Homogeneity scores indicating the same. We trained several machine learning algorithms improving the results of the best performing teams in both EVALITA and IbelEval. With respect to RQ3: since we found that tweets belonging to some of the categories are near in the features space for the data sets in the three languages, we trained and tested our models after merging two pairs of overlapping categories (in accordance with the insights from the inter-category cosine distance matrices and the confusion matrices). This allowed to improve even further the already good results that we obtained with the 5 misogynistic behaviour categories.

Acknowledgements

The work of S. Lazzardi was partially carried out at the Universitat Politècnica de València within the framework of the Erasmus+ program, Erasmus Traineeship 2018/19 funding. The work of P. Rosso was partially funded by the Spanish MICINN under the research project MISMIS-FAKEnHATE on Misinformation and Miscommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31). The work of V. Patti was partially funded by the research projects "STudying European Racial Hoaxes and sterEOTYPES" (STERHEOTYPES, under the call "Challenges for Europe" of VolksWagen Stiftung and Compagnia di San Paolo) and "Be Positive!" (under the 2019 "Google.org Impact Challenge on Safety" call).

References

Ahluwalia, R., H. Soni, E. Callow, A. C. Nascimento, and M. De Cock. 2018. Detecting hate speech against women in english tweets. In *EVALITA@ CLiC-it*.

Anzovino, M., E. Fersini, and P. Rosso. 2018. Automatic Identification and Classification of Misogynistic Language on Twitter. In M. Silberztein, F. Atigui, E. Kornyshova, E. Métais, and F. Meziane, editors, *Natural Language Processing and Information Systems*, pages 57–64, Cham. Springer International Publishing.

Bakarov, A. 2018. Vector space models for automatic misogyny identification. *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:211.

Basile, V., C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June. ACL.

Bassignana, E., V. Basile, and V. Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS.

Farrell, T., M. Fernandez, J. Novotny, and H. Alani. 2019. Exploring misogyny across the manosphere in reddit. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, pages 87–96, New York, NY, USA. ACM.

Fersini, E., D. Nozza, and P. Rosso. 2018. Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Fersini, E., P. Rosso, and M. Anzovino. 2018. Overview of the task on automatic misogyny identification at IberEval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 214–228. CEUR-WS.org.

Frenda, S., B. Ghanem, M. Montes-y-Gómez, and P. Rosso. 2019. Online hate speech against women: Automatic identification of misogyny and sexism on Twitter. *Journal of Intelligent and Fuzzy Systems*, 36(5):4743–4752.

Fulper, R., G. L. Ciampaglia, E. Ferrara, Y. Ahn, A. Flammini, F. Menczer, B. Lewis, and K. Rowe. 2014. Misogynistic language on twitter and sexual violence. In *Proceedings of the ACM Web Science Workshop on Computational Approaches to Social Modeling (ChASM)*.

García-Díaz, J. A., M. Cánovas-García, R. Colomo-Palacios, and R. Valencia-García. 2021. Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings. *Future Generation Computer Systems*, 114:506 – 518.

Gasparini, F., I. Erba, E. Fersini, and S. Corchs. 2018. Multimodal classifi-

cation of sexist advertisements. In *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications, ICETE 2018 - Volume 1: DCNET, ICE-B, OPTICS, SIGMAP and WINSYS, Porto, Portugal, July 26-28, 2018*, pages 565–572. SciTePress.

Manne, K. 2017. *Down Girl: The Logic of Misogyny*. Oxford University Press.

Megarry, J. 2014. Online incivility or sexual harassment? conceptualising women's experiences in the digital age. *Women's Studies International Forum*, 47:46 – 55.

Pamungkas, E. W., V. Basile, and V. Patti. 2020. Misogyny Detection in Twitter: a Multilingual and Cross-Domain Study. *Information Processing & Management*, 57(6):102360.

Pamungkas, E. W., A. T. Cignarella, V. Basile, V. Patti, et al. 2018. 14-exlab@unito for AMI at ibereval2018: Exploiting lexical knowledge for detecting misogyny in English and Spanish tweets. In *3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2018*, volume 2150, pages 234–241. CEUR-WS.

Plaza-Del-Arco, F.-M., M. D. Molina-González, L. A. Ureña López, and M. T. Martín-Valdivia. 2020. Detecting Misogyny and Xenophobia in Spanish Tweets Using Language Technologies. *Rossana Damiano and Viviana Patti and Chloé Clavel and Paolo Rosso (Eds.), Special Section on Emotions in Conflictual Social Interactions, ACM Transactions of Internet Technology*, 20(2).

Poland, B. 2016. *Haters: Harassment, abuse, and violence online*. U of Nebraska Press.

Waseem, Z. and D. Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June. ACL.

Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 77-87 recibido 13-12-2020 revisado 21-01-2021 aceptado 31-01-2021

Escansión automática de poesía española sin silabación

Automatic Scansion of Spanish Poetry without Syllabification

Guillermo Marco Remón, Julio Gonzalo
nlp.uned.es, Research Group in NLP & IR, UNED, Madrid, Spain
{gmarco, julio}@lsi.uned.es

Resumen: En los últimos años, han surgido diversas herramientas de análisis métrico automático de poesía española. Estos sistemas se basan en complejos métodos de silabación y asignación de acentos, los cuales se apoyan en librerías de etiquetado gramatical, cuyo coste computacional es elevado. Este coste incrementa con el cálculo de ambigüedades métricas. El presente artículo parte de la hipotésis de que es posible llevar a cabo un análisis métrico informativo y preciso sin utilizar estos métodos. Se propone un algoritmo que realiza escansiones (número de sílabas, patrón métrico y tipo de verso) sin silabación. El algoritmo resuelve ambigüedades métricas y tiene en cuenta la compensación hemistiquial. Nuestros resultados indican que obtiene una mejora relativa sobre el estado del arte de un 2 % en la clasificación de patrones métricos en poesía de medida fija y un 25 % en poesía polimétrica. También se ejecuta 21 y 25 veces más rápido, respectivamente. Por último, se ofrece una aplicación de escritorio como herramienta para los investigadores de la poesía española.
Palabras clave: patrones métricos, escansión automática.

Abstract: In recent years, several systems of automated metric analysis of Spanish poetry have emerged. These systems rely on complex methods of syllabification and stress assignment, which use PoS-tagging libraries, whose computational cost is high. This cost increases with the calculation of metric ambiguities. However, it is possible to carry out an informative and accurate metric analysis without using these costly methods. We propose an algorithm that performs accurate scansion (number of syllables, stress pattern and type of verse) without syllabification. It addresses metric ambiguities and takes into account the hemistichs compensation. Our algorithm outperforms the current state of the art by 2 % in fixed-metre poetry, and 25 % in mixed-metre poetry. It also runs 21 and 25 times faster, respectively. Finally, a desktop application is offered as a tool for researchers of Spanish poetry.
Keywords: metrical patterns, automated scansion.

1 Introducción

En los últimos años, han surgido diversas herramientas de análisis métrico de poesía española. Dado un verso, extraen automáticamente el número de sílabas y su ritmo. Por ejemplo, a partir del verso:

(1) *Amigos, el amor me perjudica*
*A-**mi**-gos-el-a-**mor**-me-per-ju-**di**-ca*
11 — 2.6.10
(Julio Martínez Mesanza)

se determina que es un verso de once sílabas y su ritmo (también llamado esquema acentual o patrón métrico) es 2.6.10.

Para resolver este problema, los sistemas publicados hasta el momento se basan en complejos métodos de silabación y asignación de acentos, los cuales se apoyan en librerías de etiquetado gramatical, cuyo coste computacional es elevado. Como la escansión de un verso puede ser ambigua –los versos del Ejemplo 2 se pueden escandir a la manera (a) o a la manera (b)–, este coste incrementa con el

ISSN 1989-7553. DOI 10.26342/2021-66-6

cálculo de las ambigüedades métricas.

(2) *dentro de su fluir los manantiales*

(a) ***den**-tro-de-su-**fluir**-los-ma-nan-**tia**-les*
10 — 1.5.9

(b) ***den**-tro-de-su-flu-**ir**-los-ma-nan-**tia**-les*
11 — 1,6,10
(Claudio Rodriguez)

Los versos por encima de las once sílabas se dividen en semiversos o hemistiquios para su escansión. Esta separación introduce fenómenos métricos que afectan al cómputo silábico, y por ello se han de considerar para la correcta determinación del ritmo. Los métodos actuales no tienen en cuenta estos fenómenos.

Este trabajo parte de la hipótesis de que es posible llevar a cabo un análisis métrico informativo y preciso sin utilizar métodos de silabación ni librerías de etiquetado gramatical. En la Sección 2 comenzaremos definiendo el problema de la escansión e introduciendo el marco teórico necesario para abordar su solución, y en la Sección siguiente resumiremos el estado del arte.

A partir del conocimiento métrico expuesto, en la Sección 4 se desarrolla el sistema análisis métrico automático que simplifica el problema de la medida del verso gracias a que prescinde de realizar silabación. El algoritmo tiene en cuenta la compensación hemistiquial y resuelve las ambigüedades métricas derivadas de sinalefas, dialefas, sinéresis y diéresis, sin perder precisión ni información sobre la decisión del analizador. La salida de este método ofrece el número de sílabas, el esquema métrico, el esquema métrico sin acentos extrarrítmicos, su nombre, su grado de coincidencia con el esquema métrico sin acentos extrarrítmicos, así como la forma del verso que ha llevado al sistema a clasificarlo en ese tipo: marca sinalefas, dialefas, sinéresis y diéresis.

En la Sección 5 se evalúa el sistema sobre conjuntos de datos de poesía de metro fijo y mixto. En nuestra experimentación, el algoritmo obtiene una mejora sobre el actual estado del arte de un 2 % en la clasificación de patrones métricos sobre poesía de medida fija y de un 25 % sobre poesía polimétrica. También se ejecuta 21 y 25 veces más rápido, respectivamente. Al final de la sección se analizan los errores cometidos por el algoritmo.

Por último, en las Secciones 6 y 7 se describe *Jumper* una aplicación de escritorio que puede servir como herramienta para los investigadores de la poesía española, y se extraen conclusiones.

2 Definición del problema

Un verso es una serie de sílabas acentuadas y no acentuadas delimitada por pausas métricas (Caparrós, 2001). La distribución de los acentos sobre esta serie determina el patrón métrico. Así, la escansión consiste en clasificar los versos por su medida (número de sílabas) y su patrón métrico.

La sílaba es la unidad estructural de la palabra. En español, el núcleo de una sílaba es siempre vocálico (RAE, 2009, 8.1a). No obstante, la vocal también desempeña funciones de margen silábico en los diptongos; se forman combinando las vocales abiertas con cerradas, cerradas con abiertas, y las cerradas entre sí. Son un conjunto de 14 realizaciones: *ai, au, ei, eu, oi, ou, ui, iu, ia, ua, ie, ue, io, uo*.

En general, las palabras en español tienen sólo un acento, con algunas excepciones como los adverbios acabados en *-mente*. En función de dónde recae su acento, se clasifican y acentúan gráficamente de la siguiente manera:

- Palabras oxítonas o agudas: cuando la sílaba acentuada es la última de la palabra. Aunque existen algunas normas adicionales (RAE, 2010, 3.4.1.2.1), principalmente se acentúan con un signo gráfico cuando terminan en los grafemas de *n*, *s* o *vocal*.
- Paroxítonas o llanas: cuando la sílaba acentuada es la penúltima de la palabra. Se acentúan con un signo gráfico cuando terminan en un grafema distinto de *n*, *s* o *vocal*.
- Proparoxítonas o esdrújulas: cuando la sílaba acentuada es la antepenúltima de la palabra. Se acentúan con un signo gráfico siempre.

De manera homónima, un verso, por la posición de su última sílaba acentuada, recibe el nombre de oxítono, paroxítono o proparoxítono. De cada uno de ellos, resultan unos

fenómenos métricos constantes que afectan al número de sílabas (Quilis, 1984):

- Cuando un verso es oxítono, se cuenta una sílaba más.
- Cuando un verso es paroxítono, se cuenta las sílabas reales existentes.
- Cuando un verso es proparoxítono, se cuenta una sílaba menos.

Además de estas denominaciones, la tradición poética ha clasificado los versos en función de su patrón métrico. Un verso cuenta con un número finito de acentos. Un endecasílabo, por ejemplo, puede contar con cinco acentos rítmicos o menos; un octosílabo hasta cuatro. Por lo tanto, el número de combinaciones posibles también son finitas y asentadas en la tradición poética. Estos esquemas métricos tradicionales son los que el poeta tiene en mente cuando elige un metro. En el presente trabajo, se toma la clasificación de Pou (2020), por ser la más sistemática y exhaustiva. En la Tabla 1 se recoge la tipología del endecasílabo.

No obstante, en la realización de un verso, la distribución de los acentos puede ser variada y no coincidir exactamente con los ofrecidos en la clasificación. En el Ejemplo 3, se da el patrón métrico 1.6.(7).10, el cual es posible y frecuente en el endecasílabo, pero no se observa en la Tabla 1. Se advierte, sin embargo, que tiene los acentos característicos en 1.6.10 del ritmo enfático puro, siendo el acento en séptima extrarrítmico.

(3) *Siempre la claridad viene del cielo*
***Siem**-pre-la-cla-ri-**dad-vie**-ne-del-**cie**-lo*
11 — 1.6.(7).10
(Claudio Rodríguez)

Siguiendo este criterio, cada verso se puede clasificar en función de la proximidad a cada uno de los tipos: se observan los acentos característicos del tipo de verso y se tratan los no coincidentes como extrarrítmicos.

Cuanto mayor es el número de acentos, más complejo y rico es el verso. Sin embargo, cuando supera las once sílabas, se rompe en semiversos o hemistiquios. Por ejemplo, un verso de 14 sílabas, también llamado alejandrino, se suele dividir en hemistiquios de 7+7. Es decir, el verso en español tiene el límite rítmico del endecasílabo.

Endecasílabos		
— Heroicos	2.6.10	puro
	2.4.6.8.10	pleno
	2.4.6.10	corto
	2.6.8.10	largo
	2.4.10	difuso
— Melódicos	3.6.10	puro
	1.3.6.8.10	pleno
	3.6.8.10	largo
	1.3.6.10	corto
— Sáficos	4.8.10	puro
	1.4.8.10	puro pleno
	1.4.6.8.10	pleno
	4.6.10	corto
	1.4.6.10	corto pleno
	4.6.8.10	largo
	2.4.8.10	largo pleno
	4.10	difuso
	1.4.10	difuso pleno
— Dactílicos	4.7.10	puro
	1.4.7.10	pleno
	2.4.7.10	corto
— Enfáticos	1.6.10	puro
	1.6.8.10	pleno
— Vacíos	6.10	puro
	6.8.10	pleno
	1.4.10	pleno
	2.4.10	heroico

Tabla 1: Clasificación del verso de once sílabas (Pou, 2020, p. 118).

Precisamente, por esa subdivisión en hemistiquios, se puede realizar una escansión separada de cada uno de los semiversos. Así, en el Ejemplo 4 se contempla que el primer hemistiquio es proparoxítono, y resta una sílaba al cómputo. El segundo es oxítono y suma una. El resultado final es un verso alejandrino.

(4) *Oh, qué frescor, qué música / de chopos de estación*
14 = 8 - 1 / 6 + 1
(Juan Ramón Jiménez)

La silabación es la segmentación de una palabra en sus sílabas constituyentes. Esta tarea viene determinada por la estructura morfológica del español (Española, 2009, 1.3.4a).

Es importante reparar en que la tarea de

silabación es distinta de la tarea de contar sílabas. Por ejemplo, dada la palabra *bolígrafo*, la silabación resultaría *bo-lí-gra-fo*, donde se ha hecho un esfuerzo morfológico por establecer la relación entre grafemas de la palabra. No obstante, el cómputo sílabico no informa de la relación entre los grafemas, sino que es una cuantificación: 4.

A parte de estas consideraciones generales, los versos son sometidos por los poetas a varios recursos métricos que afectan al cómputo silábico e introducen ambigüedad en la determinación del patrón rítmico. Los más habituales son la sinalefa, dialefa, sinéresis y diéresis.

Sinalefa es la pronunciación en una sola sílaba de dos o más vocales que se encuentran en palabras distintas (Ejemplo 5).

(5) *Creía que te había dicho adiós*
*Cre-**í**-a-que-te‿ha-**bí**-a-**di**-cho‿a-**diós***
11 — 2.6.8.10
(Amalia Bautista)

En los versos de arte mayor superiores a once sílabas, la pausa hemistiquial rompe la sinalefa. En el Ejemplo 6 se rompe la sinalefa entre *mente* y *entre*.

(6) *Escucho solamente entre las voces una*
*Es-**cu**-cho-**so**-la-**men**-te-entre-las-**vo**-ces-**u**-na*
7 + 7 = 14
(Antonio Machado)

La dialefa es la ruptura en dos sílabas de dos vocales de palabras distintas, es decir, lo contrario de la sinalefa.

La sinéresis consiste en la pronunciación en una sola sílaba de dos o más vocales que normalmente forman dos sílabas. "Hermosas ninfas que en el río metidas" (Garcilaso): "Hermosas ninfas que en el **rio** metidas" (11) (Pou, 2020, Capítulo III)

Por último, la diéresis consiste en la pronunciación de dos o más vocales que normalmente se realizan como un diptongo, convirtiéndolo en un hiato. Aunque suele venir explicitada por el poeta con el signo ¨. No siempre es así. Sirva de muestra el verso del Ejemplo 7, donde se encuentra una diéresis no indicada.

(7) *Todas las tardes se muere un niño*
*to-das-las-**tar**-des-se-**mü**-e-re‿un-**ni**-ño*
11 — 1.4.8.(9).10
(Federico García Lorca)

Además, la resolución de ambigüedades no es exacta. En el verso del Ejemplo 8, existen dos posibles escansiones, ambas válidas. En cualquier caso, de acuerdo con lo dicho anteriormente, se podría argumentar que el esquema 2.4.(7).8.10 con acento extrarrítmico en 7ª es natural del endecasílabo sáfico largo pleno y es, con un grado alto de certidumbre, el que el autor trataba de conseguir; por el contrario, el patrón 3.7.8.10 es artificial y su sonoridad está cuestionablemente alejada del endecasílabo. Hacia este tipo de decisiones se orienta el sistema presentado.

(8) *Juez Elisio, que de un verde probo*

(a) *Ju-**ez**-E-**li**-sio-que-**de‿un-ver**-de-**pro**-bo*
11 — 2.4.(7).8.10

(b) ***Juez**-E-**li**-sio-que-de-**un-ver**-de-**pro**-bo*
11 — 3.7.8.10
(Lope de Vega)

3 Trabajos relacionados

La métrica, dentro de la filología, es un tema fundamental en el análisis literario de un poema. Entre los trabajos clásicos en el campo se encuentran los manuales de Bello (1859), Tomás (1956) y Quilis (1984). Más recientemente se publicó el diccionario y manual de Domínguez Caparrós (Caparrós, 1993; Caparrós, 2001). El presente trabajo se apoya en las últimas investigaciones de métrica española expuestas por Pou (2020)

En cuanto a los algoritmos de escansión automática, en los últimos años se encuentra el trabajo de Navarro-Colorado (2017). Se trata de un sistema basado en reglas que emplea el analizador morfológico Freeling (Padró and Stanilovsky, 2012). Se centra en la resolución de ambigüedades; marca sinalefas y diéresis, pero no trata las sinéresis. Procesa las ambigüedades empleando una base de conocimiento de los distintos patrones métricos extraída en un análisis estadístico; aunque ofrece información adicional sobre el corpus, este análisis no es estrictamente necesario para realizar escansión, pues lo patrones vienen ya dados por la propia preceptiva de

la métrica. Además, el sistema se evalúa sólo sobre versos de métrica fija.

Por otro lado, Agirrezabal, Alegria, and Hulden (2017) entrenan una red de neuronas bidireccional LSTM a nivel de carácter. El sistema predice patrones métricos a partir de una transformación enriquecida de la entrada, que incluye silabación. La salida no informa sobre la decisión del sistema, por lo que no se pueden conocer las ambigüedades que ha detectado. Se reportó una tasa de acierto de 0.91 sobre el corpus de metro fijo.

El sistema más reciente publicado es Rantanplan (de la Rosa et al., 2020). Se trata de un método basado en reglas que emplea un sistema de silabación del 99.99 % de precisión. La señalización de acentos se apoya en esta silabación y en el etiquetado gramatical de la librería Spacy (Honnibal and Montani, 2017). Como indican los autores, el sistema está limitado por este modelo estadístico: comete errores ocasionales asignando los acentos y, desde el punto de vista del coste computacional, tan sólo su carga retrasa 18 segundos el análisis de acuerdo con las pruebas de los investigadores. No tienen en cuenta los efectos de compensación y dialefa hemistiquiales, lo cual perjudica la tasa de acierto (*accuracy*) sobre el corpus de poemas polimétricos.

Todos los métodos anteriores utilizan de una manera u otra métodos de silabación. Tampoco tienen en cuenta los fenómenos métricos de los versos de más de once sílabas.

4 Jumper: algoritmo de análisis métrico sin silabación

En esta Sección se presenta *Jumper*, el método de análisis métrico automático desarrollado. Se divide a su vez en tres subsecciones: análisis de palabra, verso y poema.

4.1 Módulo 1: Análisis de palabra: cómputo de sílabas y acentos

Como se ha dicho en la Introducción, en español, el núcleo de una sílaba es siempre vocálico. La unidad de la sílaba viene dada por la vocal. Por lo tanto, para contar sílabas de manera eficiente, basta con contar las vocales de una palabra, teniendo en cuenta los diptongos. De hecho, no es necesario localizar los hiatos explícitamente: los diptongos en español son un conjunto finito de 14 realizaciones. Por lo tanto, dada la aparición contigua de dos vocales, basta comprobar si es un diptongo para no contar esa siguiente vocal como sílaba.

Además, este mismo procedimiento permite localizar el acento. Se asume que la palabra sigue las reglas de acentuación gráfica del español (RAE, 2010, 3.4). Así, mientras se cuentan las vocales, se guarda la posición en la que se ha encontrado una vocal acentuada gráficamente. Si tras el cómputo no se ha encontrado el acento, se aplican las reglas de acentuación básicas del español para hallarlo.

En cuanto se conoce si la palabra es aguda, llana o esdrújula, se puede calcular fácilmente la compensación cuando se encuentra al final del verso.

Gracias a este método, al tener en cuenta la vocal acentuada explícitamente, se puede también localizar el acento cuando el poeta emplea el recurso de la sístole.[1]

Cabe decir que, de momento, se está operando a nivel de palabra. En cuanto se traten de resolver las ambigüedades métricas de un verso producidas por las sinéresis sí será necesario dar un tratamiento específico a los hiatos.

4.2 Módulo 2: Análisis de verso

Una vez se ha diseñado la función que devuelve el número de sílabas y el lugar de sus acentos individualmente, se quiere extender el análisis al verso entero. Este módulo es el núcleo del método, y es el que entraña más complejidad. Se divide, a su vez, de tres submódulos. El primero se encarga del cómputo de sílabas y acentos; el segundo, de la compensación hemistiquial para los versos mayores de once sílabas y, el tercero, de resolución de ambigüedades métricas.

4.2.1 Submódulo 1: Cómputo de sílabas y acentos de un verso

El primer submódulo funciona de la siguiente manera. Dado un verso, se convierte en una lista de palabras. Para cada palabra de la lista, se calculan, con la función expuesta en la Sección 4.1, el número de sílabas, el lugar del acento en esas sílabas y el factor de compensación. Este factor es -1 para las palabras esdrújulas, 0 para las llanas y 1 para las agudas. Se suma el número de sílabas de la palabra al número total de sílabas del

[1] Sístole consiste en adelantar una o más sílabas el acento normal de una palabra; jilguero - jílguero. (Pou, 2020, Capítulo III)

verso, teniendo en cuenta, si la hubiera, la sinalefa con la palabra siguiente. Se calcula la posición del acento y se añade a la lista del patrón métrico. Si la palabra es un adverbio en *-mente* se tienen en cuenta sus dos acentos. Si la palabra es átona, no se añade a la lista de acentos.

En este primer submódulo se calcula la medida del verso teniendo en cuenta todas las sinalefas. Posteriormente, si hay alguna que no habría que tener en cuenta, se resuelve en el módulo de ambigüedades.

4.2.2 Submódulo 2: Compensación hemistiquial

Al comienzo del análisis de un verso, no se conoce cuántas sílabas va a tener. Si tras el cómputo del primer submódulo resulta un verso de más de once sílabas, se hace una llamada recursiva sobre la función indicando que ahora se tenga en cuenta la compensación hemistiquial, activando así el segundo submódulo. Si durante este segundo cómputo se hallan hemistiquios, se tiene en cuenta el factor de compensación de cada semiverso y la ruptura de la sinalefa por la pausa hemistiquial, si la hubiera. Se considera que una palabra se encuentra en la frontera de un hemistiquio cuando el número de sílabas del verso en proceso sumado al número de sílabas de la palabra y su factor de compensación resultan la medida esperada del hemistiquio (por ejemplo, 7 para el alejandrino).

4.2.3 Submódulo 3: Resolución de ambigüedades

Para calcular correctamente el patrón métrico de un verso, hay que tener en cuenta los recursos métricos empleados en él. Es necesario detectar dialefas, sinéresis y diéresis, sus posibles combinaciones, y obtener el mejor candidato.

Por lo tanto, el submódulo de resolución de ambigüedades tiene dos tareas. La primera consiste en detectar las posibles ambigüedades del verso y generar candidatos con ellas. La segunda consiste en elegir, entre los posibles candidatos, el mejor.

La activación del submódulo de detección de ambigüedades se indica desde el módulo superior de análisis de poema. Una vez indicada, por eficiencia, la detección se hace al mismo tiempo que el cómputo del Submódulo 1. Cuando se detecta una sinalefa, se genera a la vez un candidato con dialefa. Cuando se detecta un diptongo, se genera un candidato con diéresis. Cuando se detecta un hiato, se genera un candidato con sinéresis. Una vez se han almacenado todos los candidatos con las posibles ambigüedades, se combinan: por ejemplo, un verso puede presentar, al mismo tiempo, una dialefa y una sinéresis. Todos los candidatos se etiquetan con su recurso métrico para así conservar la decisión del submódulo de resolución de ambigüedades.

En la segunda tarea, se elige el mejor candidato. Como se ha expuesto en la introducción, los patrones métricos de los versos son finitos y sus posiciones se aproximan a los tipos de versos asentados en la tradición. Por lo tanto, se compara la proximidad del vector del patrón rítmico extraído con cada uno de los tipos asentados en la preceptiva literaria; en función de esa proximidad, se elige el mejor candidato.

Aunque probar todas las posibles ambigüedades podría parecer costoso computacionalmente, es rápido por dos razones. La primera es que los versos ambiguos son sólo un pequeño subconjunto del corpus. La segunda es que, gracias al sencillo algoritmo de cómputo de sílabas y acentos, el coste de volver a computar un verso es muy bajo.

4.3 Módulo 3: Análisis de poema

Dado un poema, se convierte en una lista de versos. Se realiza la escansión de cada uno de ellos por medio del módulo de análisis a nivel de verso, con el sistema de resolución de ambigüedades desactivado.

Una vez calculadas todas las medidas y patrones rítmicos, se observa la frecuencia de los números de sílabas. Pueden darse dos casos. El primero, cuando sólo hay una medida frecuente: en este caso, se concluye que es un corpus de metro fijo. Si la lista de medidas frecuentes es mayor de uno (por ejemplo, en la poesía contemporánea, son habituales los poemas de versos polimétricos de 7, 11, y 14 sílabas), se concluye que es un corpus de poemas polimétricos. Una vez se tiene la lista de medidas frecuentes, se vuelven a computar los versos que no coincidan con ellas, ahora con el módulo de resolución de ambigüedades activado. La única diferencia entre poesía de metro fijo y mixto es que, para esta última, la lista de medidas frecuentes se actualiza en un contexto de un número n de versos. El tamaño del contexto es el único hiperparámetro del sistema. Por defecto, se establece en 14, ya que el número de versos del poema por

antonomasia, el soneto.

Una vez resueltas las ambigüedades, se almacena en una tabla cuyas columnas son: verso, verso etiquetado, número de sílabas, patrón métrico en forma de vector de enteros, patrón métrico en forma de vector de enteros del tipo de verso sin acentos extrarrítmicos, nombre del verso y, por último, ratio de coincidencia entre el patrón calculado y el patrón sin acentos extrarrítmicos. La salida del sistema se muestra en la Figura 1.

5 Evaluación

En los últimos años se ha estandarizado el uso del corpus anotado de sonetos (estrofa de 14 versos de once sílabas) del Siglo de Oro (Navarro-Colorado, Lafoz, and Sánchez, 2016) para la evaluación de los sistemas de escansión sobre poemas de medida fija. Originalmente, consistía en un conjunto de 1400 versos etiquetados manualmente con su patrón métrico; posteriormente se ampliaron a 10268. Está disponible en la herramienta de descarga de corpus Averell del proyecto POSTDATA.[2]

Asimismo, para métrica de medida mixta, de la Rosa et al. (2020) introdujeron recientemente un corpus elaborado a partir de la antología de Antonio Carvajal (1983). Se compone de 4378 versos, tanto de arte mayor como de arte menor, etiquetados manualmente. Por razones de derechos de autor, no es de dominio público; no obstante, sus desarrolladores nos permitieron su uso para esta experimentación.

Los sistemas se evalúan en función de su acierto (*accuracy*) establecido de forma binaria para cada verso: el acierto es 1 si se han identificado correctamente todos los acentos anotados del verso, y 0 en caso contrario. El rendimiento del sistema se mide como la tasa de acierto sobre el total de versos del corpus.

Todos los experimentos de nuestro sistema se han realizado sobre un ordenador equipado con la misma configuración que se empleó para las evaluaciones de los distintos métodos comparados en Rantanplan: procesador Intel® Core™ i7-8550U CPU @ 1.80GHz y 16GiB de memoria RAM DDR4.

El acierto (*accuracy*) se reporta en el intervalo [0,1] con dos cifras significativas. El tiempo se indica en segundos.

[2]https://github.com/linhd-postdata/averell

Al sistema se le ha dado el nombre de *Jumper*.

5.1 Evaluación sobre poemas de medida fija

La Tabla 2 recoge la tasa de acierto (*accuracy*) y el tiempo de cada uno de los métodos sobre el corpus de 10268 versos de medida fija elaborado por Navarro-Colorado, Lafoz, and Sánchez (2016). Nuestro sistema, Jumper, obtiene una tasa de acierto de 0,95, lo que supone una mejora del 2,2 % respecto al actual estado del arte establecido por Rantanplan. Además, se ejecuta 21 veces más rápido que éste.

Estos resultados indican que prescindir de silabación supone una mejora en el enfoque de la solución al problema. Gracias a que no se requiere la compilación de numerosas expresiones regulares para la separación de sílabas ni los pesados modelos de PoS-Tagging para la asignación de acentos, el tiempo es menor. La velocidad conseguida permite realizar una desambiguación exhaustiva en poco tiempo, lo que influye también en la mejora de la tasa de acierto.

Método	***Accuracy***	**Tiempo**
Navarro-Colorado	0,91	16787s
Rantanplan	0,93	53s
Jumper (nuestro)	**0,95**	**2,5s**

Tabla 2: *Accuracy* y tiempo sobre el corpus de 10268 versos de métrica fija de Navarro-Colorado. Los resultados de Navarro-Colorado y Rantanplan se han extraído de de la Rosa et al. (2020). Los de Jumper se han obtenido bajo la misma configuración de hardware.

Es interesante examinar el resultado de la evaluación sobre el subconjunto inicial de 1400 versos establecido por sus autores, ya que es el único corpus para esta tarea sobre el que se ha reportado una *inter-annotator agreement* (IAA) o tasa de acuerdo entre anotadores. La IAA sobre este subconjunto es de 0,96 (Navarro-Colorado, Lafoz, and Sánchez, 2016). Jumper obtiene un 0,95. No obstante, se ha detectado en el análisis de errores de la Sección 5.3.1 un sesgo en la anotación de estos primeros 1400 versos: la interjección "oh", frecuente en el Siglo de Oro, es considerada átona sólo en este subconjunto del

Análisis

N	Verso	Etiquetado	Sílabas	Acentos	Sin extrarrítmicos	Tipo	Coincidencia
1	Decir pestes de él tiene, sin duda,	decir pestes de él tiene sin duda	11	[2, 3, 6, 7, 10]	[2, 6, 10]	Endecasílabo heroico puro	80
2	un sólido prestigio literario	un sólido prestigio literario	11	[1, 2, 6, 10]	[2, 6, 10]	Endecasílabo heroico puro	90
3	-tacharlo de asesino, por ejemplo,	-tacharlo de asesino, por ejemplo,	11	[2, 6, 10]	[2, 6, 10]	Endecasílabo heroico puro	100
4	o compararlo con	o compararlo con	7	[4, 6]	[4, 6]	Heptasílabo sáfico puro	100
5	uno de esos ciclones con nombre de corista	uno de esos ciclones con nombre de corista	14	[1, 3, 6, 9, 13]	-	Alejandrino	100
6	que pasan y que dejan en los telediarios	que pasan y que dejan en los teledi~arios	14	[2, 6, 13]	-	Alejandrino	100
7	un paisaje de grandes palmeras derrocadas	un paisaje de grandes palmeras derrocadas	14	[1, 3, 6, 9, 13]	-	Alejandrino	100
8	y uralitas errantes,	y uralitas errantes,	7	[3, 6]	[3, 6]	Heptasílabo melódico puro	100
9	o simplemente lamentarlo a base	o simplemente lamentarlo a base	11	[2, 4, 8, 10]	[2, 4, 8, 10]	Endecasílabo sáfico largo pleno	100
10	de tardes y de otoños en pálidos jardines-,	de tardes y de otoños en pálidos jardines-,	14	[2, 6, 9, 13]	-	Alejandrino	100
11	pero ahora, con la mano en el poema,	pero óra con la mano en el poema	11	[2, 6, 10]	[2, 6, 10]	Endecasílabo heroico puro	100
12	os lo confieso: he sido siempre yo	os lo confieso: he sido siempre yo	11	[4, 5, 6, 8, 10]	[4, 6, 8, 10]	Endecasílabo sáfico largo	90
13	el que salió ganando de todos nuestros tratos.	el que salió ganando de todos nuestros tratos.	14	[4, 6, 9, 13]	-	Alejandrino	100

Figura 1: Salida del sistema. Observése la diéresis detectada en el décimo verso.

corpus. Así, únicamente añadiendo a la lista de palabras átonas esa interjección Jumper obtendría un 0,96. Por tanto, nuestro sistema alcanza el límite de precisión que se puede medir con este corpus, ya que iguala la tasa de acuerdo entre anotadores. Este conjunto también permite comparar el efecto del tiempo de arranque de los sistemas sobre conjuntos de datos de tamaño pequeño. Bajo la misma configuración, Jumper analiza los 1400 versos en 0,33 segundos frente a los 2356 del sistema de Navarro-Colorado y los 21 segundos de Rantanplan.

5.2 Evaluación sobre poemas de medida mixta

En la Tabla 3 se contempla la tasa de acierto (*accuracy*) sobre el corpus de métrica mixta. Jumper obtiene 0,82, lo que supone un 25 % de mejora relativa respecto al actual estado del arte establecido por Rantanplan. En cuanto al tiempo, se ejecuta 25 veces más rápido que éste.

Método	*Accuracy*	Tiempo
Navarro-Colorado	0,49	7484s
Rantanplan	0,65	27s
Jumper (nuestro)	**0,82**	**1,1s**

Tabla 3: *Accuracy* y tiempo de ejecución sobre el corpus de 4300 versos de métrica mixta de Antonio Carvajal. Los resultados de Navarro-Colorado y Rantanplan se han extraído de de la Rosa et al. (2020), donde ambos han sido ejecutados en el mismo entorno. Los de Jumper se han obtenido bajo la misma configuración de hardware.

Estos resultados son producto de los beneficios de prescindir de silabación y librerías de etiquetado gramatical, ya comentados en la Sección anterior. Pero, además, se añade otro factor en la mejora sobre este corpus: la consideración de los fenómenos métricos particulares de los versos de más de once sílabas. Esta consideración permite clasificar sus patrones métricos con mayor acierto. En el corpus de medida mixta de Carvajal se encuentran mezclados tanto versos de 7, 9 u 11 como de 14. El Ejemplo 9, tomado del conjunto de datos, Jumper lo clasifica correctamente como verso de 14 sílabas con ritmo 1.3.6.8.10.13, ya que el primer hemistiquio es oxítono y suma una sílaba. Rantanplan, sin embargo, lo clasifica como verso tridecasílabo de ritmo 1.3.6.7.9.12, al no tener en cuenta la compensación. Lo mismo ocurre con el fenómeno de ruptura de sinalefa entre hemistiquios.

(9) *una lucha común, y un descanso común*

(a) 6+1 / 6+1 = 14 — 1.3.6.8.10.13 (Jumper)

(b) 13 — 1.3.6.7.9.12 (Rantanplan) (Antonio Carvajal)

5.3 Análisis de errores

En esta Sección se profundiza en las circunstancias en las que Jumper falla en su clasificación de patrones métricos. En el corpus de medida fija se clasifican incorrectamente 546 de 10268 patrones. Sobre el corpus de medida mixta se clasifican incorrectamente 800 de 4378.

5.3.1 Errores sobre medida fija

Se ha tomado una muestra de 100 versos de los 546 clasificados incorrectamente. Tras su análisis manual, se contempla que hay principalmente cuatro fuentes de error.

La primera fuente de error (35 equivocaciones), atribuible al sistema, es la elección incorrecta del mejor candidato entre las posibles realizaciones de un verso ambiguo. Cuando un verso tiene dos escansiones posibles

igual de correctas, ambas con el mismo ratio de coincidencia con el patrón asentado en la tradición, la resolución de este empate no siempre se hace correctamente. Dado el verso del Ejemplo 10, el sistema devuelve el ritmo 3.6.8.10, que es igual de correcto que el anotado, 2.6.8.10. Sin embargo, cuando el acento en la sexta sílaba recae en un diptongo "cruel", la diéresis "crü-el" es más natural que la dialefa entre "si" y "a" elegida por Jumper.

(10) *si a Silvia la cruel pastora viere*

(a) *si-a-**Sil**-via-la-**cruel**-pas-**to**-ra-**viere**,*
11 — 3.6.8.10 (Jumper)

(b) *si͡a-**Sil**-via-la-crü-**el**-pas-**to**-ra-**viere***
11 — 2.6.8.10 (Anotado)
(Hernando de Acuña)

La segunda fuente de error es la anotación manual incorrecta, que se detecta en 26 de los 100 versos analizados. En algunas ocasiones, los errores se deben a desambiguaciones imprecisas del anotador. El Ejemplo 11 se ha anotado con el ritmo 2.3.7.10, el cual es extraño al endecasílabo. La escansión 2.(3).6.10 del endecasílabo heroico puro es la correcta para ese verso.

(11) *dolor pide a Felipe de Liaño*

(a) *do-**lor-pi**-de͡a-Fe-**li**-pe-de-Lï-**a**-ño*
11 — 2.3.6.10 (Jumper)

(b) *do-**lor-pi**-de-a-Fe-**li**-pe-de-**Lia**-ño*
11 — 2.3.7.10 (Anotado)
(Lope de Vega)

La tercera fuente de error (21 equivocaciones), también relacionada con la anotación, se debe a la consideración de palabras átonas como tónicas y viceversa. Por ejemplo, en los primeros 1400 versos se suele considerar átona la interjección "oh", mientras que, en los restantes, se considera tónica. Jumper la considera tónica. Solo debido a la arbitrariedad de la acentuación de esta interjección, tan frecuente en el Siglo de Oro, se detectan en el análisis manual 11 errores.

La cuarta fuente de error son 8 versos con errores ortográficos que provocan asignaciones de acentos incorrectas.

La atribución de los 10 errores restantes es ambigua.

5.3.2 Errores sobre medida mixta

Se ha revisado manualmente una muestra de 100 de los 800 versos clasificados incorrectamente por Jumper.

Se han hallado 21 errores de anotación. Es importante recalcar que el corpus de Carvajal ha sido recientemente introducido para esta tarea y se encuentra todavía en un proceso de refinamiento. De hecho, la herramienta puede contribuir a limpiar la anotación, puesto que si se revisan los errores de Jumper se encuentra que uno de cada cinco casos es atribuible a un error manual de anotación y no a un error del sistema.

Otra fuente de error es que en ocasiones los anotadores no consideran tónicos algunos determinantes indefinidos y demostrativos: es el caso de "un, una, unos, unas, este, esta...". La mayor parte de las veces, estos determinantes se encuentran en acentos extrarrítmicos, de modo que se anulan por el acento rítmico contiguo. Sin embargo, es un acento que realmente se encuentra en el verso, y aunque posteriormente en un análisis abstracto se elimine, consideramos que se ha de tener en cuenta en la escansión. Como Jumper, de acuerdo con RAE (2009, 9.7b), sí los considera acentuados, falla en 46 de los 100 versos analizados por esta razón. Sin embargo, como hemos explicado, podría considerarse también un error de anotación. En nuestras pruebas comprobamos que si se añaden estos determinantes tónicos a la lista de palabra átonas, el acierto sobre este corpus baja del 0,82 al 0,76, de lo que se deduce que los anotadores en algunas ocasiones consideran estos determinantes átonos y en otras ocasiones tónicos.

En menor medida (6 casos), se han hallado errores derivados de no emplear etiquetado gramatical. Por ejemplo, cuando se procesa la palabra "mientras" nuestro sistema la considera átona, ya que su función más frecuente es conjuntiva ("Estudia *mientras* yo leo") (Quilis, 1984, p. 25); sin embargo, cuando tiene función de adverbio, es tónica "estudia; *mientras*, yo leo". Este problema no es una dificultad insalvable con el sistema implementado, ya que se pueden resolver este tipo de errores tratándolos como una ambigüedad más. En cualquier caso, la frecuencia tan baja de errores atribuibles a la falta de etiquetado gramatical confirma que la estrategia de Jumper es preferible.

Del mismo modo que en el corpus de medi-

da fija, el sistema falla 23 veces por no elegir correctamente el mejor candidato entre los versos ambiguos.

Los 4 casos restantes se deben a errores ortográficos que provocaban asignaciones de acentos incorrectas.

6 Interfaz gráfica para Jumper

Se ha desarrollado una aplicación de escritorio para ofrecer una interfaz al método expuesto. Mientras el usuario escribe el poema, se ejecuta el algoritmo y se imprime el análisis en tiempo real. Para cada verso, si no tiene acentos extrarrítmicos se imprime en verde. Si no coincide plenamente con el esquema acentual sin extrarrítmicos, se imprime en negro. Si no cumple la tendencia versal del poema, en rojo. La tendencia versal se calcula de forma automática si el usuario no la introduce explícitamente. Esta aplicación puede ser de utilidad para investigadores de la poesía española y traductores. Puede descargarse en el siguiente enlace: `https://github.com/grmarco/jumper`

7 Conclusiones

En este trabajo se ha presentado un algoritmo de escansión automática de poemas en español sin necesidad de silabación. A partir del conocimiento de la métrica española, se han establecido una serie de premisas que han simplificado el problema. De manera resumida, las premisas son:

- La tarea de silabación es distinta a la de contar sílabas.
- La vocal es núcleo de la sílaba en español; por lo tanto, es lo que identifica su unidad. Así, se ha podido desarrollar un sencillo algoritmo de cómputo de sílabas y acentos (Sección 4.1), que ha mejorado notablemente la eficiencia sin necesidad de uso de librerías externas.
- Ante un verso ambiguo, las realizaciones posibles del patrón métrico son finitas y asentadas en la tradición. Por lo tanto, se trata de resolver las ambigüedades teniendo en cuenta la aproximación a los patrones naturales del verso.
- La compensación y dialefa hemistiquial se han de tener en cuenta para realizar escansiones precisas.

Gracias a estas premisas, se ha desarrollado un sistema análisis métrico automático que simplifica el problema de la medida del verso, tiene en cuenta la compensación hemistiquial y resuelve las ambigüedades métricas derivadas sinalefas, dialefas, sinéresis y diéresis, sin perder precisión ni información sobre la decisión del analizador.

Nuestro algoritmo, Jumper, mejora el actual estado del arte en un 2 % para la clasificación de patrones métricos sobre poesía de medida fija, y en un 25 % sobre poesía de medida mixta. Además, todas las evaluaciones se han ejecutado entre 21 y 25 veces más rápido que el estado del arte.

También se ha llevado a cabo un análisis de los errores cometidos por el sistema, lo que permite vertebrar trabajos futuros para solucionarlos, y también depurar los corpus utilizados, ya que entre 1/5 y 1/4 de los supuestos errores del sistema son en realidad problemas de la anotación manual.

Finalmente, se ha desarrollado una interfaz gráfica para el algoritmo de análisis métrico en tiempo real, que puede ser de utilidad para investigadores de poesía española.

Agradecimientos

Esta investigación se ha desarrollado gracias al proyecto MISMIS-BIAS (PGC2018-096212-B-C32), financiado por el Gobierno de España, Ministerio de Ciencia, Innovación y Universidades.

Para la evaluación del sistema expuesto han sido de ayuda la facilitación de los corpus empleados por parte de los autores de Rantanplan (de la Rosa et al., 2020).

References

Agirrezabal, M., I. Alegria, and M. Hulden. 2017. A comparison of feature-based and neural scansion of poetry. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, Ranlp 2017*, pages 18–23.

Bello, A. 1859. *Principios de la ortolojía i métrica de la lengua castellana.* La Opinión.

Caparrós, J. D. 1993. *Métrica española.* Síntesis Madrid.

Caparrós, J. D. 2001. *Diccionario de métrica española.* Alianza.

Carvajal, A. 1983. *Extravagante jerarquía: 1968-1981*, volume 58. Hiperión.

de la Rosa, J., Á. Pérez, L. Hernández, S. Ros, and E. González-Blanco. 2020. Rantanplan, fast and accurate syllabification and scansion of spanish poetry. *Procesamiento del Lenguaje Natural*, 65:83–90.

Española, R. A. 2009. *Nueva gramática de la lengua española*, volume 2. Espasa Libros.

Honnibal, M. and I. Montani. 2017. spacy 2: Natural language understanding with bloom embeddings. *Convolutional Neural Networks and Incremental Parsing*, 7.

Navarro-Colorado, B. 2017. A metrical scansion system for fixed-metre spanish poetry. *Digital Scholarship in the Humanities*, 33(1):112–127.

Navarro-Colorado, B., M. R. Lafoz, and N. Sánchez. 2016. Metrical annotation of a large corpus of spanish sonnets: representation, scansion and evaluation. In *International Conference on Language Resources and Evaluation*, pages 4360–4364.

Padró, L. and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *International Conference on Language Resources and Evaluation*.

Pou, P. J. 2020. *Métrica española*. Ediciones Cátedra.

Quilis, A. 1984. *Métrica española*. Ariel Barcelona.

RAE. 2009. *Nueva gramática de la lengua española. Fonética y fonología*, volume 3. Espasa Libros.

RAE, R. A. E. 2010. *Ortografía de la lengua española*. Espasa.

Tomás, T. N. 1956. *Métrica española*, volume 4. Las Americas Publishing Company.

A Apéndice: Código fuente y reproducibilidad

Se pueden reproducir los resultados del analizador en este repositorio: `https://github.com/grmarco/jumper-evaluation`

Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 89-98 recibido 29-11-2020 revisado 21-01-2021 aceptado 31-01-2021

Using Guarani Verbal Morphology on Guarani-Spanish Machine Translation Experiments

Uso de la Morfología Verbal del Guaraní en Experimentos de Traducción Automática Guaraní-Español

Yanina Borges, Florencia Mercant, Luis Chiruzzo
Universidad de la República
Montevideo, Uruguay
{yanina.borges.mijailovich, florencia.mercant, luischir}@fing.edu.uy

Abstract: This paper shows the results of a project for building computational tools and resources for processing the Guarani language, an under-researched language in the NLP community. We developed a baseline machine translation system for the Guarani-Spanish pair, and then performed a series of experiments trying to improve its quality using morphological information. In this work we focus on the analysis of verbs, which is the most complex part of speech in Guarani. We report the results of the different tools implemented for verbs analysis and detection in Guarani, as well as the experiments on machine translation carried on using different versions of the corpus augmented with morphological features.
Keywords: Guarani, Verbal Morphology, Spanish, Machine Translation.

Resumen: Este artículo muestra los resultados de un proyecto para construir herramientas y recursos computacionales para procesar el idioma guaraní, un idioma poco explorado por la comunidad de PLN. Se desarrolló una línea base de traducción automática para el par Guaraní-Español, y luego se realizaron una serie de experimentos para intentar mejorar la calidad de esta línea base utilizando información morfológica. Este trabajo se enfoca en el análisis de los verbos, los cuales componen la categoría gramatical más compleja en el idioma guaraní. Se reportan los resultados de las distintas herramientas implementadas para el análisis y detección de verbos en guaraní, así como los experimentos sobre traducción automática hechos sobre diferentes versiones del corpus aumentado con atributos morfológicos.
Palabras clave: Guaraní, Morfología Verbal, Español, Traducción Automática.

1 Introduction

Guarani is a language spoken by 12 million people in several Latin American countries, mainly in Paraguay, Argentina, Brazil and Bolivia. Paraguay is the country that speaks the Guarani language the most according to a 2002 census[1], and is one of the official languages in the country. In our case we will focus on the Jopara dialect, which is the current variant of Guarani spoken in Paraguay and includes many Spanish loanwords and neologisms (Estigarribia, 2015; Lustig, 2010).

Processing the Guarani language presents particular challenges because it is largely under-researched in the NLP community. There are very few tools or resources built for this language. Guarani is a morphologically rich language, it is agglutinative and polysynthetic (Estigarribia and Pinta, 2017), with words formed by combining prefixes and suffixes around a root, and often the roots or lemmas could be used to form different parts of speech.

In this work we focus on studying the grammar of verbs, which is the part of speech with the greatest complexity in Guarani. We also focus on building tools for the Guarani-Spanish language pair as Spanish is the second most spoken language in Paraguay and is comparatively a much more researched language in the NLP community. The main contributions of this work to the processing of Guarani can be summarized as follows:

- We propose a rule-based method for morphological analysis of verbs in Guarani.

[1] https://www.paho.org/English/DD/AIS/cp_600.htm

ISSN 1989-7553. DOI 10.26342/2021-66-7

- We implement two approaches for detecting verbs in Guarani: a rule-based method and a probabilistic system based on Hidden Markov Models.
- We create a baseline Guarani-Spanish machine translation system based on neural networks, and given the corpus we use is very small and consequently the translation quality is poor, we perform several experiments incorporating during training the linguistic knowledge extracted with the aforementioned methods to improve the results.

2 Related work

There are comparatively very few works that focus on applying NLP techniques to the Guarani language, and in particular few works on creating corpora and resources for this language. The corpus COREGUAPA (Secretaría de Políticas Lingüísticas del Paraguay, 2019) is a monolingual reference corpus of current Guarani, there is also a small corpus of Mbya Guarani sentences (which is a very different dialect from Jopara) annotated using the Universal Dependencies framework (Thomas, 2019; Dooley, 2006), and there have been attempts at creating bilingual Guarani-Spanish or Guarani-English corpora from Wikipedia, although the Guarani version of Wikipedia is itself very small.

There have been some attempts at creating machine translation or translation support systems that focus on the Guarani-Spanish pair, for example: (Gasser, 2018) describes a system for computer aided translation between Guarani and Spanish which uses morpho-syntactic rules to find translation candidates; similarly (Rudnick et al., 2014) presents a web system for collaborative translation between Guarani and Spanish with the aim of creating a parallel corpus; in (Abdelali et al., 2006) they describe a project for developing resources for a Guarani-English corpus using Spanish as a bridge language, and more recently (Alcaraz and Alcaraz, 2020) describes a web tool for analyzing Guarani sentences using a rule-based grammar approach and translating between Guarani and Spanish using an example-based method. However, none of these systems present an evaluation of results, and the systems that are readily available only perform simple translations at word level.

We differ from these works in that we collected a small corpus of parallel Guarani-Spanish documents from the web and tried to apply linguistically motivated techniques to improve the performance of a neural machine translation system over this corpus. As far as we know, this work is the first one to present a neural machine translation baseline for the Guarani-Spanish pair, and to try to improve it using morphological features. The use of morphological features to aid in machine translation for morphologically rich languages has been tried in the past for languages such as Turkish, Russian, Kazakh and Arabic (Bisazza and Federico, 2009; Myrzakhmetov and Makazhanov, 2016; El-Kahlout et al., 2019), being a helpful technique for low resourced languages.

3 Analyzing Guarani verbs

This section presents an introduction to Guarani verbal morphology and describes the tools we developed to approach the automated processing of Guarani verbs.

3.1 Introduction to Guarani verbal morphology

As mentioned before, Guarani is an agglutinative and polysynthetic language: its words con be generated by combining many different types of prefix and suffix morphemes around a root, and these morphemes generally act as independent units, i.e. they have their own meaning and do not change when they are attached to words (Academia de la Lengua Guaraní (ALG), 2018). For example, the inflected verb *"aguata"*, meaning "I walk", can be analyzed as the prefix *"a"* and the lemma *"guata"* which corresponds to the verb "walk". The prefix indicates the person who is performing the action, in this case the singular first person.

Verbs in Guarani can be classified into two groups based on the root word or lemma: proper verbs and verbalized lexical categories. Proper verbs have verbal roots, while verbalized lexical categories use a noun, adjective or adverb as root. The previous example *"aguata"* is a proper verb as it has the verbal root *"guata"* (to walk), while the verb *"amitã"* (I am a child) uses as root the noun *"mitã"* (child), so it is a verbalized noun.

The root of a verb gives meaning to the word, and other morphemes indicate the different verbal inflections. Guarani uses five

main grammar categories for inflecting verbs, indicated by different verbal affixes: **number and person** (it works as a single prefix), **form**, **voice**, **mood** and **tense**.

Number and person: Guarani uses two grammatical numbers (singular and plural) and three grammatical persons (first, second and third), while the first person plural is further subdivided in two types whether the action of the verb includes the interlocutor or not (inclusive and exclusive). There are a total of 43 prefixes that are used to denote number and person.

Form: The verb form indicates if the action of the verb is affirmative, negative or a question. Interrogative forms do not need a particular symbol to be a question, for example, to ask "Do you walk?" we would say *"reguatapa"*, with the root *"guata"* (walk), the second person plural prefix *"re"*, and the interrogative suffix *"pa"*. There are four suffixes that could be used for indicating a question. Affirmative forms use the base form of the verb and do not add an additional morpheme. Negative forms require both a prefix and a suffix around the root. For example, the phrase "I don't walk" would be *"ndaguatai"*, with the prefixes *"nd"* and *"a"* (for first person singular), then the root *"guata"* and the suffix *"i"*. The prefix *"nd"* with the suffix *"i"* indicate the negative form of the verb. There are eight combinations of prefix and suffix to indicate negative verbal forms.

Voice: Establishes the relationship between the subject and the action of the verb in a sentence, it could be either passive voice or active voice. For example, the verb *"oñembohérakuri"*, meaning "he was named", is in the passive voice indicated by the prefix *"ñe"*. First, we find the prefix *"o"* that indicates the third person singular form, then the prefix *"ñe"*, after this the root *"mbohéra"* which means "name" and finally the suffix *"kuri"* indicates recent past. This verb is in the simple indicative mood because there is no morpheme between the root and the recent past morpheme. There are ten prefixes that could be used to indicate the voice category.

Mood: Indicates the way in which the action is performed. There are two main types: indicative or imperative. For example, the verb *"opu'ãva"*, meaning "he (usually) wakes up", is in the usual indicative mood indicated by the suffix *"va"*. First, we find the prefix *"o"* that indicates the third person singular form, then the root *"pu'ã"* meaning "wake up" and finally the suffix *"va"*. This verb is in the simple indicative mood and simple present because there is no morpheme for these verbal categories. There are 46 possible suffixes for the indicative mood and eight for the imperative mood.

Tense: Indicates the time when the action of the verb is done. The basic tenses are present, past and future, but they have variants. There are 14 suffixes used to denote tense, no suffix is used for indicating present tense.

Verbal affixes always appear in the same order. To the left of the root we have the prefixes of form (only for negation), number and person, and voice. To the right of the root we have the suffixes of mood, form and tense. If the form is negative, the form prefix must match the number and person. Figure 1 shows this order graphically.

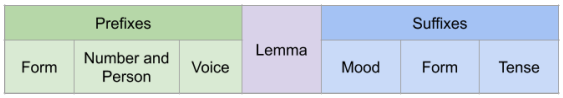

Figure 1: Canonical order of verbal affixes.

3.2 Manual annotation of the corpus

We used a small parallel corpus of around 14,500 sentence pairs extracted from the web. It has around 228,000 tokens in Guarani, corresponding to 336,000 tokens in Spanish, and it consists of articles (news, blog posts and stories) from Paraguayan websites. The corpus was aligned with a semi-automatic process: first aligning it automatically and then manually fixing the incorrect alignments (Chiruzzo et al., 2020). The news articles comprise the majority of the corpus, but they also present more noise in their translations, while the translation quality of the blog posts and stories is better. The corpus was divided into three sets for training, development and test of around 90 %-10 %-10 % keeping the same ratio between news, blog posts and stories for the three subsets.

We manually annotated a fraction of the corpus identifying all verbs and their morphological features. This was done for the blog posts and stories articles as the text was less noisy with a greater number of Guarani words and fewer Spanish loanwords or neolo-

gisms. There are 1,015 verbs in the training set, 412 in the development set, and 477 in the test set.

3.3 Verbal morphology analysis

We use the following rule-based heuristic for performing the morphological analysis of verbs: given a verb, we split it up as a concatenation of valid prefixes and suffixes in the possible order of appearance (see figure 1). We must consider some restrictions imposed by rules such as negation, where the use of a negation prefix restricts the number and person prefixes to be used.

The central part of the verb that is not recognized as a valid prefix or suffix is considered the root of the verb. This root is tested against a dictionary of valid verbs and words extracted from a web source[2].

When analyzing a word, we consider all possible combinations of prefixes, root and suffixes extracted in this way, and we sort the options based on the following priorities:

1. The resulting root or word is found in the dictionary and has a definition associated with a verb.
2. The resulting root is found in the dictionary and it is not a verb, but some grammatical rule of verbal affixes was applied.
3. The resulting root is found in the dictionary but it is not a verb and no grammatical rule of verbal affixes was applied.
4. The resulting root is not found in the dictionary, but some grammatical rule of verbal affixes was applied.
5. The resulting root is not found in the dictionary and no grammatical rule of verbal affixes was applied.

The process takes a verb and returns the analysis that has the highest priority in this list (where 1 is the highest).

We defined the following metrics to evaluate this method:

- Exact accuracy: Strict metric calculated as the number of the analyses that are exactly the same as expected in the gold standard, divided by the total number of verbs. The exact accuracy is defined in equation 1, with n verbs in the corpus, we define y_i with $i \in \{1, \ldots, n\}$ where $y_i = 1$ if the classification of the ith word is correct, $y_i = 0$ otherwise.

$$e\_accuracy = \frac{y_1 + y_2 + ... + y_n}{total\_number\_of\_verbs} \quad (1)$$

- Relaxed accuracy: It is a more relaxed metric that allows to evaluate the accuracy of the tag sequence found. To calculate this measure, we average number of hits in the tag sequence obtained for each word, that is, the number of correct tags divided by the number of tags in the sequence. The length of the tag sequence for each verb is fixed at six, since it corresponds to the five possible verbal affixes plus the lemma. The relaxed accuracy is defined in equation 2, given n verbs in the corpus, we define e_{ij} with $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, 6\}$, where $e_{ij} = 1$ if the jth tag of the ith verb was correctly classified, otherwise it is $e_{ij} = 0$.

$$r\_accuracy = \frac{e_{11}+\ldots+e_{16}+\ldots+e_{n1}+\ldots+e_{n6}}{6 \cdot total\_number\_of\_verbs} \quad (2)$$

We considered two variants of these metrics: one of them (original lemma) is considering all the possible lemmas, and the other one (tagged lemma) is substituting the lemmas for a tag `LEMAESVERBO` (the lemma is a verb) or `LEMANOESVERBO` (the lemma is not a verb) whether the expected lemma of the verb is present as a verb in the dictionary or not. Table 1 shows the values for the metrics over the development and test sets.

Accuracy		Dev	Test
Exact	Original lemma	0.436	0.310
	Tagged lemma	0.386	0.304
Relaxed	Original lemma	0.751	0.615
	Tagged lemma	0.745	0.621

Table 1: Accuracy results for the rule-based method.

As expected, in all cases the relaxed accuracy is greater than the exact accuracy. The exact accuracy for both experiments (original lemma and tagged lemma) is around 0.3 for the test set, which is quite low, so using these

[2] http://descubrircorrientes.com.ar/2012/index.php/diccionario-guarani

Method	Precision		Recall		F1 Score		Accuracy	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Rule-based method	0.574	0.611	0.838	0.716	0.681	0.660	0.874	0.862
HMM - original words	0.975	0.872	0.191	0.069	0.319	0.128	0.871	0.871
HMM - original lemma	0.975	0.891	0.191	0.083	0.319	0.153	0.871	0.871
HMM - tagged lemma	0.822	0.859	0.546	0.510	0.656	0.640	0.909	0.909

Table 2: Verb detection results.

methods for morphological analysis of verbs as the only tool would not be appropriate.

However, relaxed accuracy returns values greater than 0.6 in all cases, which indicates that in many cases, although the morphological analysis is not exactly the same as expected, several verbal affixes of the word are correctly labeled, which might indicate the method could be used to detect verbs as we will see in the following section.

Although this method has a lot of room for improvement, we consider it gives us starting point for Guarani verbs morphological analysis.

3.4 Verbs detection

We use two approaches for verb detection, first using the rule-based verb analysis method discussed above and then a machine learning method based on Hidden Markov Models (HMM).

3.4.1 Rule-based method

In order to know if a word is a verb or not, we take all the possible combinations found in the previous analysis and consider the word is a verb if at least an analysis of priority 1 or 2 is found. This means, we consider the word a verb if it can be split as a concatenation of valid prefixes, root and suffixes that honors the affixes combination rules and the root is found in the dictionary (either as a verb or another valid category).

We repeat this process for every word in the sentence, tagging all possible appearances of verbs.

3.4.2 Hidden Markov Models

We cast this problem as a sequence labeling problem, in which we take a sequence of words and we want to output a sequence of labels `VERB` or `NOT-VERB`. Notice that this can be seen as a simplified version of POS-tagging, in which we are only focusing on one part of speech (verbs). Because the size of the annotated corpus we have is rather small, we decided to use Hidden Markov Models, a statistical method that has proven to be very good at solving this kind of problems, and is comparatively less data intensive than more modern methods like Recurrent Neural Networks. We used the libraries `nltk` (Bird and Klein, 2009) and `sklearn` (Pedregosa et al., 2011) for implementing these methods. We trained the following variants:

HMM with original words: The standard HMM experiment was trained using all the original words from the sentences.

HMM with original lemma: In this experiment, instead of the words as they appear in the corpus, they are represented as a concatenation of labels representing their verbal affixes plus a label with the lemma. We used the morphological analysis described in 3.3 applied to each word.

HMM with tagged lemma: Finally, we carried out the same experiment but using a variant of the morphological analysis in which the verbal affixes are concatenated but instead of the original lemma we use the `LEMAESVERBO` or `LEMANOESVERBO` labels that represents whether or not the resulting lemma is a verb according to the dictionary. This representation might help reduce the data sparsity problem.

These last two experiments could be considered a hybrid model between the rule-based method and the Hidden Markov Model.

3.4.3 Results for verbs detection

All these experiments where trained over the training set and adjusted for the development set, then evaluated against the test set. Table 2 shows the Precision, Recall, F1 Score and Accuracy results obtained for the different verb detection methods.

Except for accuracy, the other metrics are calculated considering the positive class, that is, we calculate the performance for detecting verbs, where the majority of tokens in the corpus represent non-verb instances. Figure 2 shows a comparison of the different methods implemented for these metrics.

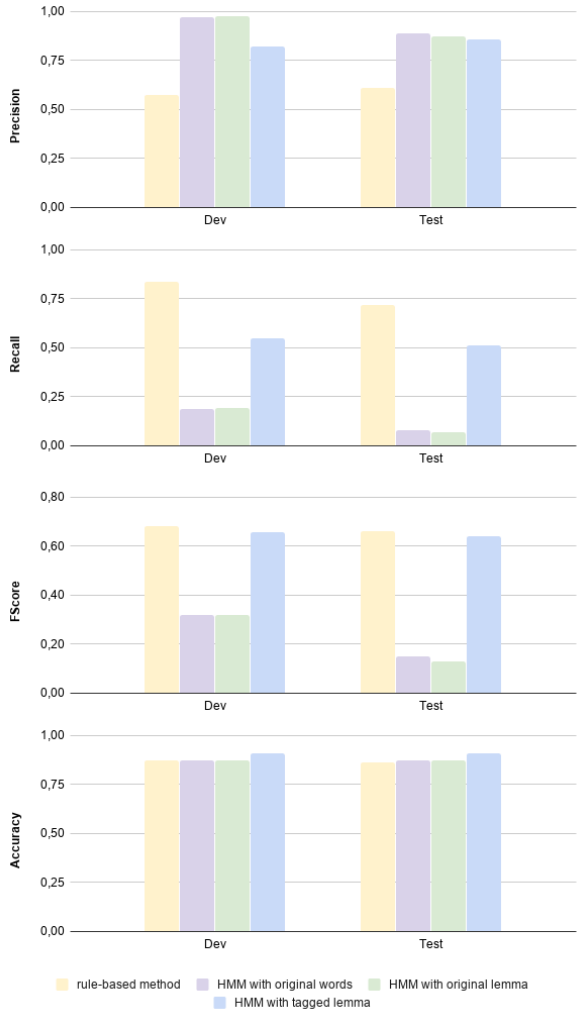

Figure 2: Performance for verbs detection.

From the figure we can see that the methods based on Hidden Markov Model with original lemma and with original words are the ones that return the best precision values. This indicates that these methods tend to get the result right when they classify a word as a verb.

For recall, the results of the rule-based method are much higher than the other methods. This indicates that it tends to classify more instances as verbs, while HMM-based methods tend to classify less (but more precisely). The recall results for the method based on HMM with original lemma and original words are much lower than the others, however with these methods the highest precision was obtained, this indicates that it is a method that tends to be correct when classifies a word as a verb, but given its recall it follows that it tends to classify very few words as verbs.

We consider that the F1 Score measure is the one that best reflects the performance for verbs detection, as it captures the trade-off between precision and recall. The values for the rule-based method and the HMM-based method with tagged lemma are similarly around 0.60. However, for HMM-based methods with original lemma and original words it gives a visibly lower result.

When evaluating the results of the methods with better F1 Score, we can see that with the rule-based method a higher recall is obtained, while with the method based on Hidden Markov Model with a tagged lemma, greater precision is obtained. These differences are balanced generating a similar F1 Score. Both methods seem to be good for verb detection. The difference is that the Hidden Markov Model-based method with tagged lemma has greater precision when classifying a word as a verb, but it tags fewer verbs. On the other hand, the rule-based method has less precision when classifying a word as a verb but classifies more instances as verbs.

In terms of accuracy, the values are very similar for all the methods. For both sets, the best result is the HMM-based method with tagged lemma. However, this measure is skewed by the high imbalance of the classes: there are many more instances of non-verbs than of verbs. For this reason, we consider this measure is not as significant for this study. The result of evaluating these metrics leads to the conclusion that the best methods for classifying verbs in the Guarani language, given the existing corpus, are the rule-based method, and the hybrid method based on Hidden Markov Model with labeled lemmas. Further research is needed to see if we can find a way of complementing both approaches in order to improve the results.

4 Translation experiments

We first performed an experiment to get a machine translation baseline for the Guarani-Spanish pair. We trained a sequence to sequence neural translation model composed by a RNN encoder with attention mechanism and a RNN decoder implemented in the OpenNMT (Klein et al., 2017) library. On our first experiments we used the default parameters for the OpenNMT model. This process ran for 100,000 iterations and took about three days to complete. We ran two variants of this experiment: the first one using all the

words as they appear in the original corpus in Guarani, the second one transforming all the verbs in the Guarani corpus to a representation based on its morphological features. For the verbs detection and analysis we used the heuristics defined in sections 3.3 and 3.4.

For example, the verb *"omoheñóiva'ekue"* ("he/she/they produced") is transformed into the tag sequence *"TPRETPLUSCUAMPERFECTO INDEFFORM 3SINPLU VACTSIMPLE MINDSIMPLE moheñói"* (meaning the verb *"moñehói"* conjugated in the pluperfect tense, indefinite form, third person plural or singular, simple active voice, simple indicative mood).

We evaluated the results of the machine translation experiments using the BLEU (Papineni et al., 2002) measure, which compares the quality between a candidate translation and one or more gold reference translations based on n-grams similarity, trying to capture at the same time notions of fluency and fidelity of the translation.

Figure 3 shows the BLEU results over the development corpus during the 100,000 iterations of these baseline experiments (with original and tagged verbs).

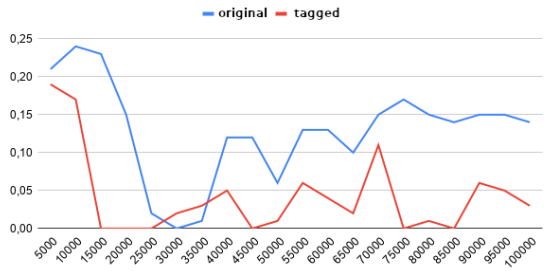

Figure 3: BLEU results for the baseline OpenNMT experiments during 100,000 iterations over the development set.

The best results seem to be achieved in the first iterations (less than 15,000 iterations) with BLEU values up to 0.237 for the original version and around 0.170 for the tagged version. After iteration 20,000 the values seem to decline and start to oscillate, but never achieving higher results. As can be seen in figure 3 after iteration 20,000 the values obtained are lower, and the model with original text performed almost always better than the tagged version.

We manually analyzed some of the predictions made by the model, and found out that they present a very low level of fluency and fidelity, which might be consistent with the low BLEU values. Table 3 shows a translation example using the model trained with original text. As we can see, the translation candidate is not semantically correct (low fidelity), and it is not fluent either since it repeats the same phrase over and over again, which seems to be a common situation in this type of neural architecture (Holtzman et al., 2019).

Original	
gn	*He'i pe reunión Gobierno representante-kuéra ndive omoîva reunión pyahu, pero nodefiníri mba'eve hikuái.*
Gold translation	
es	*Dijo que en la reunión representantes del Gobierno propusieron una nueva reunión, pero no se definió fecha para ello.*
en	*He said that at the Government representatives meeting they proposed a new meeting, but no date was set for it.*
Candidate translation	
es	*Dijo que la COMPRA en la COMPRA de la COMPRA de la COMPRA de la COMPRA de la COMPRA*
en	*He said the BUY in the BUY the BUY the BUY the BUY the BUY*

Table 3: Example of translation achieved with baseline experiment.

Since the results of these first experiments indicated that the best models were generated in the first iterations, we decided to reduce the number of iterations to 20,000 and add more validation checkpoints. Also, for the baseline experiments the morphological information did not seem to improve the translation quality, so we tried to perform different variations in the input data in order to test if these features could be used during training in an advantageous way. For this second round of experiments, we used the following variations of representation for verbs:

1. **Original:** This means using the original text as in the baseline experiment, without morphological information. In this case the verb *"omoheñóiva'ekue"* would not be changed.
2. **Separation in verbal affixes:** Each verb was converted into consecutive labels describing its affixes, and the resulting lemma was added at the end, similarly to what was done in the tagged baseline experiment. In this case, the word *"omoheñóiva'ekue"* would be transformed into the sequence *"TPRETPLUSCUAMPERFECTO INDEFFORM*

3SINPLU VACTSIMPLE MINDSIMPLE moheñoi".

3. **Separation in verbal affixes without default labels:** The third experiment is very similar to the previous one but the labels of verbal inflections that do not add affixes to the lemma (the default values) are excluded. We also decided to remove the plural/singular third person label it is the most used label by default. For example, the word *"omoheñóiva'ekue"*, was transformed into the sequence *"TPRETPLUSCUAMPERFECTO moheñói"*.

4. **Tagged with original lemma:** The fourth experiment consists in using as representation of each verb a concatenation of its verbal affixes and the lemma joined with the "++" symbol. For example, the word *"omoheñóiva'ekue"* becomes the token *"TPRETPLUSCUAMPERFECTO++INDEFFORM ++3SINPLU++VACTSIMPLE++MINDSIMPLE ++moheñoi"*. Notice that in this case it is a single token instead of a sequence of tags.

5. **Tagged with tag lemma:** The fifth experiment is similar to the previous one with the difference that instead of concatenating the verb lemma at the end, we use the tag indicating if the lemma is a verb or not in the dictionary. For example, the word *"omoheñóiva'ekue"* was transformed into the token *"TPRETPLUSCUAMPERFECTO++INDEFFORM ++3SINPLU++VACTSIMPLE++MINDSIMPLE ++LEMAESVERBO"* for this experiment. With this representation, there could be different words that correspond to the same representation, so the rationale behind this is that it could alleviate the data sparsity.

Notice that experiments 2, 3, 4 and 5 could be considered hybrid models leveraging the rule-based methods and the neural machine translation method.

Table 4 shows the BLEU scores achieved for the different experiments over the development corpus.

Figure 4 shows a comparison of the results for these experiments over the development corpus. Notice that the best scores for all models are achieved around iteration 12,000.

Iter	Experiment				
	1	2	3	4	5
2000	0.11	0.20	0.18	0.23	0.20
4000	0.21	0.15	0.20	0.24	0.22
6000	0.20	0.19	0.26	0.21	0.24
8000	0.22	0.20	0.23	0.20	0.24
10000	0.22	0.23	0.24	0.24	0.23
12000	0.24	0.20	0.26	0.24	0.28
14000	0.15	0.11	0.25	0.16	0.26
16000	0.00	0.15	0.25	0.02	0.25
18000	0.08	0.05	0.23	0.05	0.06
20000	0.09	0.00	0.09	0.14	0.15

Table 4: BLEU results for the second round of experiments over the development corpus.

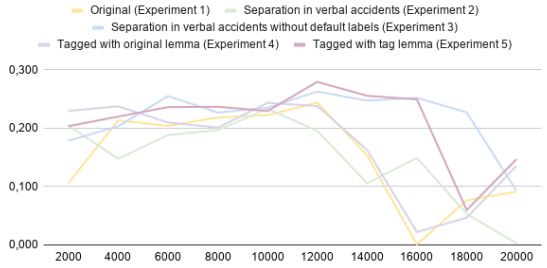

Figure 4: BLEU results over the development corpus.

The experiments with the best results were 3 and 5, reaching a BLEU score of 0.263 and 0.279 respectively for the 12,000 iteration. The experiment with the worst results in most cases is 2. On the other hand, experiments 1 and 4 behave in a similar way.

There seemed to be a certain improvement in BLEU scores using the model from experiment 5 with respect to the original model (experiment 1) over the development set. So we ran all models over the test set to check if this improvement still held on new data. Table 5 shows the results for the second round of experiments over the test set.

Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5
0.174	0.138	0.174	0.157	0.203

Table 5: BLEU results for iteration over the test corpus.

Although the BLEU scores are lower in all cases, we can see that the results keep the same general behavior, that is, experiment 5 achieves the highest result, then experiments 3 and 1, then 4 and finally 2. This seems to indicate that incorporating morphological information during training might help improve

Original	
gn	*Temimbo'e o'yta haguã ojehechauka Ayolas-pe*
Gold translation	
es	*Estudiantes de natación realizan exhibición en Ayolas*
en	*Swimming students perform an exhibition in Ayolas*
Candidate translation	
es	*Estudiantes verifican circuitos turísticos de Ayolas*
en	*Students verify tourist circuits in Ayolas*

Table 6: Example of translation achieved with experiment 5.

the translation quality.

We also performed a manual revision of some results for experiment 5, which is the one with the best results. In this case, the translations in general were more fluent and there were no longer so many repetitions. Table 6 shows a translation example achieved with the result of experiment 5. The translation candidate is clearly more fluent in the target language, and it manages to pick up some of the concepts from the original sentences. However, its semantic is very different. Most of the analyzed predictions still have errors regarding their fidelity, although in some cases they transmit similar messages.

Since no BLEU value exceeded 0.3 in our experiments, we conclude that with the existing corpus, this method for automatic translation requires other complementary methods or tools to obtain better translations. We consider that, even with the modest improvements provided by training with morphological features, there is still a long way until good translation quality can be achieved.

5 Conclusions

We presented the results of an ongoing research on developing tools for processing the Guarani language and the Guarani-Spanish pair. First we implemented a rule-based approach to Guarani verb morphology. This approach showed promising results, although it could further be improved in order to use it in larger scale experiments. For example, we left out of the scope of this project the analysis of some verbal inflections, like the ones that indicate degree. Furthermore, it would be interesting to extend this approach to other parts of speech such as nouns or adjectives, that can also be richly analyzed for Guarani.

We proposed two approaches for verbs detection: a rule-based approach and some HMM-based models. We found out that using a hybrid between the rule-base and HMM methods seems to be amongst the best performing models for detecting verbs in Guarani sentences.

Finally, we tried several experiments incorporating morphological information in order to improve the performance of a baseline neural machine translation system for the Guarani-Spanish pair. The best performing methods use some of the morphological features along with part of the original text. However, the translation results in general are still far from being perfect, so further research is needed in order to improve these systems. The corpus size was one of the major limitations, so we would like to expand the corpus and also improve the alignment quality of some pairs. Nonetheless, we consider that the augmentation of the training data using morphological features served as a proof of concept for the kind of improvements that could be done in this machine translation system without using more data.

References

Abdelali, A., J. Cowie, S. Helmreich, W. Jin, M. P. Milagros, B. Ogden, H. M. Rad, and R. Zacharski. 2006. Guarani: a case study in resource development for quick ramp-up mt. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, "Visions for the Future of Machine Translation*, pages 1–9.

Academia de la Lengua Guaraní (ALG). 2018. *Gramática Guaraní*.

Alcaraz, N. A. and P. A. Alcaraz. 2020. Aplicación web de análisis y traducción automática guaraní–español/español–guaraní. *Revista Científica de la UCSA*, 7(2):41–69.

Bird, Steven, E. L. and E. Klein. 2009. Natural language processing with python.

Bisazza, A. and M. Federico. 2009. Morphological pre-processing for turkish to english statistical machine translation. In *nnnn*.

Chiruzzo, L., P. Amarilla, A. Ríos, and G. Giménez Lugo. 2020. Development of a Guarani - Spanish Parallel Corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages

2629–2633, Marseille, France, 05. European Language Resources Association.

Dooley, R. A. 2006. Léxico guarani, dialeto mbyá com informações úteis para o ensino médio, a aprendizagem e a pesquisa lingüística. *Cuiabá, MT: Sociedade Internacional de Lingüística*, 143:206.

El-Kahlout, I. D., E. Bektaş, N. Ş. Erdem, and H. Kaya. 2019. Translating between morphologically rich languages: An arabic-to-turkish machine translation system. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 158–166.

Estigarribia, B. 2015. Guarani-spanish jopara mixing in a paraguayan novel: Does it reflect a third language, a language variety, or true codeswitching? *Journal of Language Contact*, 8(2):183–222.

Estigarribia, B. and J. Pinta. 2017. *Guarani linguistics in the 21st century*. Brill.

Gasser, M. 2018. Mainumby: un ayudante para la traducción castellano-guaraní. *arXiv preprint arXiv:1810.08603*.

Holtzman, A., J. Buys, L. Du, M. Forbes, and Y. Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Klein, G., Y. Kim, Y. Deng, J. Senellart, and A. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July. Association for Computational Linguistics.

Lustig, W. 2010. Mba'éichapa oiko la guarani? guaraní y jopara en el paraguay. *PAPIA-Revista Brasileira de Estudos do Contato Linguístico*, 4(2):19–43.

Myrzakhmetov, B. and A. Makazhanov. 2016. Initial experiments on russian to kazakh smt.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. volume 12, pages 2825–2830.

Rudnick, A., T. Skidmore, A. Samaniego, and M. Gasser. 2014. Guampa: a toolkit for collaborative translation. In *LREC*, pages 1659–1663.

Secretaría de Políticas Lingüísticas del Paraguay. 2019. Corpus de Referencia del Guaraní Paraguayo Actual – COREGUAPA. http://www.spl.gov.py. Accessed: 2019-11-01.

Thomas, G. 2019. Universal dependencies for mbyá guaraní. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 70–77.

Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 99-109 recibido 2-12-2020 revisado 22-01-2021 aceptado 31-01-2021

A morphological analyser for K'iche'

Un analizador morfológico para el idioma k'iche'

Ivy Richardson,[1] **Francis M. Tyers,**[12]
[1] Department of Linguistics, Indiana University, Bloomington, IN
[2] National Research University Higher School of Economics, Moscow
{ivrichar, ftyers}@iu.edu

Abstract: This paper describes the development of a free/open-source computational morphological description for K'iche', a Mayan language spoken in Guatemala. The language is of the agglutinative morphological type, with both prefixing and suffixing morphology. Both the nominal and verbal morphology are moderately complex. K'iche' is under-resourced and this is the first publication describing a computational tool for the language, and one of the first publications describing a computational tool for any language of the Mayan group. We use the Helsinki Finite-State Toolkit (HFST) for implementing the finite-state transducer. An automatic evaluation of the coverage of our implementation shows that the coverage is adequate, between 86% and 96% on range of freely available corpora. A manual evaluation gives a recall of over 90% over a hand-annotated test set. Both the analyser and the hand-annotated test set are available under a free/open-source licence.
Keywords: k'iche', morphological analysis, finite-state.

Resumen: Este artículo describe el desarrollo de un modelo computacional de la morfología quiché. La lengua quiché es una lengua maya que se habla en Guatemala. Es un idioma del tipo aglutinante con morfología de prefijos y sufijos. Tanto la morfología verbal como la morfología nominal son complejos a un nivel moderado. El quiché es una lengua de pocos recursos computacionales y esta publicación es la primera que describe una herramienta computacional para el idioma, y alguna de las primeras para cualquier lengua del grupo maya. La herramienta está desarrollada con HFST, una caja instrumentos para implantar transductores de estados finitos. Una evaluación indica que la cobertura de vocabulario está adecuada, entre 86% y 96% calculado sobre diversos corpus libres. Una evaluación manual indica una sensibilidad por 90% sobre un conjunto de pruebas anotadas a mano. Tanto el analizador como el conjunto de pruebas están disponibles bajo una licencia de software libre.
Palabras clave: quiché, análisis morfológico, transductores de estados finitos.

1 Introduction

3 This paper describes a morphological analyser for K'iche', a Mayan language spoken in southwestern Guatemala. Though K'iche' is the most widely spoken indigenous language in Guatemala, with nearly one million native speakers as of 2002 (INE, 2018) it is still categorised as a threatened language by the UNESCO *Atlas of the World's Languages in Danger* (Moseley, 2010).

K'iche' is a language with highly inflectional and derivational verbal morphology, which makes morphological analysers vital for both further computational research and the creation of tools such as spell-checkers for K'iche' speakers. Morphological analysers can both generate and analyse words based on a set of morphological and morphographemic rules and a list of lexemes for a language. Our morphological analyser is based on finite-state technology, which is able to map between surface forms, e.g. *nutinamit* 'my town' and lexical forms, e.g. `<px3sg>tinamit<n>` 'SG3-town-N'.

For the creation of the morphological analyser, we chose to base our analyser on the Helsinki Finite-State Toolkit (Lindén et al., 2011) due to its support for weighted finite-state transducers and the `twol` formalism (Koskenniemi, 1983). We took a freely available K'iche'–English dictionary, converted it into a machine readable format, and then converted the words into HFST-compatible lexemes. We then input morphophonemic rules from existing K'iche' grammars and

ISSN 1989-7553. DOI 10.26342/2021-66-8

teaching resources.

Most resources covering K'iche' are meant to be used as pedagogical tools rather than as sources for linguistic research, so they frequently did not cover in-depth the morphological and phonological rules of the language. In creating the analyser, we found that many aspects of the verbal morphology were not addressed in sufficient detail. In order to fill this gap, we present a diagramme of K'iche' verbal morphotactics.

The remainder of the paper is laid out as follows: section 2 describes the grammar of K'iche', section 3 describes prior computational work on K'iche' and other Mayan languages, section 4 describes the methodology of completing the analyser, section 5 provides an evaluation of the analyser, and looks qualitatively at the remaining issues. Sections 6 and 7 describe some future directions and offer some concluding remarks.

2 K'iche'

K'iche' is a language within the Quichean-Mamean branch of the Mayan language family. As of the 2018 Guatemalan census, it is documented to have over 1.5 million native speakers, however the number is likely higher now and does not account for speakers in the diaspora. There are roughly 23 dialects of K'iche' spoken throughout southwestern Guatemala (cf. Figure 1). Our work is based primarily on the Christenson dictionary (Christenson, 2006), which is based on the West dialect spoken in Totonicapan and Momostenango, and the Ixcoy grammar (Ixchajchal Batz, Cumez, and López Ixcoy, 1996), which is based on the Central dialect spoken in Santa Cruz del Quiche.

K'iche', like other Mayan languages, follows ergative-absolutive alignment. The subject and object of a given sentence are marked within the verb using what are called 'set A' markers, for the ergative, and 'set B' markers, for the absolutive. Set A markers indicate the subject in transitive verbs, as well as possessors for nouns. Set B markers indicate the subject for intransitive verbs and the object for transitive verbs. In addition, both set A and set B markers have null morpheme when referring to a formal second-person, and the verbal form is followed by a formality marker (Ixchajchal Batz, Cumez, and López Ixcoy, 1996). Table 1 gives the forms of the two sets of markers.

Verbs in K'iche' are inflected for aspect, subject, object, and voice. Verbal inflection consists of both prefixing and suffixing, although most inflectional verbal morphemes are prefixes. Finite verb forms may also contain infixes for incorporated movement. These morphemes indicate the direction of an action, for example towards or away from the speaker.

K'iche' follows a Verb-Object-Subject word order (Ixchajchal Batz, Cumez, and López Ixcoy, 1996). Nouns are not inflected for case, so K'iche' relies on a fixed word order to indicate noun function. Instead of a copula that inflects, K'iche' places a set B marker and a certain particle in place of a verb. K'iche' has a complex set of voices, including a passive, an instrumental, and various forms of antipassive. These voices are depicted through a complex system of verbal inflection (cf. Figure 2).

Most nouns that refer to human beings (as well as some nouns that refer to animals) inflect for plurality, while all inanimate nouns do not. All nominals inflect for possession (Ixchajchal Batz, Cumez, and López Ixcoy, 1996). K'iche' contains a class of nouns called relational nouns, which are used to introduce purpose clauses, show causation, and form the comparatives of adjectives, among other functions (Can Pixabaj, 2017). Relational nouns either carry Set A markers, which index the complement, or prepositions, or both.

2.1 Orthography

There is a recognised standard orthography for K'iche',[1] developed by the *Academia de las Lenguas Mayas de Guatemala* (AMLG), and most texts we have used are written in this orthography. However, the precise orthographical form, or 'spelling' of individual word forms can still vary greatly among resources and dialects, particularly with regards to vowels and the glottal stop. Some dialects distinguish between tense and lax vowels, with tense vowels lacking diacritics and lax vowels being marked with diaeresis, i.e. tense *a* vs. lax *ä*. The distinction is represented in the ALMG orthography, however many K'iche' speaking communities lack the distinction. Most of the resources we used were not written in dialects that distinguish

[1]The standard is defined in the *Acuerdo Gubernativo Número 1046-87* of the 23rd November 1987.

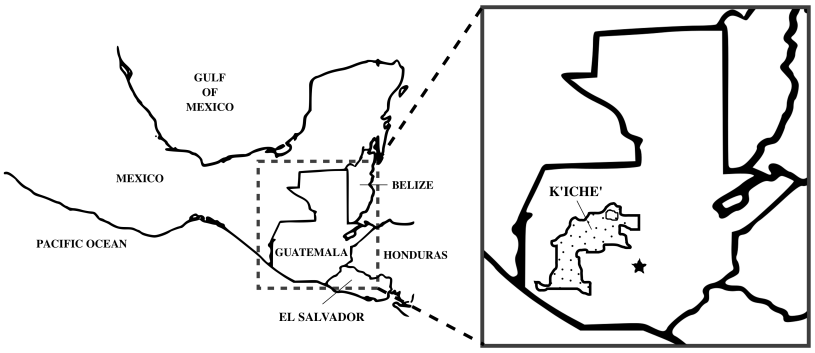

Figure 1: Dotted area represents approximate extent of the K'iche' speaking area in Guatemala.

	Singular			**Plural**		
	1	2	3	1	2	3
Set A	*nu-*, *inw-*	*a-*, *aw-*	*u-*, *r-*	*qa-*, *q-*	*i-*, *iw-*	*ki-*, *k-*
Set B	*in-*	*at-*	*ø-*	*oj-*	*ix-*	*e-*

Table 1: The Set A (Ergative) and Set B (Absolutive) person and number agreement markers for K'iche'. Set A markers are used on nouns to indicate possession and on verbs to indicate a transitive subject, and Set B markers are used on nouns for predication and on verbs for transitive object or intransitive subject. The third person singular Set B marker is null. The Set A markers have phonological variants before consonants (on the left) and vowels (on the right).

between tense and lax vowels, however the dictionary by Christenson (2006) does. We retain the diaereses where they are available in the original resource, but in order to deal with the variation, we implemented a spell relaxer module along with the analyser to accept input both with and without diaereses. The spell relaxer is implemented as a set of finite-state optional replacement rules. These are composed with the surface side of the transducer to produce the final analyser.

Another difference is between short and long vowels. Some works, e.g. Can Pixabaj (2017), indicate this distinction orthographically, by writing short vowels a single time and long vowels twice, like 'i' or 'ii' for /i/ and /iː/ respectively, but we found that most works do not make an orthographic distinction regarding vowel length.

The character for glottal stop /ʔ/ and for the ejective series of consonants is widely written using a number of punctuation symbols, i.e. ' U+0027 *Apostrophe*, ' U-2019 *Right Single Quotation Mark*, ´ U+00B4 *Acute Accent* and ‘ U+2018 *Left Single Quotation Mark*. We standardise on using the Unicode symbol ʼ U+02BC *Modifier Letter Apostrophe* and using the same spell relaxer module to accept input using any of the symbols.

3 Prior work

There is very little prior computational linguistic or natural-language processing work on K'iche', or any of the Mayan languages. Here we describe some of the research we have found. For K'iche', a limited analyser of verbs, containing 408 verb stems, was implemented as part of the Morfo project[2] (Gasser, 2009; Gasser, 2011), but was unpublished and has been unmaintained for over ten years (Gasser, p.c.). Kuhn and Mateo-Toledo (2004) describe some preliminary work on developing natural language processing tools for help in language documentation of Q'anjob'al, another Mayan language of Guatemala. They describe creating a basic finite-state grammar for the language, and train a maximum-entropy part-of-speech tagger on 4,100 tokens of manually annotated data. Their tagger gets an accuracy of 80% when doing 10-fold cross validation. Furthermore they describe some initial experiments in creating a language-model-based spellchecker. A prototype machine translation system from Spanish to Tseltal, a Mayan language of Chiapas in Mexico is

[2] https://github.com/hltdi/morfo

described in Morales Mancilla et al. (2011). The system takes a pipeline-based approach first analysing Spanish text lexically, looking up the Tseltal translations in a bilingual dictionary and then using a context-free grammar to generate Tseltal from Spanish.

4 Methodology

4.1 Lexicon

The lexicon was constructed both semi-automatically, using the dictionary by Chrinstenson (2006) and manually, based on a frequency list. When adding words manually we referred to two other dictionaries, the *Diccionario K'iche'–Español* (Conferencia Episcopal de Guatemala, 2011) and *K'iche' Choltzij* (Academia de Lenguas Mayas de Guatemala, 2004). It contains around 6,000 entries (see Table 2) categorised by part of speech and morphological paradigm.

4.2 Morphotactics

4.2.1 Nominals

Nominals in K'iche' may inflect for possession. To indicate possession, a Set A marker is added as a prefix to the possessed noun e.g. *nutinamit* 'my town' from *tinamit* 'town'. Some nouns also contain differences between their possessed and non-possessed forms. For example, the unpossessed word *kik'* 'blood' gets a suffix when possessed, for example *nukik'el* 'my blood' (Romero et al., 2018).

Relational nouns have functions similar to prepositions and some pronouns (e.g. object and direct object) in Spanish (Romero et al., 2018), but take possessive Set A markers just like nouns. Relational nouns must be possessed. Relational nouns can be combined with prepositions to form adpositional phrases. Phrases with the preposition *chi* 'to' and the relational noun *-ech*, which can have multiple definitions, are contracted to form a single word.

4.2.2 Verbs

As was previously mentioned, K'iche' verbs display a complex morphology. They inflect for the person and number of the subject (and object, in the case of transitive verbs), tense, and aspect. Additionally, they may contain infixes for incorporated movement.

There are three basic types of verb stems in K'iche': intransitive, root transitive, and derived transitive. In addition, there are morphemes of movement which can act either as intransitive verbs or infixes for incorporated movement alongside a verb stem. There are also positional stems, which function similarly to verbal stems.

Conjugated intransitive verbs contain a tense/aspect/mood prefix (hereon referred to as a TAM marker) (Can Pixabaj, 2017), a Set B marker for the subject, the intransitive stem, and a status suffix, depending on the verb's location within the phrase. Occasionally, a finite intransitive form will contain a movement morpheme between the Set B marker and the stem. Intransitive verbs cannot have passive/antipassive forms.

Root transitive verbs follow a consonant-vowel-consonant phonological structure e.g. *b'an* 'to do' (Ixchajchal Batz, Cumez, and López Ixcoy, 1996). In their most basic active conjugated forms, they contain a TAM marker, a Set B marker for the object, a Set A marker for the subject, the verb stem, and an optional phrase-final suffix. Like intransitive verbs, transitive verbs can contain a movement morpheme, although this goes between the Set A and Set B markers when both are present.

The other type of transitive verbs, derived transitive verbs, function like root transitive verbs in many cases. However, derived transitive verbs are typically longer that root transitive verbs and have infinitive forms that end in *-j* e.g *ch'ab'ej* 'to talk to'. The verb stem can be derived from the infinitive form by removing the *-j*. The verb stem for *ch'ab'ej* would be *ch'ab'e*. Basic active forms of derived transitive verbs are similar to root transitive active forms, but they add a *-j* after the stem and do not take phrase-final suffixes. Like root transitive verbs, derived transitives can contain a movement morpheme between the Set A and Set B markers (Ixchajchal Batz, Cumez, and López Ixcoy, 1996).

The valency of transitive verbs in K'iche' can be reduced in the case of passive and antipassive forms. In these cases, the subject and object respectively may still be expressed with the use of an adpositional phrase.

As was previously mentioned, verb forms may contain incorporated movement (cf. Figure 2). In the imperative mood, the TAM marker differs between verb forms with incorporated movement and verb forms without incorporated movement.

Other voices in K'iche', such as passive,

Word class	Subclasses	Entries	Word class	Subclasses	Entries
Nouns	14	2,642	Numerals	2	42
Verbs	3	1,777	Conjunctions	2	21
Adjectives	6	701	Pronouns	3	19
Proper nouns	5	365	Directionals	–	14
Prepositions	–	102	Determiners	–	4
Adverbs	2	155	Other	–	142

Table 2: The lexicon split by word class, there are a total of 5,984 lexical entries in the file, including contractions.

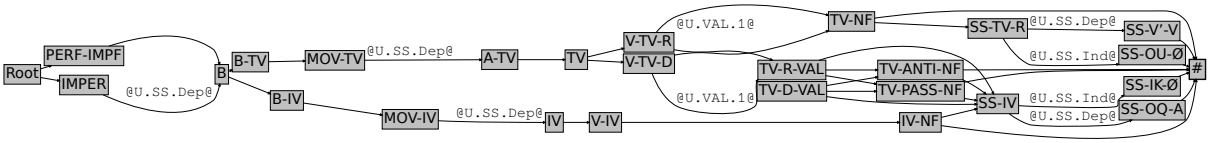

Figure 2: A graph of continuation classes modelling K'iche' verbal inflection. The node labels are the names of the continuation lexica, for example `PERF-IMPF` for the first level aspectual prefixes, *k-* and *x-*, `IMPER` for the imperative prefixes, *ch-*, and *j-*, and `TV` for transitive verb stems. The arc labels are flag diacritics which control non-adjacent morphotactic constraints. For example, in the `IMPER` and `PERF-IMPF` lexica a status suffix variable is set, either dependent, `Dep` or independent, `Ind` and this variable is used to choose the correct status suffix at status suffix lexica word finally. The graph has been lightly simplified for presentation reasons.

antipassive, and imperative, are indicated by differing TAM markers. The passive voice is used to express situations where the agent of an action is unknown/irrelevant (Can Pixabaj, 2017). In passive voice, the subject is omitted and the object is expressed with a Set B marker. The antipassive voice is used to express situations where the recipient of an action is unknown/irrelevant. Similarly to the passive voice, when forming an antipassive verb form, the object is omitted and the subject is expressed with a Set B marker (Ixchajchal Batz, Cumez, and López Ixcoy, 1996). Both antipassive and passive verb forms can contain inherent movement morphemes, but the movement always refers to the semantic agent (Romero et al., 2018).

Participles can be formed from any verb type, although the derivation process is slightly different for each verb type. For intransitive verbs, the participle is derived by adding *(i)naq* to the end of the stem. For both types of transitive verbs, the participle is formed by adding either *um* or *om* to the end of the stem, depending on the root vowel (Romero et al., 2018).

4.2.3 Other categories

There are two prepositions in K'iche', *chi* (approximately 'to') and *pa* (approximately 'in'). These are often contracted with relational nouns to form complex adpositions. For example, *chirij nutinamit* 'about my town' (lit. *chi-* 'to', *-rij* 'its-back', *nutinamit* 'my-town'). K'iche' adverbs do not inflect. Adverbs can be used to introduce purpose, temporal, reason, manner, and conditional clauses (Ixchajchal Batz, Cumez, and López Ixcoy, 1996).

4.3 Morphophonology

We used morphographemic rules in the *twol* formalism to model the phonological alternations. This formalism was first proposed by Koskenniemi (1983) and consists of finite-state constraints over possible input–output string pairs. These constraints are applied in parallel via the composition operator and the output of each of the constraints is intersected. There were a total of 10 rules, covering the preconsonantal and prevocalic forms of the agreement markers, and vowel harmony processes in suffixes. Figure 3 gives two rule examples.

4.4 Example output

Figure 4 gives the output of the system for a sentence from the Universal Declaration of Human Rights. The sentence has been analysed and tokenised by the analyser. Each token starts with a circumflex `^`. This is followed by the surface form and a forward slash

```
"Vowel harmony in root transitive status suffix"
%{U%}:Vx <=> Vy [ ? - Vow ]* %>: _ ;
      where Vx in ( u o )
            Vy in ( u o )
      matched ;
```

```
"Third-person singular possessive alternation"
%{r%}:r <=> _ %>: Vow ;
```

Figure 3: Two phonological constraints: The first deals with the vowel harmony in the status suffix applying to root transitive verbs. The second deals with the form of the third person possessive, which is *u-* preceding consonants and *r-* preceding vowels. Archiphonemes are encoded with `{…}` and `%>` is the symbol for morpheme boundary.

```
^Maj/maj<adv>$
^jun/jun/<det>/jun<adj>/jun<n>/jun<num>$
^winaq/winaq<n>$
^ya'tal/ya'tal<adj>$
^ta/taj<neg>$
^chech/chi<pr>+<px3sg>ech<n><rel>$
^xaq/xaq<adj>/xaq<adv>/xaq<n>/xäq<n>$
^k'ate'/k'ate'<pr>$
^kachapatajik/<impf><s_sg3>chap<v><tv><pass><stat>+ik<mark>$
^,/,<cm>$
^xuq/xuq<adv>$
^kokisax/<impf><s_sg3>okisaj<v><tv><pass>$
^pa/pa<pr>$
^che'/che'<adj>/che'<n>$
^we/we¹<cnjsub>/we²<cnjsub>/we<det>$
^maj/maj<adv>$
^umak/<px3sg>mak<n>$
^ub'anom/<s_sg3>b'an<v><tv><pp>$
^;/;<sent>$
^xuqe/xuqe<adv>$
^kelesaxik/<impf><s_sg3>elesaj<v><tv><pass>+ik<mark>/<s_pl3>elesaj<v><tv><pass><inf>$
^,/,<cm>$
^xuq/xuq<adv>$
^koqatax/<impf><s_sg3>oqataj<v><tv><pass>$
^b'i/b'i<adv><dir>/b'i<n>$
^chupam/chi<pr>+<px3sg>pam<n><rel>/chup<v><iv><inf>$
^pa/pa<pr>$
^ri/ri<det>/ri<cnjsub>$
^utinamit/<px3sg>tinamit<n>$
^./.<sent>$
```

Figure 4: Example output from the analyser for Article 9 of the Universal Declaration of Human Rights, *Maj jun winaq ya'tal ta chech xaq k'ate' kachapatajik, xuq kokisax pa che' we maj umak ub'anom; xuqe kelesaxik, xuq koqatax b'i chupam pa ri utinamit.* 'No one shall be subjected to arbitrary arrest, detention or exile'.

`/`. This is then followed by sequence of analyses delimited by forward slashes. The token ends with the dollar sign `$`. The analysis is composed of a lemma and a sequence of morphological tags which are surrounded by `<` and `>` symbols. The tagset used is based on that of the Apertium project (Forcada et al., 2011). A single token may be split using the `+` symbol, as in the case of contractions, e.g. *chech* is split into *chi* 'to' and a form of the relational noun *-ech* 'to, for'. Tokens are delimited with `^` and `$`, tags are encapsulated by `<` and `>` and contractions are split using the `+` symbol. The tags used are given in Table 3

5 Evaluation

We have evaluated the analyser in three ways. First we calculate the *naïve* coverage over a range of corpora to determine how many tokens receive at least one analysis.[3] Then we manually annotate a subset of 100 tokens and

[3]We consider this naïve as a token is counted as *covered* if it receives a single analysis, however it may not receive all potential analyses and the analysis it receives may not be correct.

Tag	Description	Tag	Description
<adj>	Adjective	<num>	Numeral
<adv>	Adverb	<pass>	Passive
<cm>	Comma	<pass><stat>	Stative passive
<cnjsub>	Subordinating conjunction	<pp>	Perfect participle
<det>	Determiner	<pr>	Preposition
<dir>	Directional	<px3sg>	Possession, 3rd pers. sing.
<impf>	Imperfective	<rel>	Relational
<inf>	Infinitive	<sent>	Sentence marker
<iv>	Intransitive	<s_pl3>	Subject agreement, 3rd pers. plur.
<mark>	Marker	<s_sg3>	Subject agreement, 3rd pers. sing.
<n>	Noun	<tv>	Transitive
<neg>	Negative	<v>	Verb

Table 3: The list of tags used in the analysis in Figure 4 with their descriptions. This is a subset of the full tagset.

calculate the precision and recall. Finally we analyse a randomly selected sample of tokens which do not receive any analysis and categorise the errors.

5.1 Corpora

The analyser was developed principally using the K'iche' translation of the New Testament, *Ru Loq' Pixab' Ri Dios* (Wycliffe Bible Translators, 2011). This was chosen as it was both the largest single text and also fairly orthographically and dialectally consistent. For this reason coverage of the Bible is likely to be better than texts found 'in the wild'. To account for this we also calculated coverage over two texts which we did not use in developing the analyser. The first was the K'iche' translation of the Law on Access to Public Information of Guatemala, *Q'atojtzij rajilib'al 57-2008* (Gobierno de Guatemala, 2008) and the second was a collection of traditional stories, *Tzijob'elil K'aslemal* (Tol Ciprián et al., 2016).

5.2 Naïve coverage

Our first method of evaluation was to calculate the naïve coverage and mean ambiguity on freely available corpora. Naïve coverage refers to the percentage of surface forms in a given corpora that receive at least one morphological analysis. Note that forms counted by this measure may have other analyses which are not delivered by the transducer. The mean ambiguity measure was calculated as the average number of analyses returned per token in the corpus. The results can be found in Table 4.

5.3 Precision and recall

In order to test the precision and recall of the analyser we used a test corpus created from sentences from the *Chqeta'maj le qach'ab'al K'iche'!* course (Romero et al., 2018). We first copied all the example sentences and analysed them using our transducer. We then went through and added missing analyses and removed invalid analyses according to the translations and glosses. This gave us a disambiguated corpus of 337 sentences where each of the 2,021 tokens was assigned the appropriate analysis in context.

To calculate precision and recall, for each of the tokens in the corpus we collected the valid analyses and made a gold standard where each token was associated with a list of valid analyses.

We define true positives, tp, as those analyses which were in both the transducer's output and in the gold standard list of analyses. We define false positives, fp, as those analyses that were in the transducer output but not in the gold standard list of analyses. And we define false negatives, fn as those analyses which were in the gold standard list, but not in the transducer output. Tokens which received no analyses were counted as false positives. We defined precision, P (1), recall, R (2) and F_1-score (3).

$$P = \frac{\text{tp}}{(\text{tp} + \text{fp})} \tag{1}$$

$$R = \frac{\text{tp}}{(\text{tp} + \text{fn})} \tag{2}$$

Corpus	Genre	Tokens	Coverage	Average ambiguity
Ru Loq' Pixab' Ri Dios	Religion	206,827	95.49	1.55
Q'atojtzij rajilib'al 57-2008	Legal	18,853	90.69	1.89
Tzijob'elil K'aslemal	Folklore	5,477	86.89	1.49

Table 4: The naïve coverage of the analyser over a range of texts and text types.

	Precision	Recall	F_1-score
Tokens	76.53	98.22	86.03
Types	67.98	93.17	78.61

Table 5: Precision, recall and F_1-score for the test set. The metrics are substantially higher for tokens as more frequent tokens appear more frequently in the evaluation corpus and exhibit more of the valid analyses.

$$F_1 = 2\frac{PR}{P+R} \tag{3}$$

Intuitively, precision is the likelihood of an analysis presented by the transducer being an analysis found in the gold standard, while recall is the likelihood of an analysis found in the gold standard being in the transducer. Table 5 presents the results of the evaluation.

Note that this method is only an approximation of the precision and recall of the analyser as the corpus may not contain all valid analyses for a given token. For example, the corpus has several mentions of the word *juyub'* as a noun 'mountain', but the lexicon also contains an entry as an adjective meaning 'steep'. Thus completeness in the lexicon will be penalised by the precision metric.

A more thorough evaluation would be to ask a native speaker to supply all and only the valid analyses for a given token with the aid of a concordance for each token.

We inspected the list of false negatives and found that there were some errors which were repeated. For example, in the gold standard the form of the first person singular set B pronoun, *-in-* was assimilated with a following nasal to *-im-* as in the verb form *kimb'e* 'I go', instead of the form *kinb'e*. This assimilation was not found in any of the other corpora we used and accounted for over a quarter of all false negatives. Another phonological difference between our gold standard and the other corpora we used was the form of the antipassive *-Ow-* after the verb *-to'-* 'help'. In the test corpus the suffix vowel was deleted leaving only the *-w-* of the suffix. There were also a number of idiosyncratic forms of the verb *-aj-* 'want', and some forms of other verbs which did not follow the regular patterns. A more thorough study of phonological variation would allow us to resolve these errors.

5.4 Unanalysed forms

In addition to the previous methods, we have also done an evaluation of forms (types) that do not receive any analysis, sorting them into five categories: missing stem, morphotactic error, morphophonological error, orthographic variation and tokenisation error. These forms were selected pseudo-randomly[4] from a concatenation of all of the evaluation corpora.

As can be seen from Table 6, missing stems make up the bulk of the errors. The coverage of the available corpora is impressive, given the small size of the lexicon, but there is a lot of lexicographic work to continue with. The largest number of missing stems was found in the categories of verbs and nouns.

In terms of morphotactic error, we count incorrect paradigm assignment, missing or incorrectly formed morphemes, mistakes in the way the continuation lexica are linked together, or mistakes in use of flag diacritics. For example, the wordform *ech'oko'ib'* 'cripples' was not analysed because the word *ch'oko'* 'cripple' was not assigned to the paradigm of words that have a plural form in *-ib'*.

We consider as orthographic variation any word that is equivalent to one already in our lexicon but with a different orthographic form. This does not imply any judgement as to normativity of the form, and items counted in this category could be anything from dialectal variation to typographical errors. For example, the word *rajilib'al* 'A3SG-date' appears in our lexicon as the entry, [**ajilabal** *(n)* date (calendar); number] — with the

[4] Using the GNU *sort* utility.

Error category	Frequency	Percentage (%)
Missing stem	65	61.9
- Verb	34	32.3
- Noun	18	17.1
- Proper noun	7	6.6
- Adjective	4	3.8
- Other	2	1.9
Orthographic variation	24	22.8
Morphotactic error	10	9.5
Morphophonological error	3	2.8
Tokenisation error	3	2.8
Total:	105	100

Table 6: Proportion of errors by category. Note that although there were only 100 words selected, the number of errors adds up to more than 100 as some words evinced more than one kind of error. For example if a word was both written in a way not found in our lexicon and in generation of the form from our lexicon there was a phonological error, we counted it in both categories.

vowel *a* in place of *i* before the *-b'al* (instrumental, locative) derivational suffix. In another instance we found the wordform *jastasq*, where our lexicon contains [**jastaq** *(n)* things; goods].

6 Future work

There are a number of avenues for future work. First of all we intend to fix all of the errors that we found during the evaluation. Secondly, the lexicon certainly needs to be expanded, both in terms of lexemes and dialect coverage, and improved for consistency in labelling forms as to the dialect they pertain to. We have included lexical information from a number of different dialects and although some of the entries are marked, it was not possible to mark all of them.

There are certain lacunae in terms of non-finite verb forms, it is not clear to us how the various infinitives should be categorised.

Given the fairly high ambiguity exhibited we would like to work on disambiguation for K'iche'. We have started work on manually disambiguating texts and would like to use the analyser as groundwork for bootstrapping a treebank for K'iche' under the Universal Dependencies project (Nivre et al., 2020).

The analyser can also be used to generate training data for machine learning applications, morphological analysis using data generated from finite-state machines to train neural networks has already been used in e.g. (Silfverberg and Tyers, 2019).

In terms of applications, we forsee that this work could be used in developing spellchecking and predictive text software that supports K'iche' as well as providing the basis of further language technology.

7 Concluding remarks

We have presented the first morphological analyser for K'iche', a Mayan language principally spoken in Guatemala. The analyser comprises a finite-state transducer based on the Helsinki Finite-State Tools. It covers a reasonably high percentage — 90–96% of tokens in running text over a number of freely available corpora of K'iche'. The analyser is available as free/open-source software under the GNU General Public Licence.[5]

Acknowledgements

We would like to express our thanks to Allen Christenson for the use of his K'iche'–English lexicon as the initial lexical base of the analyser. We would also like to thank Robert Henderson and Pedro Mateo Pedro for fruitful discussions about the analyser, and the anonymous reviewers for their helpful suggestions. This article is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University).

[5] https://github.com/apertium/apertium-quc/

References

Academia de Lenguas Mayas de Guatemala. 2004. *K'iche' Choltzij*. ALMG. 2nd Edition.

Can Pixabaj, T. A. 2017. K'iche'. In J. Aissen, N. C. England, and R. Zavala Maldonado, editors, *The Mayan Languages*. Routledge, Oxford.

Christenson, A. 2006. K'iche'–English dictionary. `http://www.famsi.org/mayawriting/dictionary/christenson/index.html`.

Conferencia Episcopal de Guatemala. 2011. *Diccionario K'iche'–Español*.

Forcada, M. L., M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Gasser, M. 2009. Semitic morphological analysis and generation using finite state transducers with feature structures. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 309–317, Athens, Greece, March. Association for Computational Linguistics.

Gasser, M. 2011. Computational morphology and the teaching of indigenous languages. In S. Coronel-Molina and J. McDowell, editors, *Proceedings of the First Symposium on Teaching Indigenous Languages of Latin America*, pages 52–63, Indiana University, Bloomington.

Gobierno de Guatemala. 2008. Q'atojtzij rajilib'al 57-2008: Q'atb'altzij re ukujik che uya'ik ub'ixik uwach tinamit. *[Decreto Número 57-2008: Ley de acceso a la información pública]*.

INE. 2018. XII Censo Nacional de Población y VII de Vivienda. `http://redatam.censopoblacion.gt/bingtm/RpWebEngine.exe/Portal?BASE=CPVGT2018`.

Ixchajchal Batz, E. A., L. M. Cumez, and C. D. López Ixcoy. 1996. *Gramática del idioma k'iche'*. Proyecto Lingüístico Francisco Marroquín, Guatemala.

Koskenniemi, K. 1983. *Two-level Morphology: A General Computational Model for Word-Form Recognition and Production*. Ph.D. thesis, Helsingin yliopisto.

Kuhn, J. and B. Mateo-Toledo. 2004. Applying computational linguistic techniques in a documentary project for Q'anjob'al (Mayan, Guatemala). In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisboa, Portugal.

Lindén, K., E. Axelson, S. Hardwick, T. Pirinen, and M. Silfverberg. 2011. HFST—framework for compiling and applying morphologies. *Communications in Computer and Information Science*, 100:67–85, 08.

Morales Mancilla, J. A., H. Guerra Crespo, G. B. Nango Solís, I. Valles López, and A. G. Cossio Martínez. 2011. Traductor del lenguaje español a la lengua tseltal. *Revista Tecnología Digital*, 1(1):27–39.

Moseley, C. 2010. Atlas of the world's languages in danger. `http://www.unesco.org/culture/en/endangeredlanguages/atlas`.

Nivre, J., M.-C. de Marneffe, F. Ginter, J. Hajič, C. D. Manning, S. Pyysalo, S. Schuster, F. Tyers, and D. Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4027–4036.

Romero, S., I. Carvajal, M. Sattler, J. M. Tahay Tzaj, C. Blyth, S. Sweeney, P. Kyle, N. Steinfeld Childre, D. G. Tambriz, L. E. Tambriz, M. Tahay, L. Tahay, G. Tahay, J. Tahay, S. Can, E. I. Xum, E. Guarchaj, S. M. G. Can, C. M. T. Cotiy, T. Can, T. Kingsley, C. Hayes, C. J. Walker, M. A. I. Sohom, J. Sandler, S. G. Ixmatá, M. P. Tahay, and S. Smythe Kung. 2018. Chqeta'maj le qach'ab'al K'iche'! `https://tzij.coerll.utexas.edu/`.

Silfverberg, M. and F. M. Tyers. 2019. Data-driven morphological analysis of nominal morphology for Uralic languages. In *Proceedings of the 5th International Workshop on Computational Linguistics of Uralic Languages*, pages 1–15.

Tol Ciprián, C., D. D. Oxlaj Tistoj, E. Velásquez Vicente, H. Calel Vicente, J. G. Calva Loarca, J. A. Vásquez Ajpop, J. J. Menchú Ordóñez, L. M. Calderón, M. Hernández Pocol, M. M. Batz Socop, O. O. Baten Sarat, R. L. Puac Yac, R. Gómez Pérez, S. C. Mejía Paxtor, S. Gómez Par, T. Castro Gutiérrez, and V. M. Alvarez Poncio. 2016. *Tzijob'elil K'aslemal*. USAID.

Wycliffe Bible Translators. 2011. *Ru Loq' Pixab' Ri Dios*. Wycliffe Bible Translators. `https://ebible.org/Scriptures/details.php?id=qucNNT`.

Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 111-120 recibido 30-11-2020 revisado 24-01-2021 aceptado 04-02-2021

Consumer Cynicism Identification for Spanish Reviews using a Spanish Transformer Model

Identificación del cinismo del consumidor para reseñas en español utilizando un modelo de transformador español

Samuel González-López[1], **Steven Bethard**[2],
Francisca Cecilia Encinas Orozco[3], **Adrián Pastor López-Monroy**[4]
[1]Technological University of Nogales
[2]University of Arizona
[3]University of Sonora
[4]Mathematics Research Center (CIMAT)
sgonzalez@utnogales.edu.mx, bethard@arizona.edu,
cecilia.encinasorozco@unison.mx, pastor.lopez@cimat.mx

Abstract: Companies pay close attention to how consumers react on social media to their products or services. Our work focuses on the identification of Consumer Cynicism, defined as a negative attitude that can have a broad or specific focus and comprises cognitive, affective, and behavioral components. We create a corpus of 619 Spanish-language comments on YouTube car reviews, annotated for four cynicism constructs: Dissatisfaction, Alienation, Skepticism, and Hostility. We compare different classification formulations (binary vs. multi-label) and different pre-trained models (Spanish BETO vs. multilingual BERT). We find binary classifiers derived from BETO consistently outperform multi-label classifiers and classifiers derived from BERT. Our best models achieve F1 of 0.83 for Dissatisfaction, 0.77 for Hostility, 0.71 for Skepticism and 0.70 for Alienation.
Keywords: Consumer Cynicism, binary classification model, multi-label model, social media.

Resumen: Las empresas prestan mucha atención a las reacciones de los consumidores de sus productos o servicios en las redes sociales. Nuestro trabajo se centra en la identificación del cinismo del consumidor, el cual se define como una actitud negativa que puede tener un enfoque amplio o específico y comprende los componentes cognitivo, afectivo y conductual. Creamos un corpus de 619 comentarios en el idioma español sobre reseñas de automóviles de YouTube, los comentarios se etiquetaron para cuatro constructos del cinismo: Insatisfacción, Alienación, Escepticismo y Hostilidad. Además, comparamos diferentes formulaciones de clasificación (binaria vs. multi-etiqueta) y diferentes modelos pre-entrenados (BETO-español vs. BERT-multilingüe). Encontramos que los clasificadores binarios derivados de BETO superan consistentemente a los clasificadores de etiquetas múltiples y a los clasificadores derivados de BERT. Nuestros mejores modelos alcanzan F1 de 0.83 para Insatisfacción, 0.77 para Hostilidad, 0.71 para Escepticismo y 0.70 para Alienación.
Palabras clave: Cinismo del Consumidor, modelo de clasificacion binaria, modelo multi-etiqueta, redes sociales.

1 Introduction

The need to predict customers' behavior has led companies to carry out studies on opinions in digital media. Brands seek to position themselves and provide satisfaction to their potential customers. The global adspend growth for this year is 20% reaching an investment of $84 billion in social media (Santini et al., 2020). An analysis developed from 48 brands in 8 industries (including car brands) showed that volume metrics explain brand awareness and purchase intent. The volume

ISSN 1989-7553. DOI 10.26342/2021-66-9

Components	Related Constructs
Cognitive	Suspicion Mistrust Skepticism Distrust
Affective	Alienation Dissatisfaction
Behavioral	Resistance Hostility

Table 1: Components of consumer cynicism.

metrics correspond to a collection of the number of likes, comments, and shares of posts on Facebook and YouGov (Kübler, Colicev, and Pauwels, 2020).

Many opinions users write on different social media platforms can be classified as negative, neutral, or positive. Works related to sentiment analysis in comments have received significant interest (Kauffmann et al., 2019). In some studies deep learning techniques have been used to classify reviews (Tammina and Annareddy, 2020; Kocoń, Miłkowski, and Zaśko-Zielińska, 2019). However, at a deeper level of analysis, we can find other behaviors that can differ from each other despite being negative or positive.

Our research seeks to identify behaviors related to consumer cynicism. Cynicism is defined as a negative attitude that can have a broad or specific focus and comprises cognitive, affective, and behavioral components (Chylinski and Chu, 2010). Table 1 shows the components and constructs of consumer cynicism defined by prior work (Chylinski and Chu, 2010). Our work focuses on analyzing the following subset of those constructs:

- Skepticism: Doubt of consumer related of the product or brand.
- Alienation: Consumer feels disillusioned, powerless, hopeless, detached from the product or brand.
- Hostility: Consumer attempts to force alternative or desired features on the product.
- Dissatisfaction: Unmet expectations by the consumer, negative perception of the product or brand.

We explore different configurations of the Bidirectional Encoder Representations from Transformers (BERT; (Devlin et al., 2019)) models to classify car reviews on YouTube in the Spanish language. Our contributions are the following:

- We annotate a corpus of YouTube car reviews in Spanish considering the four selected cynicism constructs.
- We train BERT models for Skepticism, Alienation, Dissatisfaction, and Hostility, reaching 0.71, 0.70, 0.83, and 0.77 F1, respectively.
- We demonstrate that binary classification models outperform multi-label models for these problems.

2 Related work

Work closely related to the study of cynicism has analyzed behaviors such as offensive language, sarcasm, irony, and aggression.

Within offensive language, we find the analysis of profanity, insults, and abuse, and explicit or implicit offensive language in German-language tweets using BERT (Risch et al., 2019). This work reports an F1 of 51.2 % for three subcategories of offensive language and 73.1 % for explicit/implicit language. Techniques such as NaiveBayes and Support Vector Machines have also been explored, obtaining accuracy results of 92 % and 90 % (De Souza and Da Costa-Abreu, 2020), but the analysis only sought to identify the offensive language category. A comparison between the Perspective tool (a Convolutional Neural network CNN) and BERT found that on the SEMEVAL2019-Offenseval dataset, the Perspective tool had better performance identifying the offensive language category, while BERT had better performance identifying the insult, threat, and attack offensive language elements (Nikolov and Radivchev, 2019). Similarly, in SEMEVAL2020-Offenseval, the use of BERT + CNN in the analysis of offensive language in social networks found results above those of traditional techniques (Safaya, Abdullatif, and Yuret, 2020).

Recent work has also analyzed the concept of sarcasm. In an analysis of 21 papers on sarcasm, 22.58 % of the cases used Support Vector Machine (SVM) as an analysis technique, followed by 19.35 with Logistic Regression, 9.67 % Naive Bayes (Sarsam et al., 2020). Recently other works have explored BERT models to identify sarcasm, first extracting local features of words in sentences

and later implementing a CNN to summarize all sentences (Srivastava et al., 2020).

Closely related to Sarcasm, we find Irony analyzed under approaches such as CNN with Embeddings (FastText, Word2vec) (Ghanem et al., 2020). This work analyzed monolingual and multilingual architectures in three languages, obtaining better performance from the monolingual configuration. Another approach, RCNN-RoBERTa, consists of a RoBERTa pre-trained transformer followed by a bidirectional Long short-term memory (BiLSTM), reaching 0.80 F1 in the SemEval-2018 dataset and 0.78 F1 in the Reddit Politics dataset (Potamias, Siolas, and Stafylopatis, 2020).

Related work has also attempted the identification of aggression. Aggression can be direct or indirect and is a feeling of anger that results in hostile behavior. Under the BERT framework and an assembly strategy, a dataset labeled as non-aggressive, covertly aggressive, and overtly aggressive was classified, and the assemblies achieved two percentage points higher F1-score than single models (Risch and Krestel, 2020). Using the same dataset but with other training features, for example, the amount of abusive/aggressive/offensive words or the presence of hash-tags, these features were incorporated into a CNN to obtain an accuracy of 73.2 % (Kumar et al., 2020).

Though we take inspiration from this prior work, our work differs in task and language. We focus on cynicism, annotating a new corpus for several previously unexplored constructs. And in contrast to most previous work that focused on the English language, our consumer cynicism analysis is on the Spanish language, and we consider both monolingual Spanish and multilingual configurations of BERT models.

3 Dataset

The training and test data set focus on user comments on YouTube channels with car review topics in the Spanish language. Reviews were pulled from 26 videos from five different car review channels. The review comments were filtered considering two requirements:

1. Comment size: comments should contain at least ten words. The goal here is to ensure sufficient text to judge the presence or absence of the four constructs of the current study. This parameter was set after the annotators performed qualitative analysis on the comments.

2. Relevance: comments should have a minimum of 5 likes. The goal here is to focus on comments that a substantial number of users find interesting.

Category	**Measure**	**Score**
Dissatisfaction	Cohen	0.88
Alienation	Cohen	0.79
Skepticism	Cohen	0.81
Hostility	Cohen	0.82
Overall	Fuzzy	0.79

Table 2: Two-annotators Kappa Agreement.

The resulting comments were tagged by two annotators who had knowledge related to consumer marketing. Each annotator was instructed to annotate each comment with one or more of Dissatisfaction, Alienation, Skepticism, Hostility, or to annotate the comment as None if none of these were present. Each annotator received a guide that defined each construct and provided annotated examples. These guidelines[1] can be found in appendix A. FindingFive[2] was used for the annotation process, and annotators' responses were automatically collected.

The filters reduced the number of comments to be tagged by annotators. For example, one of the videos with the highest number of comments exceeded 1250 but the filters reduced this to 90. After applying the filters to all comments from all 26 videos, 725 comments remained.

Table 2 shows annotator agreement on the 725 comments, measured with Kappa's Cohen (Landis and Koch, 1977) and Fuzzy Kappa (Dou et al., 2007). The fuzzy measure produces a single value for all categories, and allows comments to belong to more than one category. All measures of agreement reached 0.79 or higher, indicating good agreement.

We retained for training and evaluation data the 619 out of 725 filtered comments where annotators agreed on the labels. Table 3 shows the distribution of the categories

[1]The guide is displayed in the original language of the study. Suitable for Spanish annotators.

[2]`https://www.findingfive.com/`

Category	Percentage
Dissatisfaction	16
Alienation	20
Skepticism	18
Hostility	17
None	29

Table 3: Label distribution in the annotated corpus.

in this subset, and Table 4 shows some annotated examples from the corpus.

After the tagged process, we found comments where the annotators had disagreements. For instance, comments were tagged with the category none by the first annotator, while the second annotator selected dissatisfaction or alienation construct. Below are a couple of examples.

1. *Mi Subaru Legacy 1993 1ra Generación es mas espacioso, mas equipado y mas cómodo*/'Mi Subaru Legacy 1993 1ra Generación es mas espacioso, mas equipado y mas cómodo'. Tagged as None/Dissatisfaction.

2. *Lo bueno del march es que es el más económico de todos y en equipamiento está 'completo', digo por el precio que tiene el march*/'The good thing about the march is that it is the cheapest of all and in terms of equipment it is 'complete', I mean for the price of the march'. Tagged as None/Alienation.

These types of examples were not used in the experimentation.

4 Methodology

The collected corpus was used to train machine-learning models. We compare two different formulations of the classification problem: training one binary classification model for each construct of cynicism (Dissatisfaction, Alienation, Skepticism, and Hostility) vs. training a single multi-label model that predicts all constructs jointly. We also compare two different pre-trained transformer models[3] from which cynicism models can be fine-tuned: the multilingual BERT model (Devlin et al., 2019), and the Spanish-language BETO model (Cañete et al., 2020). We also consider adding a convolutional neural network (CNN) layer before the output of the above models. The rest of this section describes these options in detail.

BERT The Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2019) uses a transformer neural network to infer contextual representations of the words in a text. BERT uses a masked language modeling objective to pre-train the transformer network on large unlabeled data: the BooksCorpus (800M words), and English Wikipedia (2,500M words)

Multilingual BERT The BERT training paradigm has been used to train a variety of models that vary in their sizes and their training data. We select for our purposes bert-base-multilingual-cased, a model with 12 self-attention layers, 12 attention heads, 768-dimensional word representations, and 110M parameters. It was trained on cased text in the top 104 languages with the largest Wikipedias.

Spanish BERT (BETO) The BETO model (Cañete et al., 2020) is trained using the BERT training paradigm, and is similar in size to bert-base-multilingual-cased. However, BETO focuses only on the Spanish language. We use BETO-cased that has 12 self-attention layers, 16 attention heads each, 1024-dimensional word representations, and 110M parameters. It was trained on Spanish Wikipedia and the sources of the OPUS Project that had text in Spanish. These sources include the United Nations and Government journals, TED talks, subtitles, and News Stories. According to the authors of BETO, the total of corpora used is comparable to the original BERT. The comparison of BETO vs. BERT multilingual for Spanish gave favorable results to BETO (Cañete et al., 2020).

Convolutional network Though it is possible to make predictions directly from a BERT-style model, we also consider taking the contextual word representations from BERT or BETO and feeding them to a convolutional neural network (CNN) to make the predictions.

Data processing The corpus comments used were delimited by the punctuation mark

[3]In preliminary experiments we also explored bag of words and lemmatization features with naive Bayes, logistic regression, and LSA algorithms, but the BERT-based models always outperformed these.

Spanish Example	English Translation	Labels
se ve buena la suv, poco a poco parecen convencer más que otras marcas por el costo beneficio en cuestión de equipamiento, el único problema sería la confiabilidad que otras marcas como honda dan pero hasta no ver no creer.	'the SUV looks good, but slowly they seem to convince more than other brands for the cost-benefit in equipment issue, the only problem would be the reliability that other brands like honda give but until you see, not believe.'	Skepticism
Es culpa de Nissan y también de la gente, lo siguen comprando por ser la opción fácil y "segura". Mientras los consumidores no exijan mejores productos seguirán existiendo autos como estos.	'It is the fault of Nissan and the people. They continue to buy it for being the easy and safe option. As long as consumers do not demand better products, cars like these will continue to exist.'	Hostility
Nissan es una de las compañías automotriz más grandes en el mundo y aún no pueden introducir una nueva generación en todos los mercados.	'Nissan is one of the world's largest automotive companies, and they cannot yet introduce a new generation in all markets.'	Dissatisfaction Skepticism

Table 4: Label examples from the annotated corpus.

point, which indicated that the comment was finalized. The rest of the symbols and words were kept throughout the entire process. The elements of the test group were randomly selected.

5 Results and Analysis

The collected corpus was divided into three partitions: training, validation, and testing. The first dataset was used to train each of the implemented models, one binary and multilabel. The validation set was used to fine-tune the model hyperparameters. We use the validation set results to perform the analysis for section five. Finally the test set was used to run the best configuration of the BERT model. The hyper-parameters that were adjusted during the evaluation of the validation set were the batch size, the number of training epochs and the learning rate starting with a value of 5e-5.

5.1 Dissatisfaction construct

The results of experimentation for the Dissatisfaction construct are detailed in Table 5. Multi-label models with convolutional layers (rows 3 and 4) outperformed multi-label models without (rows 1 and 2). All binary classification models (rows 5 and 6) outperformed all multi-label models (first 4 rows). And for the better models (last 4 rows), the Spanish BETO models (rows 4 and 6) outperformed their corresponding multilingual BERT models (rows 3 and 5, respectively). The best configuration was Spanish BETO + binary classification, obtaining an F-measure of 0.83.

When manually reviewing the false positives for the dissatisfaction construct, we found local words or words typical to a geographic region that affected the prediction.

> *La Ram esta muy genial pero esa suspensión de aire no va a aguantar la friega...*
>
> 'The Ram is very great, but the air suspension will not withstand hard work...'

or

> *Jeep no está diseñado para chulearlo en la plaza...*
>
> 'Jeep is not designed to be displayed in the square...'

We can see in the first example the *friega* word that means hard work and represents wear for the suspension. In the second example, the *chulearlo* word appears, which means provide positive words about some object or thing.

5.2 Alienation construct

The Alienation construct results in Table 6 showed similar trends as dissatisfaction. Convolutional layers improved the performance of multi-label models, all binary classification models outperform all multi-label models, and for the better models (last 4 rows), Spanish BETO models outperformed multilingual BERT models. The best configuration was

Language	Model	P	R	F
Multi	Multi	0.18	0.52	0.27
Spanish	Multi	0.19	0.51	0.28
Multi	Multi+CNN	0.38	0.48	0.42
Spanish	Multi+CNN	0.63	0.50	0.56
Multi	Binary	0.76	0.63	0.69
Spanish	Binary	0.92	0.78	**0.83**

Table 5: Dissatisfaction construct results in terms of Precision (P), Recall (R), and F. Language is either Multilingual BERT or Spanish BETO. Model is either individual Binary models or a Multi-label model, and may include a convolutional (CNN) layer before the output.

Language	Model	P	R	F
Multi	Multi	0.34	0.51	0.41
Spanish	Multi	0.27	0.45	0.34
Multi	Multi+CNN	0.44	0.53	0.48
Spanish	Multi+CNN	0.45	0.75	0.56
Multi	Binary	0.52	0.72	0.61
Spanish	Binary	0.64	0.76	**0.70**

Table 6: Alienation construct results in terms of Precision (P), Recall (R), and F. Language is either Multilingual BERT or Spanish BETO. Model is either individual Binary models or a Multi-label model, and may include a convolutional (CNN) layer before the output.

again Spanish BETO + binary classification, obtaining an F-measure of 0.70.

Manual analysis of the development set reveals the presence of short expressions of popular wisdom annotated as Alienation. For instance:

A cualquier santo moderno le rezan...
'They pray to any modern saint...'

or

El que no conocio a Dios en el pasado...
'He who did not know God in the past...'

This construct obtained the lowest level of Kappa agreement. Incorporating a method for the detection of theses phrases in the early stages of training could help improve results.

5.3 Hostility construct

Like the preceding results, Table 7 shows that for the hostility construct binary classification outperforms multi-label models and the Spanish BETO models outperform the multilingual BERT model. However, unlike the preceding results, adding a convolutional layer to a multi-label model does not consistently yield an improvement. The best configuration was again Spanish BETO + binary classification, obtaining an F-measure of 0.77.

Language	Model	P	R	F
Multi	Multi	0.35	0.46	0.40
Spanish	Multi	0.52	0.59	0.55
Multi	Multi+CNN	0.42	0.55	0.48
Spanish	Multi+CNN	0.39	0.75	0.51
Multi	Binary	0.68	0.73	0.70
Spanish	Binary	0.76	0.79	**0.77**

Table 7: Hostility construct results in terms of Precision (P), Recall (R), and F. Language is either Multilingual BERT or Spanish BETO. Model is either individual Binary models or a Multi-label model, and may include a convolutional (CNN) layer before the output.

When manually reviewing the data, we found comments with dissatisfaction content that includes negative phrases in a pejorative word game close to sarcasm. For example:

La suspensión trasera la cagaron, mejor una suspensión trasera independiente como las generaciones anteriores. Pero los Mazdetos felices con cualquier cosa
'The rear suspension was shit, better an independent rear suspension like previous generations. But the Mazdetos are happy with anything'

If the experts' labeling of the Hostility construct includes the sarcasm subcategory, then we could identify with some level this type of word game. Negative comments focused on the presenter were not taken into account in this construct. The performance of BETO on this construct was only below the performance of the Dissatisfaction construct. It would appear that BERT models work best with negative comments.

5.4 Skepticism construct

Table 8 shows that, similar to hostility, for the skepticism construct binary classification outperforms multi-label models, the Spanish BETO models outperform the multilingual BERT model, and there is no consistent benefit to adding a convolutional layer to multi-label models. The best configuration was

Language	Model	P	R	F
Multi	Multi	0.26	0.48	0.34
Spanish	Multi	0.32	0.48	0.38
Multi	Multi+CNN	0.31	0.43	0.36
Spanish	Multi+CNN	0.53	0.64	0.58
Multi	Binary	0.56	0.77	0.65
Spanish	Binary	0.64	0.80	**0.71**

Table 8: Skepticism construct results in terms of Precision (P), Recall (R), and F. Language is either Multilingual BERT or Spanish BETO. Model is either individual Binary models or a Multi-label model, and may include a convolutional (CNN) layer before the output.

again Spanish BETO + binary classification, obtaining an F-measure of 0.71

When reviewing the false positives, we found comments with the behaviors described above in the Alienation and Hostility constructs. For instance:

> *Este auto me recuerda a nissan sentra mas bien puro humo.*
>
> 'This car reminds me of a Nissan Sentra, rather pure smoke'

We also found that many comments have a particular negative load, similar to comments of dissatisfaction, for example:

> *Ya sabemos como sera de mediocre y feo el auto con solo ver el emblema de la marca*
>
> 'We already know how mediocre and ugly the car will be just by looking at the brand's emblem'

The dissatisfaction comments specify a dislike for the price or some component of the car, while for the alienation construct, the comments are more global and show a particular dislike of the brand or the car, such as the car's reliability.

6 Discussion

The overall results are summarized in Table 9. Models for all four constructs achieve an F1 of at least 0.70, with the dissatisfaction model achieving the highest F1 of 0.83. Across all constructs, we observed similar patterns: binary classification outperformed multi-label classification and the Spanish BETO model outperformed the multilingual BERT model.

In addition to the metrics reported in the results section, we review the values produced

Category	F-measure
Dissatisfaction	0.83
Hostility	0.77
Skepticism	0.71
Alienation	0.70

Table 9: Constructs F-measure results.

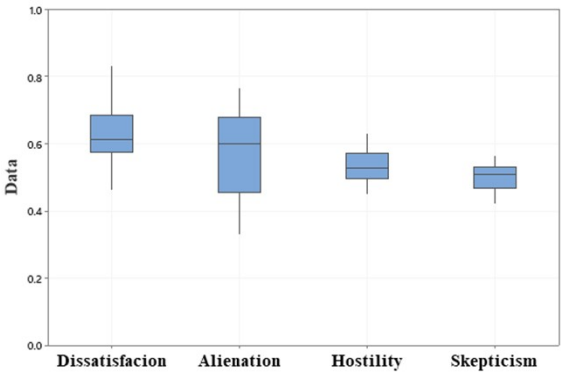

Figure 1: Test groups probabilities.

by the best configuration for all constructs: Spanish BETO + binary classification. For each construct, we gather the models' predictions over all test items and plot the distribution using a box plot. Figure 1 shows the result. An ideal result would be that all models have probabilities far from 0.5, suggesting high confidence in both the presence and absence of a construct. The Alienation model has the largest range of probabilities, while the Skepticism model has the smallest range. In fact, the Skepticism model's probabilities are all very close to 0.5, suggesting that the model is rarely confident whether a comment represents skepticism or not. Levels of agreement on labeling may need to improve. The kappa value for the Dissatisfaction construct was the highest, and in Figure 1, Dissatisfaction obtains the highest probability values.

7 Conclusion

The use of pre-trained models is beneficial since they significantly cover relationships between words due to their enormous amounts of training data. In this study, we have analyzed different configurations of BERT pre-trained models. We found the best performance from binary classification models based on BETO. We also found that for all four constructs we studied, the Spanish language BETO model outperformed the multilingual BERT model.

Of the consumer cynicism constructs, Dissatisfaction obtained the best performance (0.83 F1) and Alienation the lowest (0.70 F1). We believe these F1 scores are encouraging, given the modest amount of data we annotated and the fact that we do not perform any additional pre-processing for our models.

Our analysis suggested that detecting sarcasm and identifying typical terms for geographic regions may be important aspects of future work on cynicism models. It may also be worth exploring hierarchical models that first detect the cynicism component (e.g., Affective) before the cynicism construct (e.g., Alienation). Expanding the corpus, both in terms of size, and in terms of variety of components covered, is also an important direction for future work. Overall, we believe that consumer cynicism is a little-explored concept, which could be of great interest to brands, due to the marked trend of markets towards social media platforms.

References

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Chylinski, M. and A. Chu. 2010. Consumer cynicism: antecedents and consequences. *European Journal of Marketing*, 44(6):796–837.

De Souza, G. A. and M. Da Costa-Abreu. 2020. Automatic offensive language detection from twitter data using machine learning and feature selection of metadata. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6.

Dehghani, M., M. K. Niaki, I. Ramezani, and R. Sali. 2016. Evaluating the influence of youtube advertising for attraction of young customers. *Computers in Human Behavior*, 59:165–172.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dou, W., Y. Ren, Q. Wu, S. Ruan, Y. Chen, D. Bloyet, and J.-M. Constans. 2007. Fuzzy kappa for the agreement measure of fuzzy classifications. *Neurocomputing*, 70(4):726 – 734. Advanced Neurocomputing Theory and Methodology.

Ghanem, B., J. Karoui, F. Benamara, P. Rosso, and V. Moriceau. 2020. Irony detection in a multilingual context. In J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva, and F. Martins, editors, *Advances in Information Retrieval*, pages 141–149, Cham. Springer International Publishing.

Kauffmann, E., J. Peral, D. Gil, A. Ferrández, R. Sellers, and H. Mora. 2019. Managing marketing decision-making with sentiment analysis: An evaluation of the main product features using text data mining. *Journal of Sustainability*, 11:1 – 19.

Kocoń, J., P. Miłkowski, and M. Zaśko-Zielińska. 2019. Multi-level sentiment analysis of PolEmo 2.0: Extended corpus of multi-domain consumer reviews. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 980–991, Hong Kong, China, November. Association for Computational Linguistics.

Kumar, R., A. K. Ojha, S. Malmasi, and M. Zampieri. 2020. Evaluating aggression identification in social media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 1–5, Marseille, France, May. European Language Resources Association (ELRA).

Kübler, R. V., A. Colicev, and K. H. Pauwels. 2020. Social media's impact on the consumer mindset: When to use which sentiment extraction tool? *Journal of Interactive Marketing*, 50:136 – 155.

Landis, J. R. and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Nikolov, A. and V. Radivchev. 2019. Nikolov-radivchev at SemEval-2019 task 6: Offensive tweet classification with BERT and ensembles. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 691–695, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Potamias, R., G. Siolas, and A. Stafylopatis. 2020. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, pages 1433 – 3058.

Risch, J. and R. Krestel. 2020. Bagging BERT models for robust aggression identification. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 55–61, Marseille, France, May. European Language Resources Association (ELRA).

Risch, J., A. Stoll, M. Ziegele, and R. Krestel. 2019. hpidedis at germeval 2019: Offensive language identification using a german bert model. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 405–410, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Safaya, A., M. Abdullatif, and D. Yuret. 2020. Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media.

Santini, F. D. O., W. Ladeira, D. Pinto, M. Herter, C. Sampaio, and B. Babin. 2020. Customer engagement in social media: a framework and meta-analysis. *Journal of the Academy of Marketing Science*, 48(6):1211–1228.

Sarsam, S. M., H. Al-Samarraie, A. I. Alzahrani, and B. Wright. 2020. Sarcasm detection using machine learning algorithms in twitter: A systematic review. *International Journal of Market Research*, 62(5):578–598.

Srivastava, H., V. Varshney, S. Kumari, and S. Srivastava. 2020. A novel hierarchical BERT architecture for sarcasm detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 93–97, Online, July. Association for Computational Linguistics.

Tammina, S. and S. Annareddy. 2020. Sentiment analysis on customer reviews using convolutional neural network. In *2020 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–6.

A Appendix: Annotation Guidelines in Spanish

Guía de anotación para identificar "cinismo del consumidor" en reviews de YouTube.

Introducción: En la actualidad las redes sociales juegan un papel relevante para las empresas que ofrecen productos o servicios. Un estudio realizado a 378 estudiantes de la universidad de Sapienza Roma, se enfocó al análisis de elementos que pudieran contribuir a la intención de compra en videos de YouTube. Los usuarios promediaban de 1 a 7 horas de videos y el 98% tenía experiencia en redes sociales. El análisis reportó que la personalización y el entretenimiento percibido en los videos eran relevantes para la intención de compra (Dehghani et al., 2016). Los videos bajo esta plataforma permiten la interacción con los usuarios, a través de comentarios u opiniones de estos. En específico nuestra investigación busca identificar comportamientos relacionados al cinismo del consumidor, éste se define como una actitud negativa que puede tener un enfoque amplio y especifico, y se conforma de los componentes cognitivo, afectivo y comportamental (Chylinski and Chu, 2010). A continuación, se muestran los tres componentes y sus constructos en la Table 10.

Componentes	**Constructos relacionados**
Cognitivo	Ecepticismo
Afectivo	Alienación Insatisfacción
Comportamental	Hostilidad

Table 10: Componentes del cinismo del consumidor.

De la Table 10 el constructo "Hostilidad" se refiere cuando un cliente/usuario busca resultados alternos o deseados. Este estudio se enfoca en las opiniones y/o comentarios escritos por usuarios en canales de reviews de autos en español. Por ejemplo, en el siguiente comentario de un usuario de YouTube, se puede apreciar el deseo de que se realice una crítica del auto más a fondo: "Hola Gabo espero que cuando la camioneta llegue a México la critiques como debe ser, se entiende que los chinos te invitaron y no se puede decir más...un abrazo de chile".

Instrucciones del etiquetado: El objetivo del etiquetado es identificar si un comentario se relaciona a los siguientes constructos Escepticismo, Alienación, Insatisfacción u Hostilidad. A continuación se muestran ejemplos de cada constructo.

Ejemplo del constructo Escepticismo: Se ve buena la suv, poco a poco parecen convencer más que otras marcas por el costo beneficio en cuestión de equipamiento, el único problema sería la confiabilidad que otras marcas como honda dan pero hasta no ver no creer.

Ejemplo del constructo Alienación: Es culpa de Nissan y también de la gente, lo siguen comprando por ser la opción fácil y "segura". Mientras los consumidores no exijan mejores productos seguirán existiendo autos como estos.

Ejemplo del constructo Insatisfacción: Yo tengo un jetta, pero por 422,900 jajajajajajajaja mejor le completo y me compro un bmw XD esta muy caro el jetta.

Ejemplo de constructo Hostilidad: que equivocado está señor yo tengo una Cadillac y créame que es muy superior a Mercedes y a BMW su motor y el lujo es muy superior y es más grande que sus rivales.

A continuación se muestra un ejemplo etiquetado.

Comentario	Constructo
se ve buena la suv, poco a poco parecen convencer más que otras marcas por el costo beneficio en cuestión de equipamiento, el único problema sería la confiabilidad que otras marcas como honda dan pero hasta no ver no creer.	*Escepticismo* Alienación Insatisfacción Hostilidad
yo tengo un jetta, pero por 422,900 jajajajajajajaja mejor le completo y me compro un bmw XD esta muy caro el jetta.	Escepticismo Alienación *Insatisfacción* Hostilidad

Table 11: Ejemplo anotado.

Como se aprecia en la Table 11, el primer comentario fue etiquetado como Escepticismo, mientras que el segundo comentario se le asignó la etiqueta de Insatisfacción. El proceso de anotación se realizará a través de la plataforma FindingFive. El acceso a la plataforma se enviará por correo electrónico.

Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 121-132 recibido 11-12-2020 revisado 25-01-2021 aceptado 03-02-2021

Grammatical error correction for Spanish health records

Corrección de errores gramaticales en informes clínicos en español

Salvador Lima-López,[1,*] **Naiara Perez,**[2] **Montse Cuadros**[2]
[1]Barcelona Supercomputing Center, Barcelona, Spain
[2]SNLT group at Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
Mikeletegi Pasealekua 57, Donostia/San-Sebastián, 20009, Spain
salvador.limalopez@bsc.es, {nperez, mcuadros}@vicomtech.org

Abstract: This paper describes the first approach to Grammatical Error Correction for Spanish health records. We present a series of experiments using neural networks and data augmentation, achieving 70.89 $F_{0.5}$ score. Resources designed for this task are introduced, namely the IMEC corpus of corrected health records and the TMAE corpus of clinical texts augmented with errors.
Keywords: health records, Grammatical Error Correction, Spanish.

Resumen: Este artículo presenta el primer trabajo sobre la corrección gramatical de textos clínicos en español. En este trabajo, presentamos un conjunto de experimentos basados en redes neuronales y aumentación de datos, en los cuales conseguimos una puntuación de 70,89 $F_{0,5}$. Además, se presentan dos corpus creados para esta tarea: el corpus IMEC, un corpus médico corregido manualmente, y el corpus TMAE, un corpus de textos clínicos aumentado con errores.
Palabras clave: informes clínicos, corrección de errores gramaticales, español.

1 Introduction

Grammatical Error Correction (GEC) is a field within Natural Language Processing that deals with the correction of texts from a grammatical, lexical and orthographic point of view. As a discipline, it has traditionally focused on educational applications such as second-language learners' essays. However, there are other text genres that can also benefit from this sort of treatment. One of them is health records, a type of clinical text. Health records are documents where doctors describe patients' consultations to their office, including their impressions, diagnosis and recommendations.

As the main source of written communication between health professionals and patients, as well as among health specialists themselves, health records should be written in the most correct way possible. Still, due to the heavy time restrictions that health professionals usually work under, health records often present strange grammar structures, abbreviated words and outright spelling errors. Consider the example in Table 1. These documents sometimes end up being the source of misunderstandings on the patients' side (Terroba Reinares, 2015, p. 11).

Original:
EEII: no edemas ni singos de tvp.
Corrected:
EEII: no muestra edemas ni signos de TVP.
Translation:
LE: no signs of edemas nor TVP.

Table 1: A real sentence extracted from a health report verbatim and the proposed correction.

Given that in Spain it is legally required for doctors to write a health record for each consultation (Boletín Oficial del Estado, 2015) and that, according to the latest data available, in 2018 there were over 350 million Primary Health Care and nursing consultations (Ministerio de Sanidad, 2018, p. 11), it is safe to assume that it is not feasible to manually revise and correct health records.

In this paper, we explore for the first time

*Work done while at Vicomtech.

ISSN 1989-7553. DOI 10.26342/2021-66-10

the feasibility of making health records in Spanish clearer and more accessible by applying GEC techniques. We obtain promising results using two new corpora specifically curated for this task. These corpora are, on the one hand, a collection of manually corrected Spanish health records named IMEC (*Informes Médicos en Español Corregidos*, or Corrected Health Records in Spanish) and, on the other, a compilation of various clinical corpora artificially augmented with errors called TMAE (*Textos Médicos Aumentados con Errores* or Clinical Texts Augmented with Errors).

The structure of this paper is the following: §2 briefly discusses the history and current developments of GEC; §3 explains two different corpora for clinical GEC, while §4 introduces the different experiments performed with them and discusses their results. §5 concludes this work by making some final remarks and discussing possibilities of future work.

2 Related Work

Early GEC systems date back to the 1980s, where rule-based pattern recognisers and dictionary-based systems (Macdonald, 1983; Richardson and Braden-Harder, 1988) were initially used. Later on, statistical classifiers were also implemented, focusing on specific error types (Gamon et al., 2008; Tetreault, Foster, and Chodorow, 2010; Lee and Seneff, 2008).

The most successful approach has been to treat GEC as a Machine Translation (MT) task. An analogy can be drawn between both fields, where MT's source language corresponds to GEC's uncorrected text and the target language to corrected text. Statistical MT systems made possible the generation of an N-best list of alternative corrections for each sentence (Shen, Sarkar, and Och, 2004), which can be re-ranked using text features, classifiers or language models. Reranking helps improve overall performance and has become a staple of many state-of-the-art GEC systems even nowadays.

Neural networks have also been proposed due to their generalization potential. For a long time, the most popular architecture has been the Encoder-Decoder model, accompanied either by recurrent neural networks (Xie et al., 2016) or convolutional neural networks (Chollampatt and Ng, 2018). Lately, in the same fashion as many other NLP tasks, the state of the art has been achieved using Transformers (Vaswani et al., 2017).

It could be argued that shared tasks have played an important role in the development of this sub-field. The Conference on Natural Language Learning (CoNLL) held a shared task on GEC both in 2013 (Ng et al., 2013) and 2014 (Ng et al., 2014), releasing a different corpus each year. In 2019, the Building Educational Applications (BEA) shared task was held (Bryant et al., 2019). It saw the release of new annotated datasets (W&I (Yannakoudakis et al., 2018)+LOCNESS (Granger, 1998)), as well as the re-release of previously available corpora (FCE (Yannakoudakis, Briscoe, and Medlock, 2011), LANG-8 (Tajiri, Komachi, and Matsumoto, 2012; Mizumoto et al., 2012) and NUCLE (Dahlmeier, Ng, and Wu, 2013)) in a standardized version. This process was performed using ERRANT (Felice, Bryant, and Briscoe, 2016), a toolkit specifically designed for the annotation of GEC data. At the time of this writing, the state of the art for CoNLL 2014's dataset and W&I+LOCNESS is a Transformer model called GECToR (Omelianchuk et al., 2020) that is trained with sentences augmented with artificial errors and fine-tuned on real data.

Using artificial errors, as GECToR does, is a common technique in GEC known as Artificial Error Generation (AEG). It consists in introducing errors into error-free sentences to create parallel correct/incorrect pairs. While GECTor learns how to make these changes using Machine Learning, rule-based systems can also be used for this task. For instance, Beloki et al. (2020) created a parallel GEC corpus of 500,000 news in Basque using grammatical rules. AEG is a technique that can be performed on any text of any genre and that is very flexible since both the type of errors and how they are introduced (i.e., randomly or probabilistically) can be controlled. This method was studied in-depth, among others, by Felice (2016), Rei et al. (2017) and Grundkiewicz, Junczys-Dowmunt, and Heafield (2019).

Along the same lines, oversampling is a technique that is usually applied to balance unbalanced corpora in classification tasks. It is implemented in some low-resource GEC research due to its seemingly good results

(Náplava and Straka, 2019). In low-resource GEC, however, all sentences are duplicated regardless of whether the errors they include are from a minority class or not, as the objective is not to balance the corpus but simply to expand it. All in all, data augmentation is often a part of GEC due to the sparsity of quality parallel data.

So far, most research on Grammatical Error Correction has focused on English texts. In Spanish, the Corpus Of Written Spanish–L2 and Heritage speakers (COWS-L2H) (Davidson et al., 2020) was recently released and its authors tested its validity by training a GEC system based on an LSTM (Long Short-Term Memory) encoder-decoder. There does not seem to be any other recent GEC papers focused on Spanish. Regarding clinical texts, to the best of our knowledge there are no studies that apply GEC techniques to this domain.

3 Corpora

This section presents the two corpora developed for GEC in the clinical domain: IMEC (*Informes Médicos en Español Corregidos*, or Corrected Health Records in Spanish) and TMAE (*Textos Médicos Aumentados con Errores* or Clinical Texts Augmented with Errors).

3.1 Corrected Health Records in Spanish (IMEC)

IMEC is a collection of sentences from Electronic Health Records presented as parallel correct/incorrect sentence pairs. The sentences have been extracted from NUBes (Lima López et al., 2020), a corpus of Electronic Health Record in Spanish manually labelled with negation and uncertainty phenomena. A sample sentence taken from IMEC is shown in Table 2:

Original:
En Abril de 2003 en escreenning de cancer [...]
Corrected:
En abril de 2003 en un screening de cáncer [...]
Translation:
On April 2003, in a cancer screening [...]

Table 2: A original-corrected sentence pair from IMEC.

The corpus consists of 10,007 sentences, of which 7,801 have at least one correction. The sentences were manually corrected by a single annotator.[1] Correction guidelines were developed based on two different style guides as reference: Bello Gutiérrez (2016) and Aguilar Ruíz (2013). The principles underlying the guidelines are three:

(i) terminological and semantic errors are not be considered, as they should only be corrected by health professionals;

(ii) even if abbreviations are one of the main sources of ambiguity in clinical texts, the only changes made to them is to normalize their spelling; abbreviation disambiguation is in its own a NLP task of great difficulty, particularly in the health domain; and,

(iii) a text's clarity comes from both its content and its structure, which means that our corrections should cover orthotypographic (spelling, punctuation) as well as grammatical aspects.

The parallel corrected/uncorrected sentences have been annotated using the ERror ANnotation Toolkit (ERRANT) (Felice, Bryant, and Briscoe, 2016). ERRANT aligns parallel sentences, extracts edits and categorizes them according to a bi-axial system. The first axis corresponds to the types of changes made to the text:

- Missing: the correction consists in inserting a missing token in the incorrect sentence.
- Unnecessary: the correction consists in deleting a token from the incorrect sentence.
- Replacement: the correction consists in replacing a token with another.

The second axis classifies the errors by the linguistic properties of the token(s) involved. This classification is carried out by ERRANT with rules that involve part-of-speech tagging with the Universal Dependency tagset (Bryant, 2019). Some of the error types include NOUN for noun-related errors, DET for determiner-related errors, SVA for subject-verb agreement errors, and so on.

[1] Even though having a corpus corrected by only one annotator is not ideal, reannotating the corpus in the future is a possibility. For instance, the CoNLL-2014 test set was reannotated multiple times, up to a total of 18 overlapping annotations (Bryant, 2019).

Type	#	%
Replacement	14,414	53.20
Missing	12,505	46.13
Unnecessary	184	0.67
SPELL	12,024	44.36
DET	6,959	25.68
PUNCT	3,829	14.13
PREP	1,254	4.63
VERB	921	3.40
OTHER	787	2.90
ORTH	388	1.43
NOUN	367	1.35
MORPH	81	0.30
AUX	76	0.28
ADJ:INFL	66	0.24
NOUN:INFL	65	0.24
CONJ	57	0.21
DET:INFL	56	0.21
ADJ	42	0.15
PRON	37	0.14
VERB:SVA	25	0.09
VERB:TENSE	24	0.09
VERB:FORM	19	0.07
ADV	15	0.06
SCONJ	7	0.03
WO	4	0.01

(a) IMEC

Type	#	%
Replacement	4,728,619	60.66
Missing	3,021,303	38.76
Unnecessary	44,099	0.56
SPELL	2,387,712	30.64
DET	2,017,786	25.89
PUNCT	1,394,900	17.90
PREP	1,146,550	14.71
OTHER	322,858	4.14
VERB	236,658	3.04
DET:INFL	95,764	1.23
NOUN	50,418	0.65
MORPH	39,468	0.51
ORTH	26,665	0.34
CONJ	23,973	0.31
ADJ:INFL	10,845	0.14
PRON	9,745	0.13
ADJ	9,521	0.12
NOUN:INFL	7,668	0.10
ADV	7,174	0.09
VERB:TENSE	3,437	0.04
AUX	1,420	0.02
SCONJ	933	0.01
WO	352	0.00
VERB:FORM	197	0.00
VERB:SVA	2	0.00

(b) TMAE

Table 3: Edit and error type distribution in the corpora IMEC and TMAE.

More general types also exist, such as SPELL for spelling errors or OTHER for edits that do not fit into any other category.

ERRANT's rules and resources are originally designed for English but we have adapted them to Spanish for the annotation of IMEC. Some of the changes to the rules include:

- The category ADJ:FORM, renamed to ADJ:INFL, includes gender and number agreement errors.
- New category called DET:INFL added for determiner-noun agreement errors.
- NOUN:INFL now encompasses all noun agreement errors; NOUN:NUM is deprecated.
- NOUN:POSS category was eliminated as there is not possessive inflection for nouns in Spanish.
- Addition of new rules for specific spelling (SPELL) mistakes (e.g., accentuation).

The distribution of corrections in IMEC, both in terms of edit type and of error type, is shown in Table 3a. Regarding edit types, the most common are replacements, followed by missing tokens. Unnecessary tokens are rare, partly due to the annotation guidelines mentioned above.

When it comes to error types, there is a clear unbalance in the corpus. Most of the corrections are concentrated on the orthotypographic aspects, mainly spelling and punctuation. Grammar errors are not as common as in other GEC corpora such as FCE (Yannakoudakis, Briscoe, and Medlock, 2011) or NUCLE (Dahlmeier, Ng, and Wu, 2013). These corpora usually contain texts written by language learners. In contrast, IMEC's original authors are native speakers. Their grammar is usually correct but they are less careful in other aspects. Even then, there are many errors related to determiners, prepositions and verbs due to the health professionals' style being quite telegraphic.

Corpus	# Lines	# Tokens
IBECS	1,035,660	25,157,063
SciELO	919,553	26,706,151
PubMED	354,724	4,558,980
SPACCC	15,907	416,494
TMAE incorrect	2,325,844	51,710,613
TMAE correct	2,325,844	56,841,053

Table 4: Size of TMAE and its constituents in terms of the total number of lines and the total number of running tokens.

3.2 Clinical Texts Augmented with Errors (TMAE)

TMAE is the result of a merger of health-related texts from various sources that were pre-processed and induced with errors in order to create a synthetic corpus for GEC. Four different corpora were chosen to be augmentated: IBECS, SciELO, Pubmed (all three are part of the MeSpEn collection by Villegas et al. (2018))[2] and SPACCC (Intxaurrondo, 2018). Altogether, the resulting corpus has a size of over 2.3 million parallel sentences and 51 million tokens with almost 8 million annotations. Table 4 shows in detail the sizes of the different corpora used in TMAE. "TMAE correct" stands for the merger of the different corpora.

The aim of the augmentation was to introduce errors in a way that replicated the error types and distribution found in the IMEC corpus. For this purpose, a set of rules was handcrafted that recreated the most prominent errors in the corpus. The changes include adding or removing words based on their part-of-speech tag, introducing typos or changing the inflection of a word. A total of 24 different rules were developed, each with an assigned probability based on the frequency in IMEC of the error they generate.

The number of edits a sentence experiences is randomly chosen between 1 and 4. To avoid completely changing a sentence, sentence length was also taken into account to set a maximum threshold of edits. This corpus with introduced errors is "TMAE incorrect" in Table 4. The examples in Table 5 show some of the resulting sentence pairs.

As with IMEC, once the parallel sentences were generated, the whole corpus was annotated using ERRANT. Table 3b describes the error distribution of the corpus. When compared with IMEC, most categories are similarly distributed, although some such as PREP or OTHER have grown, and others such as SPELL have decreased in size.

Original:
Aplicación de la metodología enfermera en pacientes con úlceras por presión.
Augmented:
Aplicacion metodología enfermera pacientes con úlceras por presión
Translation:
Application of the nursing methodology in patients with pressure ulcers
Original:
También se discute la necesidad de controlar con imagen la resolución de la TEP tras el tratamiento anticoagulante, actualmente no recomendado en las guías clínicas.
Augmented:
También se discute necesidad de controlar con imagen resolución de la TEPtras el tratamiento anticoagulante, actual no recomendado en las guías clínicas.
Translation:
The need for imaging the resolution of PE after anticoagulant treatment is also discussed, currently not recommended in clinical guidelines.

Table 5: Examples of automatically induced errors.

4 Experimentation

This section documents the experimentation details followed using the resources presented above, from the development of a baseline to training a neural network. Next, we explain the experimentation setup. Results are presented in §4.2 and discussed in §4.3.

4.1 Experimentation design

Figure 1 shows the workflow of our experimentation. IMEC provides training and development data, as well as the gold standard against which to measure the results of the experiments. The sizes of these partitions are shown in Table 6. We also rely on TMAE in order to increase the volume of the training data artificially.

Two systems are evaluated: Aspell,[3] which sets the baseline, and a Multilayer Convolutional Encoder-Decoder (Chollampatt and Ng, 2018) as a more sophisticated solution. Each of these systems is extensively

[2] https://temu.bsc.es/mespen/

[3] https://aspell.net

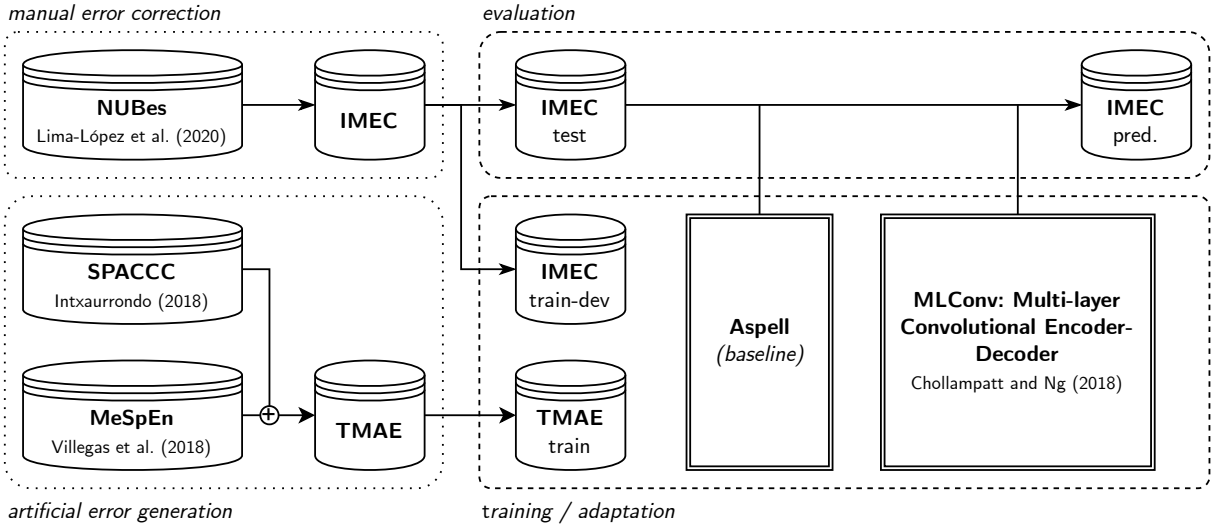

Figure 1: Overview of the different stages of this work.

Partition	# Lines	%
Train	7,507	75
Dev	1,500	15
Test	1,000	10

Table 6: Partitions' size of the IMEC corpus.

evaluated using different key-component settings and/or training set variations.

The output of each system for the test set has been annotated and scored using ERRANT. ERRANT evaluates performance in terms of the $F_{0.5}$ measure, which weights precision twice as much as recall.

4.1.1 Baseline

A baseline system was created as a benchmark for the experimentation. Given the number of spelling mistakes in the IMEC corpus, it was decided that a spell checker would be enough for the task even if it wasn't able to tackle all error types. Ultimately, the free software Aspell was chosen, as it is a renowned spellchecker that allows for some customization.

Aspell uses a dictionary to check whether a word is correctly spelled and suggests possible replacements for misspelled terms based on the metaphone algorithm and a variant of the Levenshtein distance (Atkinson, 2020).

Some of the customizations allowed include using custom dictionaries or applying filters. It is able to return a list with all possible suggestions, which can be further processed. Thus, we tweaked out-of-the-box Aspell as follows:

(i) the predetermined Spanish dictionary was expanded using a vocabulary list extracted from IMEC's train set;

(ii) a Levenshtein distance threshold for suggestions was set; and,

(iii) the suggestions provided were re-ranked using a language model.

The language model was trained on the MedlinePlus corpus, which is part of the MeSpEn collection (Villegas et al., 2018), using the KenLM toolkit (Heafield, 2011) with a window size of 5. Its size can be consulted in Table 7.

4.1.2 MLConv

As a competitive system, we performed experiments with Chollampatt and Ng (2018)'s multilayer convolutional encoder-decoder, MLConv.[4] It is a model that consists of an encoder convolutional network followed by a decoder convolutional network, each with seven layers.

The input to the network are fastText (Bojanowski et al., 2017) word embeddings pre-trained on data segmented with the byte-pair encoding (BPE) algorithm (Sennrich, Haddow, and Birch, 2016). BPE splits rare words into sub-words, helping minimize the number of out-of-vocabulary words. The output of the decoder is an N-best list of corrections. Each candidate is re-ranked using a log-linear framework that calculates features weights on the development set using min-

[4] https://github.com/nusnlp/mlconvgec2018

inum error rate training (Och, 2003). Additionally, edit operation and language model features are also used for scoring.

In this work, the fastText word embeddings were trained on a 2018 dump of the Spanish Wikipedia.[5]

Furthermore, three language models were tested to study the effect of in- and out-of-domain knowledge in the re-ranking step:

- MedlinePlus: the same medical corpus used for the Aspell baseline, MedlinePlus (Villegas et al., 2018).
- NewsLarge: a joint version of multiple NewsCrawl dumps in Spanish released as part of the 2019 Conference on Machine Translation (WMT) (Barrault et al., 2019).
- NewsSmall: in order to make the comparison fairer, we trained a language model with a subset of NewsCrawl that had the same amount of tokens as MedlinePlus.

Table 7 shows the different sizes of the corpora used to build the three language models. These language models were trained with KenLM and a window size of 5 tokens. Apart from that, the parameters documented by Chollampatt and Ng (2018) were set to train the GEC models.

Finally, we trained MLConv with different sets of training data. Given the small size of IMEC, we experimented with oversampling and the incorporation of TMAE to the training data. During the early experimentation phase, IMEC's training section was oversampled with orders of magnitude from 5 to 100, as reported later in the discussion (§4.3). In this work, we only show the 4 best performing combinations of datasets, described in Table 8. From this point on, we will refer to the different oversampling points as IMEC$_{\times N}$, N being the number of repetitions.

Corpus	# Lines	# Tokens
MedlinePlus	445,140	6,461,483
NewsSmall	220,000	6,501,721
NewsLarge	51,833,058	1,588,491,570

Table 7: Size of the corpora used to create the language models.

[5]https://dumps.wikimedia.org/

Corpus train	# Lines	# Tokens
IMEC	7,506	122,812
IMEC$_{\times 75}$	562,950	9,210,900
IMEC + TMAE	2,333,350	51,833,425
IMEC$_{\times 15}$ + TMAE	2,438,434	3,552,793

Table 8: Size of the different training sets for MLConv; the number of tokens correspond to the incorrect partitions.

4.2 Results

Table 9 shows the overall results of the experiments. For each set of experiments, the table indicates the system evaluated and the training data and configuration or re-ranking model used.

On its own, the spellchecker baseline achieves acceptable but low results. Adding specialized, in-domain vocabulary obtained from the training section of IMEC greatly boosts performance. However, any attempts at re-ranking the results, either using the language model trained on the MedlinePlus corpus or capping the suggestions at a given Levenshtein distance, seem to only interfere with Aspell's own ranking and lowers performance. In general, using Aspell returns decent precision scores but really low recall.

The results of training the MLConv network using IMEC are overall better than those obtained by the baseline, even if the corpus' size is small. This improvement is especially appreciated in terms of recall. For each experiment, the effects of re-ranking the output sentences are shown. It seems to have a positive effect, giving a performance boost in comparison to the raw output.

TMAE, which relies on data augmentation as explained in §3.2, is also a great asset. When merged with IMEC, it gave a similar boost to precision as just oversampling IMEC (IMEC$_{\times 75}$), although it seems that recall suffers a little in comparison.

In general, the best performing system is achieved when combining IMEC$_{\times 15}$ with TMAE and re-ranking with MedlinePlus.

4.3 Discussion

The overall picture of the experiments is that neural networks work much better for GEC than spellcheckers. This is something we expected, as the phenomena contained in the corpus are much wider than spelling errors.

Even then, there are some interesting re-

System	Training data / Configuration	Precision	Recall	$F_{0.5}$
Aspell	*as is*	26.44	16.44	23.57
	+ MedlinePlus	17.27	10.69	15.38
	+ train vocab	52.62	14.99	35.03
	+ MedlinePlus + train vocab	30.01	08.59	20.06
	Lev=1 + MedlinePlus	37.95	14.65	28.79
	Lev=1 + MedlinePlus + train vocab	54.80	13.23	33.65
MLConv	IMEC	42.36	41.26	42.14
	IMEC + MedlinePlus	45.23	38.79	43.78
	IMEC + NewsSmall	45.21	38.27	43.63
	IMEC + NewsLarge	46.41	41.03	45.22
	$IMEC_{\times 75}$	53.63	45.81	51.86
	$IMEC_{\times 75}$ + MedlinePlus	62.09	42.71	56.92
	$IMEC_{\times 75}$ + NewsSmall	62.01	42.94	56.95
	$IMEC_{\times 75}$ + NewsLarge	59.85	45.52	56.31
	IMEC + TMAE	62.57	35.43	54.26
	IMEC + TMAE + MedlinePlus	62.14	37.11	54.75
	IMEC + TMAE + NewsSmall	64.00	36.14	55.45
	IMEC + TMAE + NewsLarge	63.89	38.42	56.41
	$IMEC_{\times 15}$	73.71	**59.19**	70.26
	$IMEC_{\times 15}$ + TMAE + MedlinePlus	**76.00**	55.87	**70.89**
	$IMEC_{\times 15}$ + TMAE + NewsSmall	75.30	55.94	70.43
	$IMEC_{\times 15}$ + TMAE + NewsLarge	75.81	55.64	70.69

Table 9: Results of GEC in the IMEC test split. The best results of each experiment set are marked with an underline; the best results overall are highlighted in boldface.

marks that could be made about the baseline. Firstly, Aspell's own ranking system is solid enough that attempting to add any extra layer of suggestion classification only hinders it. Secondly, using in-domain vocabulary almost doubles precision. This highlights the importance of including in-domain vocabulary when dealing with such specialized texts.

An important aspect of our experiments was the re-ranking of the neural networks' output. Unfortunately, each system achieved its best results using a different language model, therefore a conclusion cannot be drawn as to exactly how much the language model's training data's size and domain matter. It is clear, though, that re-ranking is a valuable step, as it improves the system's performance in every single experiment. An argument could be made that, in some cases, precision benefits more than recall from this process. This is not inherently bad, given that in GEC it is preferable to offer good corrections than to suggest dubious corrections for every mistake.

Regarding the creation of the TMAE corpus, it could be said that it was created in a probabilistic way. Interestingly enough, the results of the model that uses it are coherent with the theory presented by Felice (2016) that states that probabilistic generation of synthetic errors increases precision while decreasing recall.

We would also like to provide some insight into the oversampling process. The training section of the IMEC corpus was repeated a different number of times (multiples of 5 up to a 100). A new model was trained with each of them to explore how performance changed. The experiments showed immediate improvement, with a steady increase as the number of repetitions increase and a peak at 75 repetitions. However, after that, the model's performance greatly decreases.

For the joint oversampled IMEC + TMAE model, however, IMEC was repeated only 15 times. After a few experiments, it seemed apparent that a higher number of repetitions seemed to not work as well when combined with more data.

ERRANT's evaluation system also allows us to look at each system's performance individually, as it returns precision, recall and $F_{0.5}$ for each error type. Due to space issues, we are not able to include full tables for each system. These are some highlights:

- As expected, our best system ($IMEC_{\times 15}$ + TMAE) has the best performance in most categories.
- All neural systems return better results for the SPELL category than the spellchecker used for the baseline, usually over 65.00 $F_{0.5}$ score. The ORTH category generally returns really good results across all systems too.
- No system is able to correct ADJ:INFL errors at all (there were few examples in the corpus to begin with), and only $IMEC_{\times 15}$ + TMAE is able to correct some DET:INFL errors. This behaviour is also shown in the VERB:FORM and VERB:SVA categories. This suggests that our convolutional neural network is not able to learn how agreement works and that it may be better suited for languages that are less morphologically rich than Spanish.
- Some of the less frequent categories, such as word order, which has only 4 instances, are not learnt at all by any of the systems.

Finally, an interesting fact that can be appreciated upon manual error analysis is that sometimes the models return correct examples that are not evaluated as such, since they differ from the gold standard. For instance, the system may insert the verb 'presentar' (*to present*) instead of 'mostrar' (*to show*) when a verb is missing. This is actually correct but, due to the lack of multiple annotations for each sentence, it is evaluated as incorrect. Extending our corpus with more data and multiple annotators is left as future work.

5 Conclusion

In conclusion, this paper presents a first approach to Grammatical Error Correction for health records in Spanish. This is a topic that has not been previously explored, but that we consider may have a great impact. Health records are the main form of communication between health specialists and patients, but their form is a flawed aspect that usually contains multiple orthographic and grammatical problems.

GEC may be a helpful solution to this problem. This work introduces the IMEC (*Informes Médicos en Español Corregidos*) corpus–which is made up of over 10,000 manually corrected sentences from health records–as well as the TMAE (*Textos Médicos Aumentados con Errores*) corpus, a parallel collection of over 2 million sentences from the clinical domain augmented with errors.

Additionally, we present extensive experimentation with a Multilayer Convolutional Encoder-Decoder (Chollampatt and Ng, 2018) and the corpora generated. The results show promising results in this line and suggest that it is possible to obtain good results even with small datasets.

As future work, one of the most important steps we would like to take is to expand the IMEC corpus, not only with more data but also with more annotators. Given the subjectivity this field shows, having multiple possibilities for a correction is almost compulsory to avoid false negatives like the ones described at the end of §4.3.

Another significant step would be to test the impact this type of correction has on other NLP tasks via extrinsic evaluation on information extraction or anonymization systems. If clinical GEC systems make text less noisy, it may prove helpful for text processing in general.

Finally, we plan on performing new experiments with more competitive systems based on the Transformers architecture and large pre-trained language models, which have achieved a widespread success in virtually every NLP task.

Acknowledgments

This work has been supported by Vicomtech and partially funded by the projects DeepText (KK-2020-00088, SPRI, Basque Government) and DeepReading (RTI2018-096846-B-C21, MCIU/AEI/FEDER, UE). We also want to thank Olatz Pérez de Viñaspre, who has collaborated in the research behind this article and whose contributions have been essential.

References

Aguilar Ruíz, M. J. 2013. Las normas ortográficas y ortotipográficas de la nueva Ortografía de la lengua española (2010) aplicadas a las publicaciones biomédicas en español: una visión de conjunto. *Panace@: Revista de Medicina, Lenguaje y Traducción*, XIV(37):101–120.

Atkinson, K. 2020. GNU Aspell 0.61 documentation.

Barrault, L., O. Bojar, M. R. Costa-jussà, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, S. Malmasi, C. Monz, M. Müller, S. Pal, M. Post, and M. Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Bello Gutiérrez, P. 2016. Aprendiendo a redactar mejor tus informes. *Curso de Actualización Pediatría*, pages 391–400.

Beloki, Z., X. Saralegi, K. Ceberio, and A. Corral. 2020. Grammatical error correction for basque through a seq2seq neural architecture and synthetic examples. *Procesamiento del Lenguaje Natural*, 65:13–20, September.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Boletín Oficial del Estado. 2015. Real decreto 9/2015, de 6 de febrero, por el que se regula el registro de actividad de atención sanitaria especializada.

Bryant, C. 2019. *Automatic annotation of error types for grammatical error correction.* University of Cambridge.

Bryant, C., M. Felice, Ø. E. Andersen, and T. Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Chollampatt, S. and H. T. Ng. 2018. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5755–5762, New Orleans, Louisiana, USA. AAAI Press.

Dahlmeier, D., H. T. Ng, and S. M. Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia. Association for Computational Linguistics.

Davidson, S., A. Yamada, P. Fernandez Mira, A. Carando, C. H. Sanchez Gutierrez, and K. Sagae. 2020. Developing NLP tools with a new corpus of learner Spanish. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 7238–7243, Marseille, France. European Language Resources Association.

Felice, M. 2016. Artificial error generation for translation-based grammatical error correction. Number 895.

Felice, M., C. Bryant, and T. Briscoe. 2016. Automatic extraction of learner errors in ESL sentences using linguistically enhanced alignments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 825–835, Osaka, Japan. The COLING 2016 Organizing Committee.

Gamon, M., J. Gao, C. Brockett, A. Klementiev, W. B. Dolan, D. Belenko, and L. Vanderwende. 2008. Using contextual speller techniques and language modeling for ESL error correction. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, pages 449–456. Asia Federation of Natural Language Processing.

Granger, S. 1998. *The computer learner corpus: a versatile new source of data for SLA research.* na.

Grundkiewicz, R., M. Junczys-Dowmunt, and K. Heafield. 2019. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In

Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 252–263, Florence, Italy. Association for Computational Linguistics.

Heafield, K. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Intxaurrondo, A. 2018. SPACCC. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

Lee, J. and S. Seneff. 2008. Correcting misuse of verb forms. In *Proceedings of ACL-08: HLT*, pages 174–182, Columbus, Ohio, USA. Association for Computational Linguistics.

Lima López, S., N. Pérez, M. Cuadros, and G. Rigau. 2020. NUBes: A Corpus of Negation and Uncertainty in Spanish Clinical Texts. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC2020)*, pages 5772–5781, Marseille, France. European Language Resources Association.

Macdonald, N. H. 1983. Human factors and behavioral science: The UNIX Writer's Workbench software: Rationale and design. *The Bell System Technical Journal*, 62(6):1891–1908.

Ministerio de Sanidad. 2018. Recursos físicos, actividad y calidad de los servicios sanitarios.

Mizumoto, T., Y. Hayashibe, M. Komachi, M. Nagata, and Y. Matsumoto. 2012. The effect of learner corpus size in grammatical error correction of ESL writings. In *Proceedings of COLING 2012: Posters*, pages 863–872, Mumbai, India. The COLING 2012 Organizing Committee.

Náplava, J. and M. Straka. 2019. Grammatical error correction in low-resource scenarios. In *Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text*, pages 346–356, Hong Kong, China. Association for Computational Linguistics.

Ng, H. T., S. M. Wu, T. Briscoe, C. Hadiwinoto, R. Susanto, and C. Bryant. 2014. The conll-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14. Association for Computational Linguistics.

Ng, H. T., S. M. Wu, Y. Wu, C. Hadiwinoto, and J. Tetreault. 2013. The conll-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12. Association for Computational Linguistics.

Och, F. J. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July. Association for Computational Linguistics.

Omelianchuk, K., V. Atrasevych, A. Chernodub, and O. Skurzhanskyi. 2020. GECToR – Grammatical Error Correction: Tag, Not Rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA. Association for Computational Linguistics.

Rei, M., M. Felice, Z. Yuan, and T. Briscoe. 2017. Artificial error generation with machine translation and syntactic patterns. *CoRR*, abs/1707.05236.

Richardson, S. D. and L. C. Braden-Harder. 1988. The experience of developing a large-scale natural language text processing system: Critique. In *Second Conference on Applied Natural Language Processing*, pages 195–202, Austin, Texas, USA. Association for Computational Linguistics.

Sennrich, R., B. Haddow, and A. Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Shen, L., A. Sarkar, and F. J. Och. 2004. Discriminative reranking for machine translation. In *Proceedings of the*

Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, pages 177–184, Boston, Massachusetts, USA. Association for Computational Linguistics.

Tajiri, T., M. Komachi, and Y. Matsumoto. 2012. Tense and aspect error correction for ESL learners using global context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 198–202, Jeju Island, Korea. Association for Computational Linguistics.

Terroba Reinares, A. R. 2015. *Mejora de la calidad del informe clínico de alta hospitalaria desde el punto de vista lingüístico*. Universidad de La Rioja.

Tetreault, J., J. Foster, and M. Chodorow. 2010. Using parse features for preposition selection and error detection. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 353–358, Uppsala, Sweden. Association for Computational Linguistics.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., pages 5998–6008.

Villegas, M., A. Intxaurrondo, A. Gonzalez-Agirre, M. Marimon, and M. Krallinger. 2018. The MeSpEN resource for English-Spanish medical machine translation and terminologies: Census of parallel corpora, glossaries and term translations. In *Proceedings of the LREC 2018 Workshop "MultilingualBIO: Multilingual Biomedical Text Processing"*, pages 32–39. European Language Resources Association.

Xie, Z., A. Avati, N. Arivazhagan, D. Jurafsky, and A. Y. Ng. 2016. Neural language correction with character-based attention. *CoRR*, abs/1603.09727.

Yannakoudakis, H., Ø. E. Andersen, A. Geranpayeh, T. Briscoe, and D. Nicholls. 2018. Developing an automated writing placement system for esl learners. *Applied Measurement in Education*, 31(3):251–267.

Yannakoudakis, H., T. Briscoe, and B. Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 133-140 recibido 11-10-2020 revisado 24-01-2021 aceptado 01-02-2021

Mejoras aplicadas a la extracción de relaciones semánticas para la Web en español

Improvements applied to Open Information Extraction in Spanish

Juan M. Rodríguez[1], Hernán D. Merlino[2], Patricia Pesado[1]
[1] Facultad de Informática. Universidad Nacional de La Plata
[2]Departamento de Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús
jmrodriguez1982@gmail.com, hmerlino@gmail.com, ppesado@lidi.info.unlp.edu.ar

Resumen: En este trabajo de investigación se presenta un método novedoso de extracción de relaciones semánticas para la Web en español llamado ECMes. Este método es comparado con otros métodos similares en el *estado-del-arte* utilizando para ello dos conjuntos de prueba diferentes: uno conocido y ya utilizado en trabajos anteriores relacionados y otro construido específicamente para este artículo utilizando fuentes heterogéneas de datos. En el primero de los conjuntos de datos ECMes obtuvo la mejor precisión, la mejor exhaustividad (*recall*) y por lo tanto la medida F1 más alta en relación a los otros métodos evaluados. En el segundo de los conjuntos de datos obtuvo la mejor exhaustividad (*recall*) y la medida F1 más alta.
Palabras clave: Extracción de conocimiento, extracción de relaciones semánticas, procesamiento de lenguaje natural.

Abstract: This paper introduces a novel method of Open Information Extraction in Spanish called ECMes. This method is compared to other similar methods in the *state-of-the-art* using two different testing datasets: one is a well known dataset, already used in previous related works and the second dataset was constructed specifically for this article using heterogeneous data sources. In the first dataset ECMes obtained the highest precision, recall and F1 measurement. In the second dataset, ECMes obtained the highest recall and the highest F1 measurement.
Keywords: Knowledge extraction, semantic relationship extraction, natural language processing, Open Information Extraction.

1 Introducción

En este documento presentamos un método novedoso de extracción de relaciones semánticas para la Web en idioma español llamado ECMes (Extractor de Conocimiento Mejorado en Español). ECMes está construido sobre la arquitectura base de TP-OIE-ES (Rodríguez y Merlino, 2020) pero implementa una serie de mejoras sobre el algoritmo original que le permiten incrementar considerablemente su precisión, elevar su exhaustividad (*recall*) y por lo tanto mejorar la medida de rendimiento F (*F-measure*).

En las siguientes secciones se detallarán cuáles fueron las mejoras introducidas en el algoritmo, cómo se construyó el conjunto de datos de prueba y cuáles fueron los criterios utilizados para evaluar los diferentes métodos.

La extracción de relaciones semánticas para la Web, u otros grandes corpus de datos, (en inglés *Open Information Extraction*) es un paradigma de extracción de información presentado por primera vez en (Banko et al., 2007). En dicho trabajo se presenta al método TEXTRUNNER, un sistema informático capaz de devolver una tupla de la forma: *(argumento 1, relación, argumento 2)* por cada relación semántica existente en una oración dada como entrada. Típicamente información fáctica del tipo: "Quién hizo qué a quién y cómo" (Rodríguez et al., 2018).

ISSN 1989-7553. DOI 10.26342/2021-66-11

Un método de OIE (*Open Information Extraction)* debe cumplir las siguientes condiciones según Banko et al. (2007).

- Hacer una sola pasada por el corpus garantizando la escalabilidad, independientemente del tamaño de este.
- Ser independiente del dominio.
- Tener un corpus como única entrada (*input*). Su salida (*output*) debe ser un conjunto de relaciones extraídas.
- Ser no supervisado.
- Extraer cualquier relación existente y no depender de relaciones previamente establecidas por el desarrollador.

2 Trabajos relacionados

Investigaciones documentales (Glauber y Barreiro, 2018; Rodríguez, Merlino y García-Martínez, 2015) muestran que fueron desarrollados diversos métodos de OIE, aunque no todos están disponibles de forma pública. De todos los métodos presentados solo algunos fueron comparados de forma experimental con métodos ya existentes para determinar su desempeño relativo. La mayoría de ellos fueron creados para el idioma inglés exclusivamente. Los más destacados en cantidad y calidad de extracciones son:

- ClausIE (Del Corro y Gemulla, 2103).
- ReVerb (Fader, Soderland y Etzioni, 2011).
- OLLIE (Schmitz et al., 2012).
- MinIE (Gashteovski, Gemulla y Del Corro, 2017).
- ArgOE (Gamallo y Garcia, 2015).
- Stanford OpenIE (Angeli, Premkumar y Manning, 2015).
- DepOE (Gamallo y Garcia, 2012).
- ExtrHech (Zhila y Gelbukh, 2013).

De la lista anterior solo trabajan con textos en idioma español ExtrHech, ArgOE y DepOE. ArgOE y DepOE fueron diseñados para soportar múltiples lenguajes (Glauber y Barreiro, 2018). ExtrHech soporta español e inglés (Zhila y Gelbukh, 2013).

Para los métodos en español: DepOE (2012), ExtrHech (2014) y ArgOE (2015), se muestra la precisión en la Tabla 1. Corresponde a las pruebas realizadas por los autores de cada método, desafortunadamente no se cuenta con el valor de exhaustividad.

Medidas	**ExtrHech**	**ArgOE**	**DepOE**
Precisión	0.55	0.55	**0.68**

Tabla 1: Resultados obtenidos de los trabajos: (Zhila y Gelbukh, 2013; Gamallo y Garcia, 2015; Gamallo y Garcia, 2012).

En (Rodríguez y Merlino, 2020) se presenta un nuevo método de *OIE* en idioma español llamado TP-OIE-ES. En el mismo trabajo se realizan pruebas para evaluar la precisión, la exhaustividad (*recall*) y la medida F1 de dicho método y de los métodos ArgOE y DepOE. Las pruebas se realizaron con un conjunto de datos compuesto por 69 oraciones extraídas de Wikipedia y propuestas por Gamallo y García (2015). En la Tabla 2 se muestran los resultados obtenidos:

Medidas	**TP-OIE-ES**	**DepOE**	**ArgOE**
Precisión	0.62	**0.89**	0.67
Exhaustividad	**0.36**	0.29	0.29
Medida F1	**0.46**	0.44	0.40

Tabla 2: Resultados obtenidos en (Rodríguez y Merlino, 2020) con el mismo conjunto de prueba utilizado en Gamallo y García (2015).

3 Problemas abiertos

Aunque TP-OIE-ES logra mejorar la exhaustividad y con ello una medida F1 más alta respecto a los otros métodos evaluados, en la prueba realizada por Rodríguez y Merlino en (2020) quedan al menos tres problemas importantes por ser resueltos. Los mismos se explican en las siguientes subsecciones.

3.1 Mejorar precisión

Si bien la medida F1, que evalúa conjuntamente la precisión y la exhaustividad con igual ponderación es más alta para TP-OIE-ES que para los otros métodos, la diferencia es de apenas 0.02 puntos con respecto a DepOE. Y la precisión de TP-OIE-ES es la más baja de los tres métodos. La precisión se convierte en el punto más débil de este método y es un problema que debe ser resuelto si se pretende mejorar al algoritmo.

3.2 Mejorar la evidencia disponible

Para soportar con mayor evidencia el desempeño de los métodos y sus medidas de rendimiento es necesario realizar por lo menos una segunda evaluación contra un conjunto de datos de prueba que no sea tan uniforme como el presentado por Gamallo y García en (2015), el cual está compuesto solo por sentencias extraídas de Wikipedia.

Además uno de los principios de los métodos de OIE es que estos son independientes del dominio. Es decir que deberían funcionar de forma similar con cualquier conjunto de textos de entrada para un mismo idioma (Banko et al., 2007).

3.3 Mejorar la informatividad de las extracciones realizadas

Muchas de las extracciones semánticas realizadas por el método TP-OIE-ES son correctas pero poco informativas, por ejemplo en la oración siguiente:

(1) *La bibliografía se estructura con los datos de las fichas bibliográficas de esos textos.*

Extrajo la siguiente tupla:

(2) *(La bibliografía; se estructura; con los datos de las fichas).*

La relación semántica es correcta ya que corresponde a lo expresado en la oración, sin embargo sería más informativa en este caso particular, si el argumento segundo contuviese al resto de las palabras. Por ejemplo:

(3) *(La bibliografía; se estructura; con los datos de las fichas bibliográficas de esos textos).*

La cantidad de información que debe contener una tupla para expresar correctamente la relación semántica expresada en una oración es una cuestión subjetiva. Dependerá en parte del problema que se esté intentando resolver y del procesamiento posterior que reciban las tuplas extraídas. Considérese la oración del ejemplo 4 a continuación:

(4) *Albert Einstein fue galardonado con el Premio Nobel en Suecia en 1921.*

Y las siguientes tuplas posibles:

1. *(Albert Einstein; fue galardonado con; el Premio Nobel).*
2. *(Albert Einstein; fue galardonado con; el Premio Nobel en Suecia en 1921).*
3. *(Albert Einstein; fue galardonado en; 1921).*
4. *(Albert Einstein; fue galardonado en; Suecia).*

Todas las tuplas expresan relaciones semánticas correctas. Sin embargo la tupla 2 es la más informativa, aunque no necesariamente es preferible por sobre las demás. Si el problema a resolver fuese, por ejemplo: "Ganadores del Premio Nobel por año" la tupla 3 sería más conveniente. No obstante, si un método devuelve una tupla como la 3, es esperable que genere al menos dos tuplas más como la 4 y la 1 para que toda la información existente en la oración original quede reflejada en las relaciones semánticas extraídas. Pero si las posibilidades del método consisten en que devuelva solo la 3 o solo la 2, la 2 es en este caso preferible para propósitos generales, ya que no hay pérdida de información.

Teniendo en cuenta lo anterior, en las pruebas realizadas por Rodríguez y Merlino en (2020) de las 49 extracciones semánticas marcadas como correctas, 19 fueron marcadas como "poco informativas". Es decir que casi un 39 % de las extracciones pueden ser mejoradas o están expresando la idea principal de la oración de forma pobre. Por lo cual mejorar la informatividad de las relaciones semánticas extraídas es un punto importante para mejorar el desempeño global del algoritmo.

4 ECMes

ECMes es una versión mejorada del método TP-OIE-ES. El algoritmo principal de TP-OIE-ES intenta identificar relaciones semánticas en una oración dada utilizando una lista de patrones. Estos patrones son generados automáticamente por el mismo método a partir de un conjunto de datos etiquetados, entiéndase en este caso una lista de oraciones con sus respectivas relaciones semánticas, expresadas estas en forma de tupla. Una vez entrenado, utiliza estos patrones para encontrar coincidencias en el árbol de dependencias sintácticas de la oración (enriquecido con información adicional: la categoría gramatical y los nombres de las entida-

des halladas). TP-OIE-ES buscará exhaustivamente todas las coincidencias existentes.

Estas coincidencias permiten identificar la relación propiamente dicha y el argumento primero (también llamado sujeto o primera entidad). El argumento segundo es obtenido mediante una serie de reglas para buscar la frase nominal más próxima, tal y como lo hace ReVerb (Fader, Soderland y Etzioni, 2011) en idioma inglés. Sin embargo, si encuentra una relación candidata y un argumento segundo candidato, pero no encuentra un argumento primero usando la lista de patrones, intentará encontrar este argumento inicial buscando la frase nominal más próxima a la izquierda de la relación.

La versión original de TP-OIE-ES fue entrenada con oraciones etiquetadas en idioma inglés. Si bien el árbol de dependencias sintácticas que se puede construir para idioma inglés es válido también en español ya que el *parser* utilizado es *depparse* de la biblioteca *Stanford CoreNLP* (Chen y Manning, 2014) el cual es universal. Esto significa que las aristas que conectan las palabras son siempre las mismas independientemente del idioma (Buchholz y Marsi, 2006). Sin embargo los ejemplos provistos en idioma inglés podrían no ser representativos de las oraciones más comunes del idioma español. Es decir, los patrones generados son válidos pero no necesariamente útiles, porque quizás son patrones que permiten identificar relaciones semánticas en oraciones válidas pero poco frecuentes, como podrían ser oraciones en idioma español con una estructura sintáctica similar a la de una frase en idioma inglés.

Por otro lado las categorías gramaticales utilizadas en la versión original de TP-OIE-ES son conocidas como *Penn Treebank POS tags* (Ratnaparkhi, 1996) y son exclusivas del idioma inglés, mientras que para español tanto TP-OIE-ES como su versión mejorada ECMes utilizan las categorías gramaticales *Universal POS tags* (Petrov, 2011). Por ende para buscar las coincidencias de los patrones, las categorías gramaticales que aparecían en estos debían ser traducidas en TP-OIE-ES de un sistema al otro (Rodríguez y Merlino, 2020). Como no existe una correlación unívoca entre estos dos sistemas, esta traducción es susceptible a introducir algún error.

4.1 Mejorar precisión y exhaustividad

Para intentar mejorar las medidas de rendimiento se implementaron tres mejoras en el algoritmo original.

4.1.1 Regeneración de los patrones de búsqueda

La principal medida que se tomó para mejorar la precisión y exhaustividad del método original fue la de reentrenar al mismo con un nuevo conjunto de ejemplos, esta vez en idioma español. Esto implicó convertir los métodos nativos que trabajaban con las categorías gramaticales en idioma inglés en el formato de *Penn Treebank POS tags* al formato *Universal POS tags*. No solo para el sistema de búsqueda de coincidencias de patrones, sino también para el sistema de puntajes, el cual utiliza información de las oraciones como la categoría gramatical para asignarle un puntaje a una relación semántica extraída. Solo aquellas que logran cierto puntaje son devueltas.

La base de datos de entrenamiento se construyó con un total de 110 oraciones en idioma español. De dichas oraciones se extrajo un total de 209 relaciones semánticas de forma manual. El resumen se puede apreciar en la Tabla 3.

Fuente	Oraciones	Relaciones
es.wikipedia.org	33	62
tweets Covid-19	13	13
tweets municipalidad	9	15
libros	23	54
periódicos de noticias	32	65

Tabla 3: Cantidad de oraciones por origen, junto a sus relaciones semánticas en el conjunto de entrenamiento.

El detalle de las oraciones y sus respectivas relaciones semánticas puede hallarse en la siguiente URL: https://bit.ly/3a5VgUT. Los tweets y las frases de periódicos de noticias fueron obtenidas al azar de tres conjuntos de datos separados disponibles públicamente en el sitio https://www.kaggle.com. Las frases de libros fueron extraídas manualmente de diversos libros.

A este conjunto de datos de entrenamiento se sumó uno adicional con 12 oraciones creadas especialmente para propósitos de prueba durante la fase de desarrollo del método. La mayoría de estas frases son traducciones de las frases

propuestas en (Del Corro y Gemulla, 2013) como ejemplos de los diferentes tipos de oraciones para el idioma inglés.

4.1.2 Mejora en el sistema de puntaje

La otra tarea que se adicionó para mejorar la precisión fue mejorar el sistema de puntaje. El sistema de puntaje asigna valores positivos o negativos a las extracciones encontradas para determinar qué tan correctas son y finalmente determinar si serán devueltas o no. Este sistema está basado en el que propuso Fader y Etzioni en (2011). Contiene las mismas reglas básicas (menos una que fue eliminada) más 7 reglas adicionales. La regla eliminada asignaba un puntaje negativo a oraciones con más de 20 palabras. Las reglas adicionadas se listan en la Tabla 4.

Regla	Puntaje
Si **e2** igual **e1**	-100
Si **e1** contiene a **r**	-100
Si **e1** termina con la misma palabra con que empieza **r**	-50
Si **e2** es un determinante	-1000
Si *longitud*(**r**) = 1 palabra y **r** es determinante	-200
Si *longitud*(**r**) = 1 palabra y **r** es verbo	10

Tabla 4: Reglas adicionales utilizadas para puntuar las relaciones semánticas extraídas, teniendo cada una la forma: (e1, r, e2).

4.1.3 Detección de sujetos tácitos

Un problema adicional del idioma español es el sujeto tácito, si bien este existe en idioma inglés es poco frecuente y en general solo se omite el sujeto si ya fue nombrado antes en la misma oración. Esta particularidad del idioma español provocaba que el algoritmo no detectase relaciones semánticas en muchas oraciones, ya que no encontraba dentro de la misma un sujeto candidato para el primer argumento. Supóngase la oración del ejemplo 5.

(5) *Jugábamos al fútbol.*

La relación semántica en este caso, indica que "*nosotros"* es el argumento primero, el sujeto que está relacionado con "*fútbol*" mediante la relación "*jugar al*". La tupla debería quedar de la siguiente forma:

(6) *(Nosotros; jugábamos; al fútbol).*

Sin embargo, el algoritmo original no encontrará nunca una palabra en la oración de entrada que pueda asociar al argumento primero, simplemente porque esa palabra no está presente. Para resolver este problema, cuando el algoritmo no es capaz de hallar un argumento primero, añadirá al comienzo de la oración una palabra comodín, que será analizada como un pronombre personal cualquiera. Si con esta nueva palabra encuentra una relación semántica tal que el comodín coincide con el argumento primero, el algoritmo devolverá una tupla con el argumento primero vacío. Siguiendo con el ejemplo anterior:

(7) *(; jugábamos; al fútbol).*

El espacio inicial vacío indica que hay un sujeto tácito en la relación semántica devuelta. Esta mejora busca no solo aumentar la precisión sino también la exhaustividad (*recall*) del método.

4.2 Añadir evidencia

Se construyó un conjunto de datos de prueba usando las mismas fuentes utilizadas para la construcción del conjunto de datos de entrenamiento. Este conjunto de datos consta de 55 oraciones diferentes y un total de 120 relaciones semánticas extraídas de forma manual, Tabla 5.

Fuente	Oraciones	Relaciones
es.wikipedia.org	16	36
tweets Covid-19	6	10
tweets municipalidad	5	9
Libros	12	29
periódicos de noticias	16	36

Tabla 5: Cantidad de oraciones por origen, junto a sus relaciones semánticas en el conjunto de pruebas.

El detalle de las oraciones y sus respectivas relaciones semánticas puede hallarse en la siguiente URL: https://bit.ly/3a5VgUT

4.3 Mejorar la calidad de las extracciones realizadas

Para mejorar la calidad de las extracciones realizadas, se implementaron tres mejoras al algoritmo original, las cuales se detallan en las secciones siguientes.

4.3.1 Expandir la relación

El algoritmo original está pensado para construir la relación propiamente dicha según patrones de coincidencia en el árbol de dependencias sintácticas. Esto implica que puede tomar palabras no consecutivas dentro de la oración para formar la relación. Teniendo en cuenta el ejemplo 1, podría construir la relación: *fue galardonado en,* aunque las palabras *galardonado* y *en* no son consecutivas. Si bien en este ejemplo esto es correcto, hay muchos casos en los cuales se generan relaciones poco informativas o bien el algoritmo no logra encontrar un argumento segundo para la relación armada. Para estos casos se decidió ampliar la relación y que esta contenga todas las palabras existentes entre su palabra inicial y final. Para el ejemplo 1, quedaría: *fue galardonado con el Premio Nobel en.*

4.3.2 Agregar nombres de entidades

Existen casos en donde una oración contiene un nombre de entidad (NER) y sin embargo esta no aparece en la extracción realizada. Por ejemplo, TP-OIE-ES para la siguiente oración:

(8) *El tío Juan nos escribió una carta.*

Extrajo la siguiente tupla:

(9) (El tío; nos escribió; una carta).

Esto se debe a que el *parser* superficial que busca un posible argumento primero, no detecta a la entidad (en este caso Juan) como parte de la frase nominal. Se agregó en este caso una corrección al algoritmo para que no ignore entidades detectadas que están junto a frases nominales candidatas.

4.3.3 Tener en cuenta múltiples verbos

El método original fallaba al extraer la relación en oraciones donde aparecen dos o más verbos seguidos, ya que, por lo general, solo extrae uno solo de los verbos, respetando las coincidencias del patrón de extracción. Para ilustrar este punto, supóngase que una oración como la del ejemplo 10, (cuyo árbol de dependencias sintáctico se muestra en la Figura 1), fue utilizada para entrenar al algoritmo.

(10) La ciencia mejoró la sociedad.

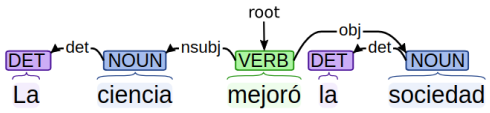

Figura 1: Árbol de dependencias sintáctico y categorías gramaticales. Oración del ejemplo 10.

La relación que en este caso es: *mejoró,* es el verbo raíz en el árbol de dependencias sintácticas. Este ejemplo generará un patrón que servirá para identificar a cualquier verbo raíz como una posible *relación* candidata para la tupla. Con lo cual, una oración como la del ejemplo 11, cuyo árbol de dependencias sintácticas se muestra en la Figura 2, detectará como *relación* candidata la palabra: *permitido* e ignorará el verbo *mejorar*.

(11) La ciencia ha permitido mejorar la sociedad.

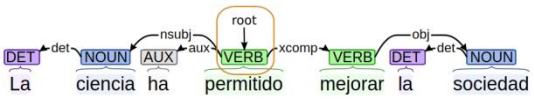

Figura 2: Árbol de dependencias sintáctico y categorías gramaticales, oración del ejemplo 11. En naranja se muestra la coincidencia del patrón utilizado.

En este ejemplo el patrón tampoco está considerando al verbo auxiliar: *ha*, aunque este podría ser detectado por un patrón diferente.

La mejora que se introdujo al método original implica mantener unidos los verbos que están juntos en la oración (incluidos los auxiliares). Esto ha permitido mejorar la informatividad en las relaciones extraídas.

5 Resultados

Para medir la precisión, la exhaustividad (*recall*) y la medida F1 se utilizaron los dos conjuntos de pruebas mencionados en las secciones 2.1 y 4.2. En el primer conjunto se comparó ECMes con otros tres métodos de Open IE en español: ArgOE (Gamallo y Garcia, 2015), DepOE (Gamallo y Garcia, 2012) y TP-OIE-ES (Rodríguez y Merlino, 2020). Y en el segundo conjunto se lo comparó solo contra ArgOE y DepOE.

Se detallan a continuación las fórmulas utilizadas para calcular la precisión, la exhaustividad y la medida F:

$$\text{Precisión} = \frac{\text{extracciones correctas}}{\text{total de extracciones}} \quad (1)$$

$$\text{Exhaustividad} = \frac{\text{extracciones correctas}}{\text{total extracciones manuales}} \quad (2)$$

$$F_\beta = \frac{(1+\beta^2) \cdot \text{Precisión} \cdot \text{Exhaustividad}}{(\beta^2 \cdot \text{Precisión}) + \text{Exhaustividad}} \quad (3)$$

En fórmula (2) para la variable: *total extracciones manuales* se asume que las relaciones semánticas extraídas de forma manual conforman la totalidad de las existentes. Para el conjunto de datos de prueba de Gamallo y García es 137 según se indica en el trabajo de (2015). Para el nuevo conjunto de datos, descripto en la sección 4.2 es de 122.

En la fórmula (3) el parámetro β se estableció igual a 1, para que la precisión y la exhaustividad tuviesen el mismo peso en la fórmula. Por ello nos referimos a la medida F directamente como *medida F1* o simplemente *F1*.

Los resultados obtenidos se resumen en las tablas 6 y 7.

Medidas	**DepOE**	**ArgOE**	**TP-OIE-ES**	**ECMes**
Precisión	0.89	0.67	0.62	**0.92**
Exhaustividad	0.29	0.29	0.36	**0.42**
Medida F1	0.44	0.40	0.46	**0.57**

Tabla 6: Resultados obtenidos sobre el conjunto de datos de prueba de Gamallo y García (2015).

Medidas	**DepOE**	**ArgOE**	**ECMes**
Precisión	**0.81**	0.68	0.68
Exhaustividad	0.18	0.22	**0.34**
Medida F1	0.30	0.33	**0.45**

Tabla 7: Resultados obtenidos sobre el conjunto de datos de prueba presentado en la sección 4.2.

Como puede observarse en los resultados mostrados en ambas tablas ECMes supera a los otros métodos con un margen de al menos 10 puntos porcentuales para la medida F1. Y en el primero de los conjuntos de prueba supera incluso a DepOE en precisión. Sin embargo la precisión disminuye bastante en el segundo conjunto de pruebas, el cual tiene oraciones mucho menos uniformes, aunque en compensación logra mantener una exhaustividad relativamente alta en relación a los otros dos métodos.

Respecto a la informatividad de las relaciones semánticas extraídas, en el conjunto de pruebas de la tabla 6, solo 4 relaciones semánticas fueron identificadas como poco informativas a diferencia del método original TP-OIE-ES que tenía un total de 19 relaciones semánticas poco informativas. Además como extrajo mayor cantidad de relaciones semánticas el porcentaje de relaciones poco informativa cayó de 39% a solo 7%. Para el conjunto de datos presentado en la tabla 7 el número de extracciones semánticas poco informativas es de 13.

El código fuente del método junto a la totalidad de las pruebas realizadas pueden encontrase en *Github* en la siguiente URL:

https://github.com/juanma1982/ECMes

6 Conclusiones

A partir de los resultados presentados en la sección 5 podemos concluir que el método propuesto en este artículo: ECMes supera al método original sobre el cual está construido: TP-OIE-ES y que puede ubicarse entre los métodos de *Open Information Extraction* en idioma español en el *estado-del-arte*.

Si bien ECMes está construido como una serie de mejoras sobre TP-OIE-ES, al haber reentrenado al algoritmo desde cero utilizando datos en español, y al haber adaptado todo el algoritmo de búsqueda de coincidencias por patrones a idioma español nos hemos alejado un poco de la propuesta original de TP-OIE que pretendía construir patrones universales capaces de ser utilizados en diferentes idiomas con solo algunas pequeñas adaptaciones. TP-OIE-ES es la adaptación a español de TP-OIE. Sin embargo, ECMes muestra que la focalización exclusiva en idioma español arroja mejores resultados.

Bibliografía

Angeli, G., M. J. Premkumar, y C. D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint*

Conference on Natural Language Processing, 1, págs. 344--354.

Banko, M., M. J. Cafarella, S. Soderland, M. Broadhead, y O. Etzioni. 2007. Open information extraction for the web. *IJCAI , 7*, 2670-2676.

Buchholz, S. y E. Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. *In Proceedings of the tenth conference on computational natural language learning*, (págs. 149-164).

Chen, D. y C. Manning. 2014. A fast and accurate dependency parser using neural networks. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, (págs. 740--750).

Del Corro, L. y R. Gemulla. 2103. ClausIE: clause-based open information extraction. *22nd international conference on World Wide Web*, (págs. 355-366).

Fader, A., S. Soderland, y O. Etzioni. 2011. Identifying relations for open information extraction. *Association for Computational Linguistics*, (págs. 1535-1545).

Gamallo, P. y M. Garcia. 2012. Dependency-based open information extraction. En A. f. Linguistics (Ed.), *Proceedings of the joint workshop on unsupervised and semi-supervised learning in NLP*, (págs. 10--18).

Gamallo, P. y M. Garcia. 2015. Multilingual open information extraction. En Springer (Ed.), *Portuguese Conference on Artificial Intelligence*, (págs. 711--722).

Gashteovski, K., R. Gemulla, y L. Del Corro. 2017. Minie: minimizing facts in open information extraction. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, (págs. 2630--2640).

Glauber, R. y D. Barreiro Claro. 2018. A systematic mapping study on open information extraction. *Expert Systems with Applications* (págs. 372--387). Elsevier.

Petrov, S. D. 2011. A universal part-of-speech tagset.

Ratnaparkhi, A. 1996. A maximum entropy model for part-of-speech tagging. *In Conference on Empirical Methods in Natural Language Processing*.

Rodríguez, J. M. y H. D. Merlino. 2020. TP-OIE-ES: Método autónomo de extracción de relaciones semánticas para la Web en Español. *Conferencia Iberoamericana de Complejidad, Informática y Cibernética: CICIC 2020*. Orlando, Florida, USA. Manuscript submitted for publication.

Rodríguez, J. M., H. D. Merlino, P. Pesado, y R. García-Martínez. 2018. Evaluation of open information extraction methods using Reuters-21578 database. En ACM (Ed.), *2nd International Conference on Machine Learning and Soft Computing (ICMLSC '18)*, (págs. 87--92).

Rodríguez, J. M., H. D. Merlino, y R. García-Martínez. 2015. Revisión Sistemática Comparativa de Evolución de Métodos de Extracción de Conocimiento para la Web. *XXI Congreso Argentino de Ciencias de la Computación (CACIC 2015)*. Buenos Aires, Argentina.

Schmitz, M., R. Bart, S. Soderland, y O. Etzioni. 2012. Open language learning for information extraction. *2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, (págs. 523-534).

Zhila, A. y A. Gelbukh. 2013. Comparison of open information extraction for English and Spanish. *Computational Linguistics and Intelligent Technologies*, *12, number 19*, págs. 714--722.

Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 141-152 recibido 13-12-2020 revisado 24-01-2021 aceptado 04-02-2021

Computational Reproducibility of Named Entity Recognition methods in the biomedical domain

Reproducción computacional de métodos de reconocimiento de entidades nombradas en un dominio biomédico

Ana Garcia-Serrano,[1] Sebastian Hennig[2] and Andreas Nürnberger[2]
[1] ETSI Informatica - UNED
[2] Computer science Department - OVGU
agarcia@lsi.uned.es, sebastian.hennig@st.ovgu.de, andreas.nuernberger@ovgu.de

Abstract: Unsupervised Named Entity Recognition (NER) approaches do not depend on labelled data to function properly but rather on a source of knowledge, in which promising candidates can be looked up to find the corresponding concept. In the biomedical domain knowledge source like this already exists; namely the Unified Medical Language System (UMLS). In this paper, three different unsupervised NER models using UMLS, namely MetaMap, cTakes and MetaMapLite are evaluated and compared from the results published by Demner-Fushman, Rogers and Aronson (2017) and Reategui and Ratte (2018). The Unsupervised Biomedical Named Entity Recognition framework (UB-NER) is developed, with which the results of the experiments of the three models, five datasets and two NER tasks are presented.
Keywords: Named Entity Recognition (NER), Biomedical, supervised and unsupervised models, Unified Medical Language System.

Resumen: Los enfoques para reconocimiento de entidades nombradas no supervisados (NER, por sus siglas en inglés) no dependen de corpus con datos etiquetados, sino de una fuente de conocimiento donde buscar candidatos prometedores para encontrar el concepto correspondiente. En el ámbito biomédico existe la fuente denominada "Sistema Unificado de Lenguaje Médico" (UMLS, por sus siglas en inglés). En este artículo, se evalúan y comparan tres modelos diferentes de NER no supervisados que utilizan UMLS, a saber, MetaMap, cTakes y MetaMapLite, a partir de los resultados publicados por Demner-Fushman, Rogers y Aronson (2017) y Reategui y Ratte (2018). Para ello se desarrolla el entorno *Unsupervised Biomedical Named Entity Recognition* (UB-NER), con el que se presentan resultados de los experimentos en los modelos, cinco datasets y dos tareas NER.
Palabras clave: Reconocimiento de Entidades Nombradas (NER), Modelos biomédicos, supervisados y no supervisados, Sistema de Lenguaje Médico Unificado.

1 Introduction

The task of automated detection and the correct mapping of entities to a concept is called Named Entity Recognition (NER). Unsupervised approaches do not depend on labelled data but rather on a source of knowledge in which candidates can be looked up to find the corresponding concept. In the biomedical domain this knowledge source exists, the metathesaurus *Unified Medical Language System* (UMLS)[1], a metathesaurus in which the concepts have an associated *Concept Unique Identifier* (CUI). Three different unsupervised NER models using UMLS, namely MetaMap (Aronson, 2001), cTakes (Savova, 2010) and MetaMapLite are replicated and compared in this paper.

This research work follows the NISO Standard[2] recommendations subscribed to by

[1] https://www.nlm.nih.gov/research/umls/index.html
[2]https://www.niso.org/standards-committees/reproducibility-badging

ISSN 1989-7553. DOI 10.26342/2021-66-12

the ACM[3] to reproduce the three models in the developed framework called the *Unsupervised Biomedical Named Entity Recognition framework* (UB-NER), whose objective is to find the same results as the experiments published by Reategui and Ratte (2018) and Demner-Fushman, et al. (2017).

A section is included in the following with related work, as well as a section which describes the developed framework. Section four is devoted to the setting and description of the two kinds of experiments. A comparison of the results and some considerations on reproducibility are given when some of the configuration details are missing, unknown software versions, external resources which are no longer available or when other difficulties arise.

2 Related work

For the literature review on NER methods in the biomedical domain it can be discriminated between supervised, unsupervised and hybrid approaches (Table 1). Supervised models rely heavily on data as opposed to unsupervised models. Hence supervised approaches rely on the quality of the data and how well they represent the reality. The data needs to be labelled so that supervised models can use it for training, meaning that the model fits parameters to the underlying distribution of the data. However, the acquisition of data can usually be offset by an increased performance in contrast to unsupervised models.

Properties	**Sup.**	**UnS.**
Need for labeled data	yes	no
Domain independent	no	yes
Knowledge Source	no	yes
Arbitrary filtering of sem. types	no	yes
Restricted filtering of sem. types	yes	yes
Recognize entities	yes	yes
Metaconcept recognition	no	yes
Better accord. quality measures	yes	no
Explainability	some	yes

Table 1: Features of supervised versus unsupervised NER approaches in the biomedical domain.

Recent supervised approaches adapt the state-of-the-art approaches of neighbouring fields to the biomedical domain, giving rise to high quality NER models. For example, Lee et al. introduced BioBERT (Lee et al., 2020), a variation of the standard BERT (Devlin et al., 2019) model. The default model is additionally trained on PubMed abstracts and PubMed Central full-text articles, to fit the model to the biomedical vocabulary.

The resulting BioBERT model can solve different tasks such as NER, relationship extraction and question answering. The authors establish a new state-of-the-art performance in all three tasks. Furthermore, Cho et al. (2020) used an LSTM-CRF (Lample, 2016), to generate the embedding that is fed into the LSTM-CRF, each token goes through a bi-directional LSTM character embedding and a convolutional neural network character embedding. Instead of using the standard LSTM-CRF, the authors have inserted an attention layer between the LSTM output and the CRF, which enables the CRF to attend to the relevant parts of a sequence and put less weight on the features deemed irrelevant.

(Yu et al. 2020) published a *Generative Adversarial Network* (GAN) combined with an active learning approach, to utilize unlabelled data for training. This approach finds the different semantic types of mentions in the entity. Supervised approaches perform better in general by considering measures of quality such as precision, recall and the f1-score compared to unsupervised approaches. However, the supervised approaches rely heavily on the dataset for both the coverage of domains and the semantic filtering of the mentions.

An NER tool is considered as hybrid if it is a mixture of supervised and unsupervised methods. Supervised models may have some steps based on unsupervised methods (or vice versa), thus the model is considered hybrid. Some approximations are provided below, and some functionalities are named to show their hybrid approach. Gimli (Campos, Matos and Oliveira, 2013) is a combination of dictionary consultation and pre-processing steps usually used for unsupervised models. They use a linguistic processing tool called GDep (Sagae and Tsujii, 2007) to carry out tokenization, lemmatization, POS tagging, chunking and dependency parsing. The entities found in the dictionary consultation process are not the final output as in unsupervised settings, but rather serve as an additional feature for multiple CRF models. Another hybrid approach (Bhasuran et al. 2016) extended the CRF model

[3] https://www.acm.org/publications/policies/artifact-review-and-badging-current

and uses fuzzy matching to find rare concepts in a self-made dictionary. Instead of using one CRF model in a forward chain, they also employ a CRF model in a backward chain which reads the input sequence in reverse order. Finally in (Lara-Clares and Garcia-Serrano, 2019) a Few-Shot Learning approach is described for NER on a hybrid Bi-LSTM and Convolutional Neural Network model with four input layers to recognize multi-word entities improving precision by nearly 10%, with the addition of Wikidata entities in the vocabulary.

3 Developed Platform

This section explains the UB-NER[4] java-framework developed (Hennig, 2020). One main contribution is to bring together datasets and models and comparison functionalities in quality measures of NER experiments, to simplify its access and processing of further researchers. The second main contribution is the computational reproducibility of the most used unsupervised NER approaches (MetaMap, MetaMap Lite and cTAKES) and compare it was in (Demner-Fushman, Rogers, and Aronson, 2017) and (Reategui and Ratté, 2018), so we not include any detailed description of these approaches.

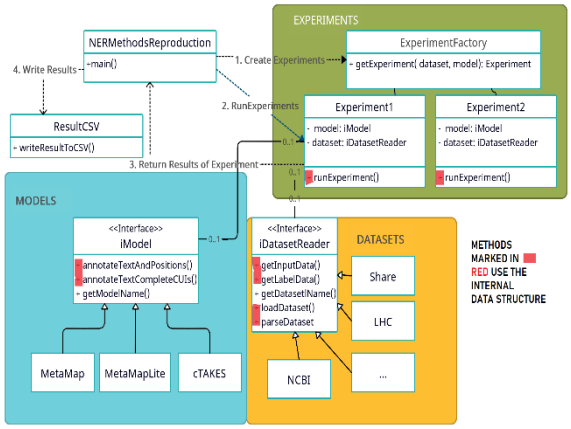

Figure 1: UB-NER High Level View. The dotted lines display the order of implementation.

All three models provide publicly available java APIs, thus facilitating the implementation of UB-NER that supports 5 datasets and 2 different NER tasks.

UB-NER consists of four components: the models, the datasets, the experiments, and the internal data structure (see Figure 1). The *annotateTextAndPositions* method first solves the NER task, giving the specific start and end position as a character offset throughout the information on the entity. The *annotateTextCompleteCUIs*, just gives a set of all entities found in the input text without any positional information. The output of the *annotateTextAndPositions* produces triplets with (start offset, end offset, concept name/ CUI). For example *'The patient has hyperlipidemia and is known to have dementia as previously stated.*' is parsed:

annotateTextAndPositions →
{(16, 30, hyperlipidemia), (45, 53, dementia)}
or {(16, 30, C0020473), (45, 53, C0497327)}
annotateTextCompleteCUIs → {C0020473, C0497327}

in which 'C0020473' is the CUI for 'hyperlipidemia' and 'C0497327' the CUI of the concept 'dementia' according to the UMLS.

Each dataset needs to implement the data set reader interface. After reading and parsing the data files, both the input and the labels can be accessed from a uniform structure. There is no pre-processing included in UB-NER because all the models implemented so far carry out the pre-processing as part of their process.

	UMLS	cTakes	Meta Map	M. M. Lite
First Experiment	2016AA	3.2.2	2016 Release	3.0
Second Experiment	2018AA[5]	3.2.0	2015 Release	-
UB-NER	2020AA	4.0.0	2020 R.	3.6.2rc5

Table 2: Versions of the UMLS and the models used in UB-NER.

A *UB-NER Experiment* is an instance of one model and one dataset, built with the *Experiment Factory*. The latest UMLS and model versions were chosen (table 2) because the two different experiments presented used different versions. Implementing a reproducible framework that automatically switches between versions would be unsuitable for the scope of this work (deviations induced by the different versions are covered in the following sections).

Apart from the semantic groups which are defined for each experiment in the configuration subsection, there are no additional configurations for MetaMap. The only

[4] Implementation technical details and reproducibility process are detailed in (Hennig and Garcia-Serrano, 2020).

[5] The UMLS version used in the original experiment is not mentioned.

additional configuration for MetaMap Lite is the segmentation method, which is set to LINES (reading each line separately instead of the complete text) for the i2b2 2010, ShARe and i2b2 2008 datasets. For the other datasets the segmentation method is not set and hence the default is used.

The *TokenProcessingPipeline* and the *FastDictionaryLookup* is used for cTAKES. Furthermore, the outputs of cTAKES are filtered to only return matches that are part of the semantic class *DiseaseDisorderMention*, since both experiments and all five datasets only contain disease and disorder references.

Although all three models contain a functionality that supports negation detection, it is not used in UB-NER, since our main goal is to reproduce the results published previously and neither of them used negation detections. However, negation can be activated by configuring the models accordingly.

The reproducible protocol is published at (Hennig, Garcia-Serrano, 2020). The framework developed is as light-weight as possible and extensible with new datasets and models following the experimental line of research established in works (Lastra and Garcia-Serrano, 2015a and b) or (Benavent et al., 2010).

4 Experiments and Evaluation

This paper's one main goal is to reproduce the two sets of NER experiments, the published by Demner-Fushman, Rogers and Aronson (2017) and the reported by Reategui and Ratte (2018). In the former, the outputs contain the name and the start and end position of each entity found. They are then compared to the gold standard and the precision, recall and f1-score are computed.

The latter experiments collate all of the entities found in a document. The entity list returned as a result is compared to a set of reference labels to locate all relevant matches. If a match is found, the candidate is added to a final output set, which is compared to the annotated gold-standard label set (subset of the reference label) and then the precision, recall and f1-score are computed.

In UB-NER each annotated concept is stored with its positional information as *AtomStringLabel*. A text usually contains more than one medical concept; hence we need a data structure to save all annotations that appear. So, each ground truth and each output consists of a set of *AtomStringLabels* and these can be compared to each other. Let L be the ground truth labels and M the labels suggested by the NER model, then

- $I = L \cap M$
- $OL = L \setminus M$
- $OM = M \setminus L$

where I is the intersection, OL are the concepts that only appear in ground truth labels and OM are the concepts that only appear in the output. These three sets can now be used to compute the *set of retrieved documents* (as $I \cup OM = M$) and the set of *relevant documents* (as $I \cup OL = L$) which are needed to calculate the precision and recall as can be seen in the following formulas. So, the calculation of OM and OL could be omitted and M and L could be used to get the retrieved and relevant document sets. In MetaMap Lite implementation OL and OM are employed for the evaluation, subsequently the precision and recall are calculated using the following formulas.

$$recall = \frac{\sum_{d \in D} |I|}{\sum_{d \in D} |\{retrieved\ documents\}|} \quad (1)$$

$$recall = \frac{\sum_{d \in D} |I|}{\sum_{d \in D} |\{relevant\ documents\}|} \quad (2)$$

$$F1 = 2\frac{precision\ .\ recall}{precision + recall} \quad (3)$$

To obtain the overall performance on a document's dataset D we do not compute the precision and recall of each document $d \in D$ and take the average, but rather accumulate all intersection set sizes and all retrieved and relevant set sizes.

In the second experiment, the multi-label classification problem, let Y be the set of all classes. Usually for a multi-label classification problem, a binary vector of size $|Y|$ for each document of D is defined, which indicates its classes. However due to the modality of the experiments, an alternate representation is used instead. For each class $y \in Y$ there is a set $L_y \subset D$, so that every $d \in L_y$ is an instance of class y. The set of leftover documents which are not an instance of y will be referred to as A_y. So, for each class $y \in Y$ there exists L and A, so that $L \subset D$, $A \subset D$ and $L \cup A = D$. We use L and A for the sets that represent the ground truth labels. Similarly, there is a set LM_y, containing all of the documents that the model predicts to be an instance of y. In the case of NER, a model

predicts a class *y* for a document *d* when an entity that is associated with *y* appears in *d*.

Let's assume MetaMap Lite is the model and we currently want to find LM for the class *Asthma*. Then we process each document $d \in D$ with MetaMap Lite. If MetaMap Lite recognizes a concept in d and assigns the CUI *C0004096* for *Asthma* to it, then d will be added to LM_{Asthma}. Thus, AM_{Asthma} is the set containing all documents in which *Asthma* is not part of the concepts detected by MetaMap Lite. At the same time, there is an LM and AM for each class $y \in Y$ where $LM \subset D$, $AM \subset D$ and $LM \cup AM = D$. Using these sets we can define the true positives (TP, entities recognized by the system that are also present in the ground truth), false negatives (FN, entities recognized that are not present in the ground truth) and false positives (FP, entities not recognized but present in the ground truth) for each class: TP = $L \cap LM$; FN = $L \setminus LM$ and FP = $LM \setminus L$. This leads to the calculation of the final score:

$$precision = \frac{|TP|}{|TP|+|FP|} \quad (4)$$

$$recall = \frac{|TP|}{|TP|+|FN|} \quad (5)$$

$$F1 = 2\frac{precision*recall}{precision+recall} \quad (6)$$

4.1 First experiments: Exact position

The four corpuses used in the exact position experiments are:

The NCBI Disease Corpus (Dogan, Leaman and Lu, 2014) consists of annotated titles and abstracts from 793 PubMed articles, annotated with MeSH and OMIM concept identifiers. As these identifiers are part of the UMLS, they can be mapped to CUIs.

Lister Hill Center (LHC) test collection is a mixture of annotated PubMed abstracts in which 150 are clinically oriented and another 150 are biology oriented. A total of 2,242 disorders are annotated and normalized to their UMLS CUIs. There exists a version of NCBI, which is also belongs to the LHC collection, that contains additional manual annotations.

The i2b2 2010 is a collection of 871 clinical notes, which provides various annotations. In this work we ignore the treatment and test annotations, following the MetaMap Lite author's evaluation strategy.

ShARe corpus contains 300 clinical notes, annotated with disorder references and normalized to a CUI if possible.

All datasets are in text-form and for each document there is a file with the text and a file with the corresponding annotated entities. The authors of the original experiments parsed the labels to brat standoff format[6] and the CUIs are omitted as the preferred concept names, the human readable identifier in UMLS, are used for comparison as they can be interchanged. They compare the concept name as well as the start and the end positions in the text. In this work the labels are not parsed to the brat standoff format, but the concept name and offsets are equally compared using the *AtomStringLabel* format.

A typical label could look like *"lung cancer | 14 | 25*", in which lung cancer is the preferred name, 14 is the number of the starting character and 25 the ending one. The character offsets are all relative to the first character of the document. Each label has a semantic type assigned to it and, following the work to be reproduced, we only consider labels that fall under one of the semantic types *Disorder* or *GeneralDisorder*. The main reasons for this choice are that the datasets and tools are heavily skewed toward these semantic types and also because of their importance in clinical text processing and downstream applications, such as the extraction of phenotypes or adverse reactions to drugs (Segura-Bedmar and Martínez, 2017).

As mentioned before, the concept names are used for the gold-standard labels instead of the CUIs. Therefore, in this work we need the output of the models to be a concept name, too (to make then iniformal). Each of the three models MetaMap, MetaMap Lite and cTakes, can output the multiple formats of a found concept. Namely the CUI, the preferred concept name as saved in the dictionary (UMLS) and the concept name found in the text. We decided to use the concept name found in the text, which is also used in the ground-truth labels. Using the preferred concept names as defined in a dictionary, would lead to problems in assigning correct offsets in the model outputs as well as in the labels, since the length of the dictionary entry can vary from the length of the corresponding phrase found in the text.

Although it is not mentioned in the original paper, but directly influences the results, the

[6] http://brat.nlplab.org/standoff.html

MetaMap and MetaMap Lite output is restricted to a list of semantic types. The nine semantic types mentioned in the code kindly provided by the authors of MetaMap Lite are: *Acquired Abnormality (acab), Congenital Abnormality (cgab), Injury or Poisoning (inpo), Pathologic Function (patf), Disease or Syndrome (dsyn), Anatomical Abnormality (anab), Neoplastic Process (neop), Mental or Behavioral Dysfunction (mobd), Sign or Symptom (sosy)*. Restricting the output to these semantic types increases the precision of MetaMap and MetaMap Lite, since any entities found that are not annotated in the gold standard as disease, e.g. entities of the semantic type plant, are discarded.

4.2 Second experiments: Classification

The authors (Reategui and Ratte, 2018) ran two experiments identifying whether a comorbidity[7] is present or not in a discharge summary. They differ in the labels used. In the first experiment (Single CUI experiment), single UMLS concepts are assigned to each comorbidity which should be predicted by the models. For the Multiple CUIs experiment, additional UMLS concepts are added to some of the comorbidities thus forming an aggregation of CUIs. This task is easier since the models only need to find one of the CUIs mentioned in a concept aggregate of a comorbidity to get a successful match.

The i2b2 2008 obesity challenge dataset used in this experiment (Uzuner, 2009) contains 1,237 medical discharge summaries of obese and diabetic people. It is annotated with 15 possible comorbidities of obesity. The labels indicated for each comorbidity in the underlying medical record are: **present** (the patient has/had the disease); **absent** (the patient does/did not have the disease); **questionable** (the patient may have the disease) and **unmentioned** (the disease is not mentioned in the discharge summary).

Aiming exactly at reproducing the results of the authors, we selected the subset of 412 summaries which had obesity as a comorbidity and the annotated gold standard was taken and changed into a binary classification task. We discriminate between present and absent, where a comorbidity is present if and only if it is tagged as present in the gold standard. If it is tagged as either absent, questionable, or unmentioned we consider it as absent. With this new binary presentation two sets of documents can be created for each comorbidity, namely *L* and *A*, as mentioned in section 4. In the original experiments of Reategui and Ratte (2018), *D* corresponds to the set of all 412 obesity discharge summaries, and the set of classes of comorbidities (*Y)* considered are: *Pathologic Function (patf)*; *Disease or Syndrome (dsyn); Therapeutic* or *Preventive Procedure (topp); Mental or Behavioral Dysfunction (mobd)*. We refer to the original publication for explanations on the aggregation process and the reasoning behind the choices for the aggregations used in the second experiment.

The two experiments are carried out using the precision and recall calculations stated in section 4. The only difference is the creation of the *LM* sets for the Multiple CUIs experiment. In the Single CUI experiment, a summary is only part of $LM_{\text{Depression}}$, if the model detects an entity with CUI *C0011570* in it. For the Multiple CUI experiment it is enough for a summary to be included in $LM_{\text{Depression}}$, if the model manages to detect either the concept *C0011570* or *C0011581*.

Since the two classification experiments described are different from the first one based on the work of the MetaMap Lite, we have created an additional *ExperimentCompleteDoc* class in UB-NER, where instead of looking at each concept found separately, we create a list of all concepts found. The resulting list is checked against the 14 available comorbidities considered (*Hypertriglyceridemia* was excluded due to a lack of sufficient examples). If a comorbidity is found in the document, it is added to the *LM* set.

Changes were also needed in the dataset loading. Instead of creating *AtomStringLabels* for each document, we assigned a document to the *L* set if a comorbidity was annotated as present in the gold standard. The scores are then computed after all *L* and *LM* sets are calculated.

5 Results Comparison

In this section the results obtained by UB-NER reproducing the two original experiments are compared with the results published. Furthermore, the delta between the two evaluations is calculated by subtracting the original from UB-NER score. Hence a positive

[7] Comorbidity refers to the presence of more than one disorder (co-existing) in the same person.

entry means that our model is better than the original and a negative means that is not.

5.1 Exact Position Experiments Results

We found similar results (table 3) as published in the original article as set out in table 4. MetaMap Lite outperforms the others in terms of precision, recall and f1 and only on the ShARe dataset, MetaMap marginally beat MetaMap Lite.

In the original experiment the *AggregatePlaintextUMLSProcessor* were used in the cTAKES pipeline. Unfortunately, we could not run it since it took more time to process one datapoint, than it took MetaMap Lite to process the complete dataset. Hence, we used the fast pipeline provided by cTAKES.

Datasets	MetaMap		
	P	R	F1
LHC NCBI	0.546	0.583	0.564
LHC-Bio Cits	0.396	0.608	0.479
ShARe	0.444	0.662	**0.532**
LHC-Clin Cits	0.561	0.635	0.596
i2b2 2010	0.364	0.347	0.355

Datasets	MetaMapLite		
	P	R	F1
LHC NCBI	0.664	0.714	**0.688**
LHC-Bio Cits	0.468	0.711	**0.564**
ShARe	0.483	0.585	0.529
LHC-Clin Cits	0.635	0.711	**0.671**
i2b2 2010	0.395	0.349	**0.371**

Datasets	cTAKES		
	P	R	F1
LHC NCBI	0.483	0.607	0.538
LHC-Bio Cits	0.443	0.549	0.490
ShARe	0.464	0.417	0.440
LHC-Clin Cits	0.517	0.549	0.533
i2b2 2010	0.315	0.202	0.246

Table 3 (a), (b), (c) UB-NER results on the exact position experiments.

Most of the deviations detailed in table 4 (on average our scores are 0.035 worse than originals) can be explained by the differences in comparing the model output to the label set. We parsed the labels and the model output to *AtomStringLabel*, whereas in the original experiments the brat standoff format was used.

There are some cases in which the output of UB-NER identifies the positions correctly, but the entity name does not exactly match the label name. For example, "524 555 glucose/galactose malabsorption" is the output and "524 555 glucose malabsorption" is the gold standard. Parsing the model output to the brat standoff, changes those cases to be mapped correctly. However, only 0.34% of all labels are affected.

	METAMAP		
DATASETS	P	R	F1
LHC NCBI	-0.057	-0.1	-0.077
LHC-BIO CITS	-0.072	-0.148	-0.099
SHARE	-0.151	0.181	0
LHC-CLIN CITS	-0.027	-0.137	-0.072
I2B2 2010	-0.017	-0.01	-0.013

	METAMAP LITE		
DATASETS	P	R	F1
LHC NCBI	-0.067	-0.005	-0.037
LHC-BIO CITS	-0.207	-0.068	-0.16
SHARE	-0.259	0.164	-0.009
LHC-CLIN CITS	-0.059	-0.038	-0.029
I2B2 2010	-0.075	0.03	-0.009

	CTAKES		
DATASETS	P	R	F1
LHC NCBI	0.013	0.069	0
LHC-BIO CITS	-0.028	-0.057	-0.04
SHARE	0.001	-0.045	-0.022
LHC-CLIN CITS	0.091	-0.05	0.035
I2B2 2010	-0.004	-0.139	-0.083

Table 4 (a), (b), (c): Delta to Original Results.

MetaMap and MetaMap Lite differ in precision and recall on the ShARe dataset, but in such proportions that they offset each other and the f1 score stays the same. The ShARe dataset does not have the name of the entity as a label, but instead each entity is tagged with its CUI. MetaMap Lite converted those CUI labels to the brat standoff format. In UB-NER the outputs of the models were adapted, mapping the entity to the corresponding CUI, allowing the output to be matched against the gold labels given by the ShARe dataset, containing positional information and the CUI.

The differences of MetaMap for the LHC-Bio Cits and LHC-Clin Cits are induced by the aggregation of variations from the original experiment. In addition to the differences between the brat standoff and the *AtomStringLabel*, the output of MetaMap is also different. The original experiment uses the fielded MetaMap (MMI) output. Unfortunately, the MetaMap API does not support this output format. We approximate the MMI output as closely as possible with the API available tools.

However, there are limits which cannot be easily overcome. For example, the phrase "*transposition of the great vessels*" is recognized

as *Transposition of Great Vessels* and CUI *C0040761* when the fielded MMI output is used. When the API is used, two independent concepts, namely *Transposition* with CUI *C0040759* and *Great vessels* with CUI *C0225991* are returned by MetaMap from which the latter is removed from the output since the semantic type of *Great vessels* is not part of the list used for the experiment.

Therefore, without adapting the semantic types, it is not possible to get the same output with the API as the console version with a fielded MMI output. Both output forms identify abbreviations but only the fielded MMI output returns the short form mentioned in the text, which is also the one used in the labels most of the time. The API on the other hand, only returns the full name of the corresponding concept instead of the abbreviation. A heuristic is implemented in UB-NER to map those complete matches back to the abbreviations found but is unable to produce the same output as the fielded MMI.

Changes in the UMLS versions also causes some entities, that were previously found, to no longer be recognized. For example the concept *HIV* is part of the semantic group *Disease or Syndrome (dysn)* in former UMLS versions, while the current version of the UMLS used in UB-NER maps HIV to the semantic group *Virus (virs)* which is not part of that list.

Theoretically these problems are present in all datasets processed by our implementation of MetaMap. The greater influence of these factors on the LHC-Bio Cits and LHC-Clin Cits among others stems from the fact that these datasets are relatively small compared to the other, and hence single errors have a greater impact on the overall score. The deviation of the precision of MetaMap Lite on the LHC-Bio Cits is because MetaMap Lite was able to recognize many more abbreviations with the UMLS 2020AA than with older versions.

Unfortunately, the texts in the LHC-Bio Cits dataset contain a lot of abbreviations for phrases that are not diseases. For example, the phrase "*Corticotropin-releasing factor (CRF)",* where the abbreviation *CRF* is used for all other occurrences in the text, is identified by MetaMap Lite with the UMLS 2020AA. Even though the correct concept *C0772289* belonging to this phrase, can be found by MetaMap Lite, the resulting semantic type for this match is not contained in the list of accepted semantic types. This would result in the *CRF not being matched*. Unfortunately, the abbreviation CRF is also used for the concept *Cancer-related fatigue* with CUI *C4274302*. Hence MetaMap Lite outputs a wrong interpretation of CRF.

Naturally biological abstracts contained in the LHC-Bio Cits dataset, also contain abbreviations for biological phrases, and some concepts are mapped to the same abbreviation, even though they are completely uncorrelated. These false positives, who's weighting to the total score is enhanced by the fact that an abbreviation is frequently used, results in a lower precision. So, while it is a good idea to include abbreviations to increase recall, it can decrease the precision disproportionately.

5.2 Classification Experiments Results

The best model for each comorbidity shows that MetaMap Lite cannot outperform MetaMap and cTakes, in contrast to results in previous section, but it can match their performance.

Precision and recall of this task are higher than in the exact position because: (1) No position tagging is required; (2) The task is aligned with the dataset: nearly all biomedical entity mentions belong to one of the 14 target concepts; and (3) The entities are not verified one by one but count as a match if the entity appears at least once in the document.

In general, our UB-NER results (tables 5, 6 and 7) match closely with the results of the original work and differences can be attributed to the use of different versions of UMLS and that: (1) no configuration details of MetaMap are given, (2) neither are cTAKES and (3) neither was the UMLS version mentioned for MetaMap nor cTAKES.

	MetaMap					
	Single CUI Exp.			Multiple CUI Exp.		
Entity	P	R	F1	P	R	F1
CHF	0.864	0.927	**0.895**	0.864	0.921	0.891
Hypertension	0.95	0.97	0.96	0.949	0.967	0.958
Venous Insufficiency	1	0.316	0.48	1	0.316	0.48
Gout	0.945	0.963	0.954	0.945	0.963	0.954
CAD	0.839	0.672	0.746	0.821	0.688	0.749
Gallstones	0.982	0.7	0.818	0.97	0.8	0.877
Depression	0.932	0.932	**0.932**	0.931	0.92	**0.926**
Asthma	0.885	0.906	0.895	0.884	0.894	0.889
GERD	0.911	0.947	0.929	0.911	0.947	0.929
OA	0.866	0.816	0.84	0.866	0.816	**0.84**
Hypercholesterolemia	0.935	0.571	**0.709**	0.948	0.84	**0.891**
Diabetes	0.885	0.686	0.773	0.888	0.895	0.892
OSA	0.924	0.758	**0.833**	0.923	0.75	**0.828**
PVD	0.974	0.974	**0.974**	0.974	0.974	**0.974**
Average	0.921	0.796	0.838	0.92	0.835	0.863

Table 5: Results for MetaMap Experiments.

MetaMap Lite

	Single CUI Exp.			Multiple CUI Exp.		
Entity	P	R	F1	P	R	F1
CHF	0.873	0.915	0.893	0.873	0.915	**0.893**
Hypertension	0.95	0.976	0.963	0.95	0.976	0.963
Venous Insufficiency	1	0.316	0.48	1	0.316	0.48
Gout	0.944	0.944	0.944	0.944	0.944	0.944
CAD	0.826	0.892	0.858	0.81	0.892	0.849
Gallstones	0.979	0.588	0.734	0.966	0.7	0.812
Depression	0.761	0.943	0.843	0.761	0.943	0.843
Asthma	0.892	0.871	0.881	0.892	0.871	0.881
GERD	0.913	0.961	**0.936**	0.913	0.961	**0.936**
OA	0.897	0.598	0.717	0.897	0.598	0.717
Hypercholes-terolemia	0.949	0.537	0.686	0.959	0.811	0.879
Diabetes	0.886	0.725	0.797	0.885	0.899	0.892
OSA	0.919	0.711	0.802	0.919	0.711	0.802
PVD	0.97	0.842	0.901	0.97	0.842	0.901
Average	0.911	0.772	0.817	0.91	0.813	0.842

Table 6: MetaMap Lite Experiments Results.

cTAKES

	Single CUI Exp.			Multiple CUI Exp.		
Entity	P	R	F1	P	R	F1
CHF	0.924	0.661	0.77	0.924	0.661	0.77
Hypertension	0.943	0.991	**0.966**	0.943	0.991	**0.966**
Venous Insufficiency	1	0.316	0.48	0.633	1	**0.776**
Gout	0.945	0.963	0.954	0.945	0.963	0.954
CAD	0.828	0.903	**0.864**	0.828	0.903	**0.864**
Gallstones	0.984	0.788	0.875	0.959	0.888	**0.922**
Depression	0	0	0	0.719	0.989	0.833
Asthma	0.867	1	**0.929**	0.867	1	**0.929**
GERD	0.862	0.987	0.92	0.862	0.987	0.92
OA	0.906	0.667	0.768	0.906	0.667	0.768
Hypercholes-terolemia	0.941	0.549	0.693	0.954	0.829	0.887
Diabetes	0.888	0.826	**0.855**	0.877	0.915	**0.896**
OSA	0.921	0.727	0.812	0.921	0.727	0.812
PVD	0.969	0.816	0.886	0.969	0.816	0.886
Average	0.856	0.728	0.769	0.879	0.881	0.87

Table 7: UB-NER results for cTAKES experiments.

The greatest discrepancy in table 8 is the single CUI *Depression* which is mapped to the CUI *C0011570*. cTAKES maps all occurrences of *C0011581*. In the multiple CUIs experiment in which the CUI *C0011581* is added, the results match up again for cTAKES and a marginal improvement in precision is achieved.

The MetaMap implementation used in UB-NER also gave rise to better precision results in the single CUI and multiple CUIs experiment. In the literature, *Depression* is hard to recognize correctly, because usually refers to a mental disorder, but in the biomedical domain it can also refer to a "reduction".

There is also a significant difference for *Atherosclerotic Cardiovascular Disease (CAD)* in the single CUI experiment, whereas the difference in the multiple CUIs experiment is negligible. In former UMLS versions, instances of CAD were solely mapped to the concept *Coronary arteriosclerosis*, CUI *C0010054*, however in the current version it has a new one, the *Coronary artery disease* with CUI *C1956346*. The CUI mapping table shows that *C1956346* was used in the Single and *C0010054* was added for the Multiple experiment.

	MetaMap			cTAKES		
Entity	P	R	F1	P	R	F1
CHF	-0.006	0.037	0.015	0.064	**-0.259**	**-0.12**
Hypertension	-0.01	-0.02	-0.02	-0.007	0.001	-0.004
Venous Insufficiency	0	0.026	0.04	0	0.026	0.04
Gout	-0.005	-0.017	-0.006	-0.005	-0.017	-0.006
CAD	-0.021	**0.222**	**0.156**	-0.012	-0.017	-0.016
Gallstones	-0.018	-0.03	-0.022	-0.016	0.008	-0.005
Depression	**0.232**	0.042	**0.142**	**-0.71**	**-0.99**	**-0.82**
Asthma	-0.015	-0.024	-0.015	-0.013	0	-0.011
GERD	0.021	-0.023	-0.001	-0.018	-0.003	-0.01
OA	-0.004	0.056	0.03	-0.044	-0.003	-0.012
Hypercholes-terolemia	-0.005	-0.019	-0.021	-0.009	0.039	0.033
Diabetes	-0.025	0.036	0.013	-0.022	-0.004	-0.015
OSA	-0.016	-0.022	-0.017	-0.019	-0.033	-0.028
PVD	0.004	0.004	0.004	-0.001	-0.024	-0.014

Table 8: Delta Single CUI.

	MetaMap			cTAKES		
Entity	P	R	F1	P	R	F1
CHF	-0.006	0.031	0.011	0.064	**-0.259**	**-0.12**
Hypertension	-0.011	-0.023	-0.022	-0.007	0.001	-0.004
Venous Insufficiency	**0.296**	**-0.589**	**-0.312**	-0.067	0	-0.048
Gout	-0.005	-0.017	-0.006	-0.005	-0.017	-0.006
CAD	-0.009	0.088	0.059	0.018	-0.017	-0.006
Gallstones	-0.02	-0.065	-0.043	-0.011	-0.002	-0.008
Depression	**0.225**	-0.01	**0.124**	0.009	-0.001	0.013
Asthma	-0.016	-0.036	-0.021	-0.013	0	-0.011
GERD	0.021	-0.023	-0.001	-0.018	-0.003	-0.01
OA	-0.004	0.056	0.03	-0.044	-0.003	-0.012
Hypercholes-terolemia	-0.012	-0.04	-0.029	-0.006	0.019	0.007
Diabetes	-0.022	0.005	-0.008	-0.013	-0.005	-0.014
OSA	-0.017	-0.03	-0.022	-0.019	-0.033	-0.028
PVD	0.004	0.004	0.004	-0.001	-0.024	-0.014

Table 9: Delta Multiple CUIs.

Table 9 shows that *Venous Insufficiency* has significant differences for the multiple CUIs experiment. This stems from the addition of the concept *Postthrombotic syndrome* with CUI *C0277919*. In the former versions of the UMLS, *venous stasis* is mapped to the CUI *C0277919*, which explains the improved performance in the original. In the 2020AA version of the UMLS a new concept for *venous stasis* was introduced with CUI *C441518*. Hence all instances that were previously mapped to *C0277919* are now mapped to *C441518*. If we

substituted the *C0277919* with *C441518*, we would likely get the same results.

The discrepancy in the single CUI and multiple CUIs experiment for *CHF* in cTAKES is also brought about by software versions. In the current one, like *Venous insufficiency*, instances that can be mapped to more specific concepts are no longer mapped to the general concept *C0018802*, resulting in a lower recall. It is necessary to have notice that even if the three models were processed by the UB-NER for this experiment and results explained, MetaMap Lite it is not shown in delta tables 7 and 8 because it was not included in the comparison of the original work in (Reategui and Ratté, 2018, thus it is not possible to calculate any delta for MetaMap Lite.

5.3 Computational reproducibility

UB-NER was able to reproduce the results published in (Demner-Fushman, Rogers, and Aronson, 2017) and (Reategui and Ratte, 2018) with no significant differences according to the student's t-test, proving the published findings as reproducible and correct.

For the student's t-test, the p-value is computed by using a two-sided t-distribution on two paired sample sets. Our null hypothesis H_0 states that the average performance of the compared implementations is equal, whilst the alternative hypothesis states that their average performance is different. We choose a 5% significance level and say that the performance differs significantly if we must reject H_0 i.e. the p-value is smaller than 0.05. On the other hand, if the p-value is larger or equal to 0.05 H_0 holds and the differences in performance are considered insignificant.

If we take the values from table 3 and the original results from in (Demner-Fushman, Rogers, and Aronson, 2017) the calculation of the p-value yields 0.198 and hence shows that our results are not significantly different from the original results. Analogous the p-value for the Single CUI experiment (table 5) is 0.199 and 0.373 for the Multiple CUI experiment (table 6) respectively, indicating that the differences to the results published in (Reategui and Ratte, 2018) are also not significant.

6 Conclusions

The two NER in the biomedical domain widely used are the following unsupervised models: MetaMap with just a few supervised parts in its pipeline (the POS-tagger) and cTAKES which has more pre-trained supervised parts in its pipeline. Both use the UMLS to identify and extract medical entities from text and were compared in (Reategui and Ratté, 2018) showing very similar behaviour using the i2b2 2008 dataset. In (Demner-Fushman, Rogers, and Aronson, 2017) MetaMap and cTAKES were compared with MetaMap Lite, a Java implementation of MetaMap focusing on real-time processing speed.

We presented the UB-NER framework to validate published results in the original comparisons of (Demner-Fushman, Rogers, and Aronson, 2017) and (Reategui and Ratté, 2018), with a discussion justifying the differences found and explaining how the different versions of UMLS, the abbreviations considered and other related features, impact on the results.

UB-NER enables the computational reproduction of scientific research results, bringing together biomedical datasets and models for NER models, so removing barriers in the dataset access and NER processing to the researchers, i.e. all models in the original papers have different input/output formats and not in UB-NER. To configure an experiment in UB-NER you only must do some database and model selection to obtain results and quality measures.

We plan to extend UB-NER to support more datasets, models, and experiments for unsupervised as well as supervised approaches. Furthermore, we want to create a novel NER method that uses a supervised approach, exploiting additional information provided by UMLS, to enhance the usability of entities found for downstream tasks.

Acknowledgements

Thanks to Juan J. Lastra-Díaz y Alicia Lara-Clarés for their initial comments. We also want to thank Dina Demner-Fushman and Willie Rogers. The feedback provided by them were really helpful.

Bibliography

Aronson, A.R. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: theMetaMap program. *Proc. AMIA Annual Symposium*, pages 17–21, ISSN 1531605X.

Benavent, J., X. Benavent, E. de Ves, R. Granados, and A. Garcia-Serrano. 2010. Experiences at ImageCLEF 2010 using CBIR and TBIR Mixing Information Approaches. CLEF CEUR-WS, vol 1176.

Bhasuran, B., G. Murugesan, S. Abdulkadhar, and J. Natarajan. 2016. Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases. *Journal of Biomedical Informatics* 64 (Dec), pp. 1–9. doi: 10.1016/j.jbi. 2016.09.009.

Campos, D., S. Matos, and J. L. Oliveira. 2015. Gimli: Open source and high-performance biomedical name recognition. *BMC Bioinformatics* 14.1 Feb, p. 54. doi: 10.1186/1471-2105-14-54

Cho, M., J. Ha, C. Park, and S. Park. 2020. Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition. *Journal of Biomedical Informatics* 103 (Mar) p. 103381. doi: 10.1016/j. jbi.2020.103381.

Demner-Fushman, D., W. J. Rogers, and A. R. Aronson. 2017. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. *J. of the American Medical Informatics Association* 24.4, pp. 841–844. doi: 10.1093/jamia/ocw177.

Devlin, J., M.Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Technical report. https://github.com/tensorflow/tensor2tensor.

Dogan, R.I., R. Leaman, and Z. Lu. 2014. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10, doi: 10.1016/j.jbi.2013.12.006.

Hennig, S. 2020. An experimental survey of Named Entity Recognition methods in the biomedical domain. Master Data and Knowledge Engineering. Faculty of Computer Science. OVGU. A. Garcia-Serrano and A. Nürnberger supervisors.

Hennig, S. and A. Garcia-Serrano. 2020. Reproducible experiments on the master thesis: An experimental survey of Named Entity Recognition methods in the biomedical domain, UNED *e-cienciaDatos, V1* (dec) https://doi.org/10.21950/DYAZRE.

Lample, G., M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. 2016. Neural architectures for named entity recognition. *Proc. of NAACL HLT 2016*, pp. 260–270.

Lara-Clares, A., A. Garcia-Serrano. 2019. LSI2_UNED at eHealth-KD Challenge 2019: A Few-shot Learning Model for Knowledge Discovery from eHealth Documents. *CEUR-WS*, vol 2421, IberLEF. Bilbao, Spain.

Lastra-Díaz, J.J. and A. Garcia-Serrano. 2015a. A novel family of IC-based similarity measures with a detailed experimental survey on WordNet. *Engineering Applications of Artificial Intelligence* 46, 140-153.

Lastra-Díaz, J.J. and A. Garcia-Serrano. 2015b. A new family of information content models with an experimental survey on WordNet. *Knowledge-Based Systems* 89, 509-526.

Lee, J., W. Yoon, S. Kim, D. Kim, S. Kim, C.Ho So, and J. Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36.4 (Feb), pp. 1234–1240. doi: 10.1093/bioinformatics/ btz682.

Merkel, D. 2014. Docker: lightweight Linux containers for consistent development and deployment. https://dl.acm.org/doi/10.5555/ 2600239.2600241.

Mowery, D. 2013. ShAReCLEF eHealth Evaluation Lab 2014 (Task 2): Disorder Attributes in Clinical Reports. *PhysioNet* https://doi.org/10.13026/0zgk-9j94.

Reategui, R. and S. Ratte. 2018. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. *BMC Medical Informatics and Decision Making* 18.3, p. 74. doi: 10.1186/s12911-018-0654-2.

Sagae, K. and J. Tsujii. 2007. Dependency Parsing and Domain Adaptation with LR

Models and Parser Ensembles. In *Proc. of the EMNLP-CoNLL, 2007*, pp. 1044–1050 https://www.aclweb.org/anthology/D071111

Savova, G., 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 American Medical* Informatics Association 17(5):507–513.
DOI: 10.1136/jamia.2009.001560

Segura-Bedmar, I. and P. Martínez. 2017. Simplifying drug package leaflets written in Spanish by using word embedding. *Journal of Biomedical Semantics* 8, 45. https://doi.org/10.1186/s13326-017-0156-7

Uzuner, A. 2009. Recognizing Obesity and Comorbidities in Sparse Data. *Journal of the American Medical Informatics Association*, 16(4):561–570, 7.

Gang, Y., Y. Yang, X. Wang, H. Zhen, G. He, Z. Li, Y. Zhao, Q. Shu, and L. Shu. 2020. Adversarial active learning for the identification of medical concepts and annotation inconsistency. *Journal of Biomedical Informatics 108* (Aug), p. 103481.
https://doi.org/10.1016/j.jbi.2020.103481.

Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 153-164 recibido 23-11-2020 revisado 26-01-2021 aceptado 01-02-2021

Classifying Spanish *se* constructions: from bag of words to language models

Clasificación de construcciones con se en español: de modelos de bolsa de palabras a modelos de lenguaje

Nuria Aldama García, Álvaro Barbero Jiménez
Universidad Autónoma de Madrid
Instituto de Ingeniería del Conocimiento
nuria.aldama@estudiante.uam.es, alvaro.barbero@iic.uam.es

Abstract: Spanish *se* constructions are a complex linguistic phenomenon that challenges Natural Language Processing (NLP) tasks such as part-of-speech or dependency relation tagging. *Se* is a high-frequency word that appears in nine different types of syntactic constructions and adds information of diverse nature depending on the context. Thus, to solve the problem Spanish *se* constructions poses in an efficient way, this study proposes a tagging system for *se* applied to a corpus composed of 2,140 sentences. This corpus is used in a classification experiment where 9 classifiers based on machine learning models and a dependency parser are tested. Results show that pre-trained language models based on transformers architecture reach the highest accuracy (0.83) and f-score (0.70) values.
Keywords: Spanish *se* constructions, multiclass classification, machine learning.

Resumen: Las construcciones con *se* en español son un complejo fenómeno lingüístico que desafía tareas de Procesamiento del Lenguaje Natural (PLN) como el etiquetado automático de categoría gramatical (*POS tagging*) o de relaciones de dependencias. *Se* es una forma de alta frecuencia que aparece en nueve tipos de construcciones sintácticas del español, aportando información de diferente naturaleza en función del contexto. Por ello, para tratar el problema de clasificación que plantean las construcciones con *se* de manera eficiente, este estudio propone un sistema de etiquetado de *se* aplicado a un corpus de 2.140 oraciones y probado con 9 clasificadores basados en modelos de aprendizaje automático y un parser de dependencias. Los resultados muestran que los modelos pre-entrenados basados en arquitectura de *transformers* alcanzan los valores más elevados de exactitud (0,83) y de F-score (0,70).
Palabras clave: Construcciones con *se*, clasificación multiclase, aprendizaje automático.

1 Introduction

Spanish *se* constructions are a well-known and complex linguistic topic within the study of Spanish. *Se* constructions challenge Natural Language Processing (NLP) tasks such as automatic part-of-speech-tagging (POS) and dependency parsing for three main reasons. First, *se* is a high-frequency Spanish word. According to CORPES XXI (Real Academia Española de la Lengua, 2020), *se* is in the eleventh position of the ranking of most common grammatical elements and most common lemmas in Spanish and it is placed in the ninth position in the ranking of most common Spanish forms. Second, *se* may appear in nine different syntactic constructions where it conveys diverse semantic meanings and bears several syntactic roles (if any). Third, the form *se* does not bear any specific morphosyntactic feature that helps disambiguating one type of *se* from another.

The main goal of this study is to evaluate the performance of different classification strategies that are intended to solve the task of *se* disambiguation based on an adaptation of the analysis of *se* presented by Moreno Cabrera (1997, 2002). To do so, a corpus containing 2,140 sentences, the *SE*-corpus, is built as a means of training and evaluating nine classifiers and a state-of-the-art parser. A secondary

ISSN 1989-7553. DOI 10.26342/2021-66-13

objective is to understand the kind of information (lexicon, colocations, semantics, syntax, any other contextual information) that is needed by a machine learning model to best disambiguate Spanish *se* constructions.

The paper is structured as follows. Section 2 summarizes Spanish *se* constructions. Section 3 describes the *SE*-corpus. Section 4 presents *se* tag distribution. Section 5 deals with corpus quality. Section 6 introduces the classification strategies used in this study. Section 7 shows experimental results. Conclusions and future work are drawn in section 8.

2 Spanish se constructions

Se may appear in nine different syntactic constructions where it conveys diverse semantic meanings and bears several syntactic roles (if any). This section makes a brief theoretical review of this kind of constructions based on (Sánchez, 2002), (Sánchez, 2015), (Mendikoetxea, 1999 a), (Mendikoetxea, 1999 b), (Campos, 1999), (Fernández-Montraveta and Vázquez, 2017), (Moreno 1997) and (Moreno, 2002).

Se constructions may be classified as paradigmatic (if the concrete construction can be built with all the pronominal forms of the paradigm) or non-paradigmatic (if the concrete construction can only be built with the form *se*). Within the class of paradigmatic constructions, *se* may appear in transitive constructions like (1), (2) and (3). *Se* functions as an indirect object in (1) and (3) and it has a benefactive or recipient semantic role. In (2), *se* is the internal argument of the main predicate *comb*, it is accusative case assigned and bears the semantic role commonly known as patient. (1) is a ditransitive construction, (2) is a transitive reflexive construction and (3) is a transitive reciprocal one.

(1) Se lo dije a Juan ayer .
Him-DAT it-ACC tell-PST.1SG to Juan yesterday.
'I told it to Juan yesterday.'

(2) Juan se peina .
Juan himself-ACC comb-PRS.3SG.
'Juan combs himself.'

(3) Ellos se envían cartas.
They them-DAT send-PRS.3PL letters.
'They send letters to each other.'

Example (4) corresponds to a *pure* pronominal construction (the pronoun is inherent to the predicate and its semantic meaning) where *se* does not bear a syntactic function. *Se* in (5) is an emphatic pronoun that is sometimes called emphatic or interest dative and that may be elided because it does not bear any semantic or syntactic function.

(4) Juan se desmayó de repente.
Juan him faint-PST.3SG suddenly.
'Juan suddenly fainted.'

(5) Juan (se) comió un bocadillo.
Juan him eat-PST.3SG a sandwich.
'Juan ate a sandwich.'

Se in examples in (6), (7), (8) and (9) does not behave as a pronoun bearing a syntactic, semantic, emphatic or discursive function like the ones in (1) - (5), but it works as a valency reduction mark signaling that the number of arguments of the main predicate is reduced. More concretely, the agentive external argument of the constructions in (6) – (9) is elided due to different linguistic strategies. (6) is an active construction where the agent is not present because it is the inchoative variant of the causative-inchoative alternation duplicity allowed by the predicate *romper*. (7) is a *media* voice construction where the agent *washer* is not present and where the property of *washing-well* is assigned to the shirt itself. (8) is a reflexive passive where the *looker* is not present for some reason. (9) is an impersonal construction where a general gone-without-saying subject is understood to perform the action of eating.

(6) El jarrón se rompió.
The vase - break-PST.3SG.
'The vase broke.'

(7) La camisa se lava muy bien.
The shirt - wash-PRS.3SG very well.
'The shirt washes very well.'

(8) Se buscan camareros.
- look.for-PRS.3PL bartenders.
'Bartenders are required.'

(9) Se come bien aquí.
- eat-PRS.3SG well here.
'It's a good place to eat in here.'

Moreno (1997, 2002) presents a unifying analysis that treats *se* constructions as a continuum of transitivity. Transitive constructions are placed at one end of the continuum where *se* can behave as an internal argument of the main predicate. All those *se* bearing the syntactic functions of direct and indirect objects are at this end of the continuum. Those *se* constructions that are traditionally considered paradigmatic (they belong to paradigmatic class) but where *se* does not bear any syntactic/semantic function, that is, *se* part of pure pronominal predicates and emphatic *se* are placed in the mid part of the continuum. Those *se* signaling main predicate valency reduction are placed at the other end of the continuum, that is, impersonal, passive *se* and those *se* that appear in *media* voice and inchoative constructions.

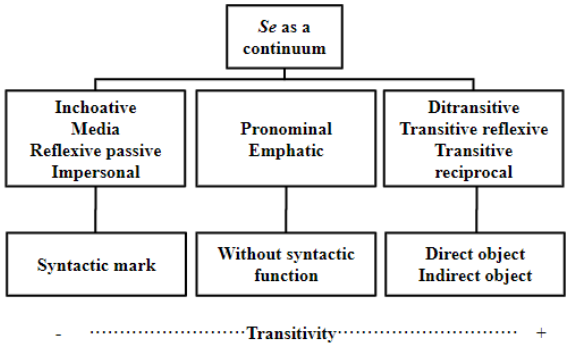

Figure 1: *Se* constructions as a continuum, of transitivity.

Following Moreno (1997, 2002) an annotation scheme for *se* constructions is proposed in the following section.

3 SE-corpus reduced version[1]

The *SE*-corpus reduced version (from now on *SE*-corpus) is composed of 2,140 sentences that come from CORPES XXI (*Real Academia Española de la Lengua, 2020*). The sentence selection procedure starts picking up, from the whole CORPES XXI, every sentence that contains the word *se* and belongs to the *news, leisure and daily life* domain in the European Spanish variant. From the output of this retrieval query, 3,000 sentences (complete SE-corpus) are randomly selected. Through the last filter, those sentences having more than one instance of *se* are eliminated.

Summing-up, the corpus used to carry out this research is composed of 2,140 sentences containing a single instance of *se*. The corpus is representative of the *news, leisure and daily life* domain in the European Spanish variant because it maintains real usage distribution of *se* constructions.

The annotation process is carried out following the next annotation criteria:

- *se-mark*: Cases of valency reduction (6) - (9).
- *expl*: Pure pronominal predicates or emphatic contexts (4) – (5).
- *iobj*: *Se* as indirect object of the main predicate (1) and (3).
- *obj*: *Se* as direct object of the main predicate (2).

4 Se tags distribution

The distribution of *se* tags presented in the corpus is quite unbalanced, as shown in table 1. The most prominent category (*se-mark*) is twelve times more frequent than the less prominent category (*obj*). Besides the intermediate categories, *expl* and *iobj* are quite extreme too: *expl* is close to the most frequent category (*se-mark*) whereas *iobj* is close in volume to the less frequent category *obj*. Thus, the corpus is unbalanced with two very frequent categories and two very infrequent ones. This distribution challenges the classification task.

Tag	Volume	%
se-mark	964	45.05
expl	946	44.21
iobj	154	7.2
obj	76	3.55
TOTAL	2,140	100

Table 1: *Se* tag distribution in the *SE-corpus.*

5 SE-corpus quality

The *SE*-corpus is annotated by a single annotator (annotator 1) due to human and time resources restrictions.[2] However, for the sake of

[1] The original *SE*-corpus is composed of 3,000 sentences that include one or more *se* per sentence. The reduced version of the *SE*-corpus is built from sentences that include a single instance of *se*.

[2] Annotation processes take quite a long time. Besides, it is not easy to find annotators with a

consistency and annotation quality, 100 sentences are annotated by the main annotator and two experts in the field of theoretical study of Spanish *se* (annotator 2 and annotator 3). All the annotators had the same annotation information, followed the same annotation guidelines and were aware of the 9 *se* types this classification experiments are focused on. The average inter-annotator agreement value[3] is 76.90%. The f1-score obtained by an average expert annotator against the gold standard is 0.85.

Pair of annotators	Agreement (%)
Anno1-Anno2	75.71
Anno1-Anno3	83.57
Anno2-Anno3	71.43

Table 2: Inter-annotator agreement.

Having a look at table 2, it can be observed that annotation agreement experiments some variations. The agreement value between annotator 1 and annotator 3 is higher in nearly 8 points than the agreement value between annotator 1 and annotator 2. The agreement value between annotator 1 and annotator 3 is also higher in 12.14 points than the agreement value between annotator 2 and 3. However, it is important to mention that the agreement value between annotator 2 and annotator 3 differs in 7.1 points with the next lowest agreement value, meaning that there are no significant differences in annotation quality nor consistency among the three annotators. Main disagreement cases come from media constructions that are not always easy to tell apart from pronominal predicates.[4] It is important to say that neither pronominal nor media constructions are part of the under-represented categories. The less frequent categories are those where *se* displays argument functions, namely, *obj* and *iobj*. These differences and similarities in agreement values may point towards the complexity of classifying Spanish *se* constructions and the possible alternative interpretations that may arise despite consistent annotations.

certain level of knowledge of the object of study, annotation, and computer skills.

[3] Raw or observed agreement (Bayerl and Paul, 2011), (Artstein, 2017).

[4] Pronominal predicates also called 'pure' pronominal predicates or inherently pronominal predicates in the literature introduced in section 2.

6 Classification strategies

To test whether the annotation scheme is efficient and can be easily learnt, and, whether the *SE*-corpus is big enough to deal with this classification problem, the *SE*-corpus is automatically segmented in train (1,713 sentences) and test (427 sentences) corpora.[5] Except for the es-BERT and UD-Pipe models, all text processing, vectorization steps and classifiers were implemented using Scikit Learn (Pedregosa et al., 2020). The tags of both the train and test corpora are preprocessed and turned into numbers using *LabelEncoder*. The classification task is performed by eight different models and a state-of-the-art parser. Precision (10), recall (11) and F1-score (12) are calculated per tag. Macro average F-score (13) and Accuracy, that is, the percentage of correct answers, show overall performance.[6] Model hyperparameters are tuned using a standard grid search with 5 folds stratified cross-validation. Parameter ranges are detailed in Appendix B. Different strategies are carried out for each concrete model to reduce the effect of unbalanced tag distribution:

- No balancing: models are trained using the unaltered training dataset.
- Search scoring (SC): Grid Search is configured to optimize the f1 macro scoring function.
- Class weight balancing (CW): the models are configured to give more relevance during training to patterns belonging to underrepresented classes.[7]
- Oversampling (OS): synthetic samples from underrepresented classes are added to the training dataset until all classes are balanced. The new samples are duplicates of samples already present in the training data.

[5] The test corpus remains the same along the experimental procedure for the sake of comparison between the different models and parser. The train corpus is expanded up to 3,195 sentences to run oversampling experiments.

[6] Equations taken from (Shmueli, 2019) and (Shung, 2018).

[7] For all the lineal models, a combination of search scoring and class weight strategies is also tested with similar results to the search scoring and class weight strategies applied independently.

$$\frac{True\ positives}{True\ positives + False\ positives} \quad (10)$$

$$\frac{True\ positives}{True\ positives + False\ negatives} \quad (11)$$

$$2 * \frac{Precision * Recall}{Precision + Recall} \quad (12)$$

$$\frac{\sum F\text{-}Score}{Total\ number\ of\ tags} \quad (13)$$

6.1 Baseline

The baseline model is generated by annotating the whole test corpus with the most frequent tag *se-mark* and comparing it to the gold standard. The accuracy and macro average f-score raise up to 0.49 and 0.16 respectively.

Metric	expl	iobj	obj	Se-mark
Precision	0.00	0.00	0.00	0.49
Recall	0.00	0.00	0.00	1
F-score	0.00	0.00	0.00	0.66
Acc/MAF[8]	**0.49 / 0.16**			

Table 3: Baseline results.[9]

6.2 Non-linear bag of words model

The first bag of words model is built with *CountVectorizer* and a *Random Forest Classifier* (Breiman, 2001) model. Random Forest is an ensemble of classification trees that has been shown to perform well on a wide range of problem. The best grid search parameters include pentagrams of characters from text inside word boundaries. The highest accuracy value obtained goes up to 0.63 and the highest macro average f-score reaches 0.34 points.

Metric	expl	iobj	obj	Se-mark	BS[10]
Precision	0.55	0.00	0.00	0.68	OP
Recall	0.73	0.00	0.00	0.65	
F-score	0.63	0.00	0.00	0.67	
Acc/MAF	0.61 / 0.32				
Precision	0.55	0.00	0.00	0.68	SC
Recall	0.73	0.00	0.00	0.65	
F-score	0.63	0.00	0.00	0.67	
Acc/MAF	0.61 / 0.32				
Precision	0.58	1.00	0.00	0.66	CW
Recall	0.65	0.03	0.00	0.74	
F- score	0.61	0.06	0.00	0.70	
Acc/MAF	**0.63** / **0.34**				
Precision	0.57	0.50	0.00	0.65	OS
Recall	0.61	0.03	0.00	0.75	
F-score	0.59	0.06	0.00	0.70	
Acc/MAF	0.62 / **0.34**				

Table 4: Non-linear bag of words model results.

6.3 Linear bag of words model

The second bag of words model is built with *CountVectorizer* and a Linear Support Vector Classification model (Fan et al. 2008). Such model has been widely used in text classification problems; however, it lacks the ability to deal with multiclass problems. Hence, an *OneVsRestClassifier w*rapper is applied to split the problem into 4 one-versus-rest binary problems. *GridSearch* best parameters include groups of n-grams from 5 to 7 characters within word boundaries. As shown in table 5, there is no result variation. The highest accuracy and macro average f-score values are 0.61 and 0.32, respectively.

Metric	expl	iobj	obj	Se-mark	BS
Precision	0.58	0.00	0.00	0.65	OP
Recall	0.65	0.00	0.00	0.71	
F-score	0.61	0.00	0.00	0.68	
Acc/MAF	**0.61** / **0.32**				
Precision	0.58	0.00	0.00	0.65	SC
Recall	0.65	0.00	0.00	0.71	
F-score	0.61	0.00	0.00	0.68	
Acc/MAF	**0.61** / **0.32**				
Precision	0.57	0.00	0.00	0.64	CW
Recall	0.64	0.00	0.00	0.71	
F-score	0.60	0.00	0.00	0.67	

[8] *Acc* stands for accuracy and *MAF* for macro average F-score.

[9] For baseline, *CountVectorizer*, *HashingVectorizer*, *TF-IDF* and *UD-Pipe* models, precision, recall, f-score and acc/MAF results are obtained from training and testing procedure using the *SE-corpus* and the original parameter configuration (OP); SC method results are obtained using the Search scoring strategy; CW results are obtained using the Class Weight balancing strategy; OS results are obtained using the Oversampling strategy.

[10] Balancing strategy.

Acc/MAF	**0.61 / 0.32**				
Precision	0.55	0.00	0.00	0.66	OS
Recall	0.68	0.00	0.00	0.66	
F-score	0.61	0.00	0.00	0.66	
Acc/MAF	0.60 / **0.32**				

Table 5: Linear bag of words model results.

6.4 Non-linear hashing model

As a variant of the non-linear bag of words model, a vectorization through the Hashing trick (Weinberger, 2009) was also explored. This vectorization is able to produce more space-efficient representations that can lead to better results. Table 6 shows the classification results of the first model composed of a *Hashing Vectorizer* and *Random Forest Classifier* algorithms. Using 100 classification trees, and a n-gram range of 5-7 characters from text inside word boundaries, the model achieves 0.62 accuracy points and 0.32 macro average points.

Metric	**expl**	**iobj**	**obj**	**Se-mark**	**BS**
Precision	0.55	0.00	0.00	0.66	OP
Recall	0.66	0.00	0.00	0.69	
F-score	0.60	0.00	0.00	0.67	
Acc/MAF	0.61 / **0.32**				
Precision	0.55	0.00	0.00	0.66	SC
Recall	0.66	0.00	0.00	0.69	
F-score	0.60	0.00	0.00	0.67	
Acc/MAF	0.61 / **0.32**				
Precision	0.54	0.00	0.00	0.63	CW
Recall	0.63	0.00	0.00	0.68	
F-score	0.58	0.00	0.00	0.65	
Acc/MAF	0.59 / 0.31				
Precision	0.6	0.00	0.00	0.64	OS
Recall	0.6	0.00	0.00	0.78	
F-score	0.6	0.00	0.00	0.70	
Acc/MAF	**0.62** / **0.32**				

Table 6: Non-linear *hashing* model results.

6.5 Linear hashing model

Similar to the previous model, the fourth model is a hashing version of the linear bag of words model. Again, a *GridSearch* algorithm extracts the best training parameters, that are 100 classification trees and a range of 5 to 7 n-grams of characters from text inside word boundaries. The accuracy goes up to 0.61 points and the macro average to 0.32.

Metric	**expl**	**iobj**	**obj**	**Se-mark**	**BS**
Precision	0.53	0.00	0.00	0.65	OC
Recall	0.68	0.00	0.00	0.64	
F-score	0.60	0.00	0.00	0.65	
Acc/MAF	0.59 / 0.31				
Precision	0.53	0.00	0.00	0.65	SC
Recall	0.68	0.00	0.00	0.64	
F-score	0.60	0.00	0.00	0.65	
Acc/ MAF	0.59 / 0.31				
Precision	0.56	0.00	0.00	0.66	CW
Recall	0.66	0.00	0.00	0.70	
F-score	0.61	0.00	0.00	0.68	
Acc/ MAF	**0.61** / **0.32**				
Precision	0.55	0.00	0.00	0.68	OS
Recall	0.69	0.00	0.00	0.68	
F-score	0.61	0.00	0.00	0.68	
Acc/MAF	**0.61** / **0.32**				

Table 7: Linear *hashing* model results.

6.6 Non-linear TF-IDF

The fifth model is formed by a combination of *TF-IDF* and *Random Forest Classifier*. *TF-IDF* is a weighed variant bag of words, that promotes words that are highly specific of the document under analysis. The best training parameters extracted by a *GridSearch* algorithm convey 100 classification trees and a range of 5 to 7 n-grams of characters from text inside word boundaries. The highest accuracy value goes up to 0.63 points and the macro average to 0.35.

Metric	**expl**	**iobj**	**obj**	**Se-mark**	**BS**
Precision	0.55	0.00	0.00	0.67	OC
Recall	0.67	0.00	0.00	0.70	
F-score	0.61	0.00	0.00	0.68	
Acc/MAF	0.61 / 0.32				
Precision	0.55	0.00	0.00	0.67	SC
Recall	0.67	0.00	0.00	0.70	
F-score	0.61	0.00	0.00	0.68	
Acc/MAF	0.61 / 0.32				
Precision	0.58	1.00	0.00	0.67	CW
Recall	0.67	0.03	0.00	0.72	
F-score	0.62	0.06	0.00	0.70	
Acc/MAF	**0.63** / **0.35**				
Precision	0.59	0.50	0.00	0.64	OS
Recall	0.60	0.03	0.00	0.77	
F-score	0.59	0.06	0.00	0.7	
Acc/MAF	0.62 / 0.32				

Table 8: Non-linear *TF-IDF* model results.

6.7 Linear TF-IDF model

The second TF-IDF model is built up with *TF-IDF* and a linear SVC classifier. The

GridSearch algorithm yields that the best parameters include 100 classification trees and a n-gram range between 5 and 7 characters found within word boundaries. The accuracy reaches 0.64 points, and the macro average goes up to 0.34.

Metric	expl	iobj	obj	Se-mark	BS
Precision	0.57	0.00	0.00	0.72	OC
Recall	0.75	0.00	0.00	0.7	
F-score	0.65	0.00	0.00	0.71	
Acc/MAF	**0.64** / **0.34**				
Precision	0.57	0.00	0.00	0.72	SC
Recall	0.75	0.00	0.00	0.7	
F-score	0.65	0.00	0.00	0.7	
Acc/MAF	**0.64** / **0.34**				
Precision	0.57	0.00	0.00	0.72	CW
Recall	0.75	0.00	0.00	0.70	
F-score	0.65	0.00	0.00	0.71	
Acc/MAF	**0.64 / 0.34**				
Precision	0.55	0.00	0.00	0.69	OS
Recall	0.71	0.00	0.00	0.68	
F-score	0.62	0.00	0.00	0.69	
Acc/MAF	0.62 / 0.33				

Table 9: Linear *TF-IDF* model results.

6.8 Recurrent network with embeddings

All the models presented above confront the learning task with no prior knowledge of the Spanish language, a trait that might limit the performance on some applications. A common approach to inject some semantic and syntactic knowledge is to make use of word embeddings (Mikolov, 2013), whereby a numerical vector representative of each word is pre-trained with a large unannotated corpus, then used as inputs for the task at hand instead of the original words. In this work we use the Spanish embeddings provided by fasttext project (Bojanowski, 2016).

The simplest way to use word embeddings is to compute a document embedding as the average of embeddings the words in the document, then feed such sentence vector into a machine learning model (e.g. Random Forest). However, this approach turned out to produce very poor results for our task. Instead, we resort to implementing a small recurrent neural network with GRU layers (Cho, 2014) to obtain a better mixing of the embedding vectors.

The network is comprised of an Embedding layer, 1 to 3 GRU layers (the first one bidirectional), global average pooling and 1 to 3 dense layers with ReLU activations. Dropouts are added at the embeddings, GRU and dense levels to prevent overfitting. We do not fine-tune the embedding vectors. Since many parameters in the network design are susceptible to tuning, we run a Bayes Search optimization strategy, as implemented in scikit-learn. With this, we are able to attain an accuracy of 0.62 and macro average f1 of 0.41.

Metric	expl	iobj	obj	Se-mark
Precision	0.56	0.29	0.50	0.71
Recall	0.71	0.19	0.07	0.64
F-score	0.63	0.23	0.12	0.68
Acc/MAF	0.62 / 0.41			

Table 10: *Recurrent network with embeddings model* results.

6.9 es-BERT[11]

Recent advances in statistical NLP are mainly based on making use of a fully pre-trained deep neural network that models the conditional distribution of tokens in a specific language: a language model. In particular, the BERT model has proven very successful in many applications (Devlin, 2018). Such model is adapted to specific NLP tasks through a so-called fine-tuning procedure. The first BERT-based model for Spanish is *es-BERT* (Cañete, Chaperon and Fuentes, 2020). We used the Transformers library (Wolf et al, 2019) to train an es-BERT classifier. Following a similar approach to the previous model, to perform the hyperparameter tuning we follow a Bayes Search strategy. The resulting accuracy goes up to 0.83 points and the macro average raises to 0.70.

Metric	expl	iobj	obj	Se-mark
Precision	0.75	0.71	0.50	0.95
Recall	0.89	0.65	0.36	0.84
F-score	0.82	0.68	0.42	0.89
Acc/MAF	0.83 / 0.70			

Table 11: *es-BERT* results.

6.10 UD-Pipe

UD-Pipe (Straka and Straková, 2017) (Straka, Hajič and Straková, 2016) is a state-of-the-art, embedding-based,[12] dependency parsing tool,

[11] An adaptation of Barbero (2020) was used to train transformer-based models.

[12] UD-Pipe is a neural network parser based on embeddings. Form embeddings are adjusted from

capable of analyzing different linguistic aspects (lemma, PoS, morphological features, dependency relations) of each token of a sentence encoded in CoNLL-U (Universal Dependencies, 2020) format.[13] The model used in text classification is *Spanish-gsd-ud-2.5-191206.udpipe* (Ballesteros et al., 2019).[14] To predict the tags assigned to each instance of *se* the whole architecture (tokenizer, tagger and parser) is re-trained using the default parameter configuration. The results achieved go up to 0.62 points of accuracy and 0.45 points of macro average F-score.

Metric	**expl**	**iobj**	**obj**	**Se-mark**	**BS**
Precision	0.56	0.64	0.00	0.70	OC
Recall	0.72	0.29	0.00	0.62	
F-score	0.63	0.40	0.00	0.66	
Acc/MAF	**0.62** / 0.42				
Precision	0.61	0.35	0.10	0.70	OS
Recall	0.55	0.39	0.21	0.70	
F-score	0.58	0.37	0.14	0.70	
Acc/MAF	0.60 / **0.45**				

Table 12: *UD-Pipe* results.

7 Results

Table 13 shows the highest accuracy and macro average F-score values obtained for each of the models and the parser in the different training experiments. The highest accuracy value is reached by *es-BERT* model (0.83). The highest macro average f-score is also achieved by *es-BERT* model (0.70).

It is important to mention that the value accuracy reached for most models doubles the macro average F-score. However, in the case of models that make use of some kind of transfer learning (recurrent network with embeddings, BERT and *Spanish-gsd-ud-2.5-191206.udpipe)* the difference between accuracy and macro average F-score values is around 0.13-0.20 points. This might mean that, whereas classical classification models always pay more attention to the most frequent tags, models making use of prior knowledge seem to take more into consideration the whole tag set distribution. This hypothesis is supported by the precision, recall and f-score values obtained for the less frequent tags *iobj* and *obj*: whereas BERT-like and UD-Pipe model learn to discriminate the four categories, the non-linear bag of words and non-linear TF-IDF models learn to discriminate the three most frequent categories *se-mark, expl* and *iobj,* but ignore the category *obj*. Linear bag of words, hashing and linear TF-IDF models together with non-linear hashing model only learn to discriminate the two most frequent categories *se-mark* and *expl*, paying no attention to *iobj or obj* cases. Besides, it is important to mention that the best performing models make use of transfer learning: they use and adjust already learnt information whereas classic models need to learn to disambiguate from scratch without any other additional information. Furthermore, the very best results are obtained by BERT, showing that doing transfer learning of not just the word representations but also the mixing layers contributes positively to this task. Our hypothesis is that syntactic knowledge of the Spanish language is required to perform *se* classification correctly, and so the pre-trained Transformer layers are providing critical contextual information to expose such syntactic elements. It is also remarkable how the performance of BERT is close in accuracy to that of an expert human annotator, though a gap still exists in f1-score due to misclassifications in minority classes.

Having a look at the confusion matrix obtained from the best classification model, es-Bert, it can be seen that class frequency is directly related to the higher accuracy results: the model learns better to classify the most frequent classes *expl* and *se-mark*. On the contrary, the model gets worse results for the less represented classes *iobj* and *obj*.

It is important to mention that the model never predicts the tag *iobj* in front of direct object *se* or valency reduction values of *se*. Besides, the model rarely predicts the tag *se-mark* for argumental *se cases.* However, the model gets confused and sometimes assigns the tag *expl* to argumental uses of *se* (14)-(15) and the other way round (16).

Spanish word2vec embeddings. The rest of embedding layers are randomly started and adjusted along the training procedure. See appendix C for more information on UD-pipe architecture.

[13] There are other state of the art parsing tools such as FreeLing (Padró & Stanilovsky, 2012), Ixa-pipes (Agerri, Bermudez & Rigau, 2014), Stanza (Qi et al., 2020) or Spacy (Honnibal & Montani, 2017). Usability and training ease have been key aspects for the selection of UD-Pipe.

[14] See Straka and Straková (2017) and Straka and Straková (2019) for a detailed description of the training and hyperparameter adjustment procedure.

Model	Accuracy	Macro Avg
Baseline model	0.49	0.16
CountVectorizer + RandomForestClassifier + GridSearchCV	0.61	0.33
CountVectorizer + OneVsRestClassifier + LinearSVC	0.61	0.32
HashingVectorizer + RandomForestClassifier + GridSearchCV	0.62	0.33
HashingVectorizer + OneVsRestClassifier + LinearSVC	0.61	0.32
TF-IDF + RandomForestClassifier + GridSearchCV	0.65	0.34
TF-IDF + OneVsRestClassifier + LinearSVC	0.65	0.34
Recurrent network with embeddings	0.62	0.41
es-BERT	**0.83**	**0.70**
Spanish-gsd-ud-2.5-191206.udpipe	0.62	0.45
Expert human annotator (average)	0.88	0.85

Table 13: Summary of best results.

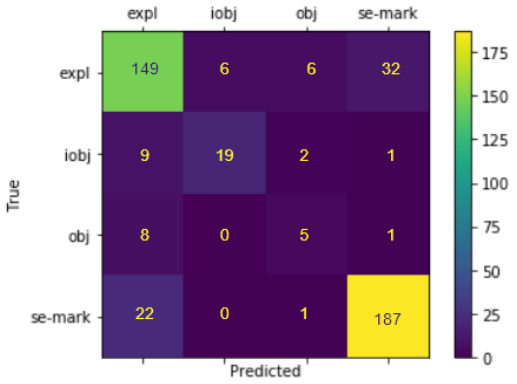

Figure 2: es-BERT confusion matrix.

(14) El maestro José Fernández
The maestro José Fernández
se ha propuesto redescubrir
him-DAT have-PRS.3SG propuesto redescubrir […].
propose-PTCP rediscover-INF […].
'Maestro José Fernández has proposed himself to rediscover […].'

(15) Aquí cerca , el joven Tomás
Here nearby, the young Tomás
Rodaja […] , **se**
Rodaja […] , **him-ACC**
ofrecía como
offer-PST.IMPV.3SG as
criado […] .
servant […] .

(16) […] atraca su velero […] , se
[…] docks his boat […] , him
alquila una villa
rent-PRS.3SG a vaccation-house
o dos y juega con […].
or two and play-PRS.3SG with […].

8 *Conclusions and further work*

Se constructions constitute a complex linguistic phenomenon that challenges annotation criteria creation, annotation and automatic classification tasks. Transformer-based models entail exceptional advantages for complex classification problems like the one posed by *se* constructions, obtaining the highest accuracy and f-score classification values. Corpus unbalance is an important factor affecting the results, which prevents attaining automated annotations on par with those of an expert human annotator. Thus, future work needs to be done into the following research lines: first, enlarging the existing *SE*-corpus while maintaining the real distribution of *se*-constructions, and second, evaluating whether this enlarged version of the *SE-corpus* may palliate category unbalance improving classification results. Another open research line is to study how to integrate a *se* construction classifier as an extra module into a NLP pipeline to turn it into a general use tool.

Acknowledgements

We would like to thank Cristina Sánchez López and Amaya Mendikoetxea Pelayo for their help and dedication in the development of this study. The authors acknowledge financial support from PID2019-106827GB-I00 / AEI / 10.13039/501100011033 and from the European Regional Development Fund and from the Spanish Ministry of Economy, Industry, and Competitiveness - State Research Agency, project TIN2016-76406-P (AEI/FEDER, UE).

References

Agerri, R., J. Bermudez and G. Rigau. 2014. IXA pipeline: Efficient and Ready to Use

Multilingual NLP tools. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014), May, 2014, Reykjavik, Iceland.*

Artstein, R. 2017. Inter-annotator Agreement. In Ide, N. and J. Pustejovsky (Eds). *Handbook of Linguistic Annotation*. Springer: Dordrecht, 297-314.

Bayerl, P. and K. Paul. 2011. What determines inter-coder agreement in manual annotations? Ametaanalytic investigation. *Computational Linguistics*. 37(4): 699-725.

Ballesteros, M., H. Martínez, R. McDonald, E. Pascual, N. Silveira, D. Zeman and J. Nivre. 2019. Spanish-gsd-ud-2.5-191206.udpipe https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3105. Accessed date: 23/09/2020.

Barbero, A. 2020. Training transformer models. https://github.com/Spain-AI/transformers. Accessed date: 26/09/2020.

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

Breiman, L. 2001. Random Forests. *Machine Learning 45*, 5-32. https://doi.org/10.1023/A:1010933404324.

Campos, H. 1999. Transitividad e intransitividad. In Bosque, I. and V. Demonte (Eds.) *Gramática descriptiva de la Lengua Española*. Madrid: Espasa. V2, 1519-1574.

Cañete, J., G. Chaperon and R. Fuentes. 2020. Spanish pre-trained BERT model and evaluation data. To appear in *ICLR 2020 workshop*. https://users.dcc.uchile.cl/~jperez/papers/pml4dc2020.pdf. Accessed date: 22/09/2020.

Cho, K., B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Devlin, J., M. Chang, K. Lee and K. Toutanova. 2018. BERT: Pre-trained of Deep Bidirectional Transformers for Language Understanding. *Computer Science*. https://arxiv.org/abs/1810.04805. Accessed date: 22/09/2020.

Fan, R., K. Chang, C. Hsieh, X. Wang and C. Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 9, 1871-1874.

Fernández Montraveta, A., and G. Vázquez. 2017. *Las construcciones con se en español (Cuadernos de lengua española 130).* Madrid: Arco/Libros-La Muralla.

Honnibal, M. and I. Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Mendikoetxea, A. 1999 a. Construcciones con se: Medias, pasivas e impersonales. In Bosque, I. and V. Demonte (Eds.) *Gramática descriptiva de la Lengua Española*. Madrid: Espasa. V2, 1631-1722.

Mendikoetxea, A. 1999 b. Construcciones inacusativas y pasivas. In Bosque, I. and V. Demonte (Eds.) *Gramática descriptiva de la Lengua Española*. Madrid: Espasa. V2, 1575-1630.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Moreno, J.C. 1997. *Introducción a la lingüística general: un enfoque tipológico y universalista.* Madrid: Editorial Síntesis.

Moreno, J.C. 2002. *Curso Universitario de Lingüística General*. Madrid: Editorial Síntesis.

Padró, L. and E. Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA May, 2012. Istanbul, Turkey*.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830. https://scikit-learn.org. Accessed date: 22/09/2020

Qi, P., Y. Zhang, Y. Zhang, J.Bolton, and C.D. Manning. 2020. Stanza: A Python Natural

Language Processing Toolkit for Many Human Languages. In *Association for Computational Linguistics (ACL) System Demonstrations*, 101-108.

Real Academia Española de la Lengua. 2020. Banco de datos (CORPES XXI) [online]. Corpus del Español del Siglo XXI (CORPES) https://www.rae.es/recursos/banco-de-datos/corpes-xxi. Accessed date: 20/09/2020

Sánchez, C. 2015. Se y sus valores. In Gutiérrez Rexach, J. (Ed.) *Enciclopedia de Lingüística Hispánica*, Vol. 2, 1-12.

Sánchez, C. 2002. Las construcciones con se (Gramática del Español, 8). Madrid: Visor Libros.

Scikit optimize. https://scikit-optimize.github.io/stable/. Accessed date: 15/11/2020.

Shmueli, B. 2019. Multi-Class Metrics Made Simple, Part II. *Towards Data Science*. https://towardsdatascience.com/multi-class-metrics-made-simple-part-ii-the-f1-score-ebe8b2c2ca1. Accessed date: 20/09/2020.

Shung, K. 2018. Accuracy, Precision, Recall or F1? *Towards Data Science*. https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9. Accessed date: 20/09/2020.

Straka, M. and J. Straková. 2019. Universal Dependencies 2.5 Models for UDPipe (2019-12-06). https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3131. Accessed date: 21/11/2020

Straka, M. and J. Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, Canada, August.

Straka M., J. Hajič and J. Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, May.

Universal dependencies. 2020. *CoNLL-U format*. https://universaldependencies.org/format.html. Accessed date: 23/09/2020.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998-6008.

Weinberger, K., A. Dasgupta, J. Langford, A. Smola, and J. Attenberg. 2009. Feature hashing for large scale multitask learning. In *Proceedings of the 26th annual international conference on machine learning*, 1113-1120.

Wolf, T. L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest and A. M. Rush. 2019. Transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.

A Supplementary material

To aid in the reproducibility of the results presented in this manuscript, both the SE-corpus and a Jupyter notebook with the experimental procedure are available as supplementary material at the following link: https://github.com/albarji/sepln-spanish-se-constructions.

B Hyperparameter ranges

The following hyperparameter ranges were used for searching for optimal model parameters.

Model	**Parameter**	**Values**
Count Vectorizer / Hashing Vectorizer	Analyzer	char_wb
	Binary counts	[False, True]
	N-gram range	[(1,1), (1,2), (1,3), (1,4), (1,5), (2,2), (3,3), (3,5), (5,5), (5,7), (7,7), (10,10)
Random Forest	Number of estimators	[1, 10, 100]
LinearSVC	C	[1e-4, 1e-3, …, 1e4]
Recurrent network with embeddings	Spatial dropout	[0.0, …, 0.9]
	GRU layers	[1, 2, 3]
	GRU units	[16, 32, …, 1024]
	GRU dropout	[0.0, …, 0.9]
	Dense layers	[1, 2, 3]
	Dense units	[16, 32, …, 1024]
	Dense dropout	[0.0, …, 0.9]

	Training epochs	[50, …, 200]
es-BERT	Model casing	[cased, uncased]
	Learning rate	[10^{-6}, …, 10^{-4}]
	Training epochs	[1, …, 10]
	Batch size	[4, 8, 16, 32, 64]
	Attention dropout	[0.0, …, 0.9]
	Hidden dropout	[0.0, …, 0.9]

Table 14: Hyperparameter ranges.

C Spanish-gsd-ud-2.5-191206.udpipe params

The following parameters are the ones used along the training procedure of Spanish-gsd-ud-2.5-191206.udpipe. The same params are used to perform the experiments in 6.9.

Module	**Parameter**	**Values**
Tokenizer	Dimension	24
	Epochs	100
	Segment_size	200
	Initialization_range	0.1
	Batch_size	50
	Learning_rate	0.005
	Learning_rate_final	0
	dropout	0.2
	early_stopping	1
Tagger	es_gsd models	2
	templates_1	tagger
	guesser_suffix_rules_1	10
	guesser_enrich_ dictionary_1	6
	guesser_prefixes_ max_1	0
	use_lemma_1	1
	use_xpostag_1	1
	use_feats_1	1
	provide_lemma_1	0
	provide_xpostag_1	1
	provide_feats_1	1
	prune_features_1	0
	templates_2	lemmatizer
	guesser_suffix_rules_2	4
	guesser_enrich_ dictionary_2	4
	guesser_prefixes_ max_2	4
	use_lemma_2	1
	use_xpostag_2	1
	use_feats_2	1
	provide_lemma_2	1
	provide_xpostag_2	0
	provide_feats_2	0
	prune_features_2	0
Parser	es_gsd iterations	30
	embedding_upostag	20
	embedding_feats=20	20
	embedding_xpostag	0
	embedding_form	50
	embedding_form_file	=../ud-2.5-embeddings/ es_gsd.skip. forms.50. vectors
	embedding_lemma	0
	embedding_deprel	20
	learning_rate	0.01
	learning_rate_final	0.001
	l2	0.5
	hidden_layer	200
	batch_size	10
	transition_system	Link2
	transition_oracle	Static
	structured_interval	8

Table 15: Spanish-gsd-ud-2.5-191206.udpipe params.

Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 165-175 recibido 29-11-2020 revisado 19-01-2021 aceptado 30-01-2021

Generación automática de meta-resúmenes para la evaluación del manejo de estructuras discursivas y coherencia en el alumnado

Automatic generation of meta-summaries for evaluation of the handling of discursive structures and coherence in students

Unai Atutxa[1], **Alejandro Molina-Villegas**[2], **Mikel Iruskieta**[1]
[1] HiTZ Center - Ixa, University of the Basque Country UPV/EHU, Spain
[2] Conacyt - Centro de Investigación en Ciencias de Información Geoespacial, Mexico
atutxaunai@gmail.com, amolina@centrogeo.edu.mx, mikel.iruskieta@ehu.eus

Resumen: La técnica de crowd-sourcing puede ser una herramienta de gran ayuda tanto para evaluar los resúmenes de los alumnos como para poder ofrecerles un *feedback* que ayude a mejorar sus destrezas para resumir. En este trabajo, se propone un enfoque para la generación de meta-resúmenes en euskera, con el objetivo de diseñar y desarrollar una evaluación automática de los resúmenes de extracción. Se presenta un nuevo algoritmo que permite usar los meta-resúmenes generados con las siguientes finalidades: *i*) comparar los resúmenes elaborados por alumnos de diferentes edades y cursos educativos (primaria y universidad), *ii*) evaluar los resúmenes creados en clase (evaluación de la clase) y *iii*) evaluar a cada alumno (evaluación individual). Los resultados muestran que el método propuesto, el cual se ha elaborado basándose en aspectos cualitativos (estructura discursiva de la coherencia) y cuantitativos (kappa de Fleis y distancia de Hamming), es apto para comparar grupos e individuos.
Palabras clave: Evaluación de Resumen Automático, Análisis del discurso, PLN en Euskera.

Abstract: Crowd-sourcing can help teachers to evaluate student summaries and give them feedback to improve their summarization skills. In this paper, we propose an approach for meta-summaries generation, to design and develop the automatic evaluation of extractive summaries for the Basque language. We propose a novel algorithm that allows to use the generated meta-summaries to *i*) compare students meta-summaries at different ages and education stages (elementary and undergraduates), *ii*) evaluate classroom meta-summaries (classroom evaluation) and *iii*) evaluate each student (individual evaluation). The results show that our proposed method, based on qualitative (coherence discourse structure) and quantitative (Fleis kappa and Hamming distance) measures, is accurate to compare both: groups and individuals.
Keywords: Summarization Evaluation, discourse Analysis, Basque NLP.

1 Introducción

En la actualidad, el exceso de información puede conducirnos a no distinguir lo que es realmente relevante. En ese contexto, el resumen es de vital importancia en áreas como la educación, ya que un resumen muestra la capacidad de comprensión y de síntesis de quien lo ha elaborado. El hecho de que un alumno disponga de una gran cantidad de información no significa que vaya a entender más y mejor el tema de estudio, lo cual puede provocar que el proceso de aprendizaje no sea óptimo. Esto sugiere que trabajar el resumen de una manera eficaz en el aula puede ser de gran ayuda para lograr los objetivos curriculares.

Cuando se resume, primeramente se debe entender lo que se ha leído, y después, plasmar las ideas más importantes del texto adecuando el lenguaje al conocimiento propio.

ISSN 1989-7553. DOI 10.26342/2021-66-14

Las ideas más relevantes deben ser extraídas del texto, manteniendo la coherencia entre dichas ideas. Sin embargo, los resúmenes producidos por los estudiantes no siempre coinciden con los resúmenes producidos por los profesores o evaluadores. En el ámbito educativo, es habitual que los libros de texto (o la información de los sistemas de gestión de aprendizaje) empleados en la escuela no dispongan de actividades que desarrollen la destreza de resumir, al menos en la educación del País Vasco. En consecuencia, raramente se trabajan estrategias que ayuden a desarrollar la capacidad de resumen. Además, ni siquiera se trabaja lo suficiente el resumir los textos, poco comprensible, ya que como indica Sanz (2005) resumir en sí es una estrategia muy útil para mejorar la comprensión lectora y la expresión escrita. Ante esta situación, utilizar una evaluación basada en crowd-sourcing y emplear técnicas de Procesamiento de Lenguaje Natural (PLN) puede ser de gran ayuda para que profesores y estudiantes evalúen de manera automática los resúmenes y obtengan un *feedback* que les haga mejorar su capacidad de resumen.

El objetivo de este trabajo es lograr modelos fiables mediante crowd-sourcing que sirvan para: i) comparar los resúmenes elaborados por alumnos de diferentes edades y cursos educativos (primaria y universidad), ii) evaluar los resúmenes modelo creados en clase (evaluación colectiva de la clase) y iii) evaluar a cada alumno (evaluación individual). Según los resultados, cuando los alumnos de primaria resumen textos previamente trabajados en clase, la capacidad de resumir que muestran en los textos más fáciles es muy similar a la de los universitarios. Sin embargo, cuando se trata de resumir textos más complejos, queda patente que los universitarios muestran una capacidad superior. En cuanto al método, el cual se ha elaborado basándose en aspectos cualitativos (estructura discursiva de la coherencia) y cuantitativos (kappa de Fleis y distancia de Hamming), los resultados indican que el método empleado para comparar tanto grupos como individuos es adecuado.

En cuanto a la evaluación de resúmenes, este siempre ha sido un tema complejo y controvertido en la lingüística computacional (Saggion et al., 2010). ROUGE (CY, 2004), BLEU (Papineni et al., 2002), Pyramid (Nenkova y Passonneau, 2004) y SummTriver (Cabrera-Diego y Torres-Moreno, 2018), son métodos que permiten evaluar resúmenes de manera automática. Sin embargo, tal y como lo explican Molina y Torres (2015), muchas de estas métricas pueden acarrear ciertas desventajas, ya que resúmenes no gramaticales pueden obtener puntuaciones muy altas. Además, no se debe olvidar que algunos de estos métodos necesitan modelos de referencia hechos por expertos, lo cual es un problema a la hora de aplicarlo en un escenario real como puede ser la escuela. Por un lado, los profesores carecen de tiempo para poder crear resúmenes modelo de varios textos y después corregir todos los trabajos realizados (todos los resúmenes de 25-35 estudiantes). Por otro lado, como bien indican Radev y Tam (2003), existe más de un modelo adecuado para realizar un resumen, pudiendo ser dos resúmenes igualmente buenos aun estando elaborados con frases totalmente distintas. Esto muestra la necesidad de contar en el aula con varios modelos de un mismo resumen, labor que es difícilmente realizable por un único profesor. Por lo tanto, es necesario trabajar con varios modelos de referencia, para poder ofrecer un modelo tanto para la máquina, como para dar un *feedback* adecuado al docente y estudiante. La medida *Relative Utility* (Radev y Tam, 2003) utiliza más de un modelo de referencia para evaluar los resúmenes. Aun así, poder lograr varios modelos de referencia de todos los textos curriculares, es una tarea dificil para el docente, lo que deja en evidencia las limitaciones que puede llegar a tener dicho modelo.

En consecuencia, además de los desafíos habituales de la comprensión de un texto, es fundamental adaptar las herramientas, y así proporcionar un entorno propicio para el aprendizaje (de los estudiantes) y la evaluación (para los profesores). Para ello, es necesario tener en cuenta los recursos y las limitaciones (lingüísticas y no lingüísticas) que existen en las escuelas vascas si se quiere crear un entorno eficiente para desarrollar habilidades de síntesis y procedimientos de evaluación.

2 Estado de la cuestión

Como indica Molina-Villegas (2013b), una posible solución para recopilar resúmenes de referencia es recurrir a la participación de voluntarios no expertos en tareas científicas que no requieran experiencia en el tema. Invo-

lucrar ciudadanos en actividades científicas puede ser interesante y útil, ya que permite ahorrar recursos humanos; aumentar la cantidad y velocidad del procesamiento de datos; o simplemente acercar las personas a la ciencia. Molina-Villegas (2013b) desarrolló un sistema de crowd-sourcing para compilar un corpus de resumen automático en español donde se obtuvieron cerca de tres mil resúmenes en tan solo 10 semanas, lo que demuestra su utilidad para reunir un corpus amplio en un espacio de tiempo muy concreto. Siendo considerables las ventajas, es imprescindible centrarse en las posibles desventajas, para poder después contrarrestarlas. La principal desventaja es que los participantes pueden llegar a actuar al azar, o es posible que carezcan de criterios específicos. En consecuencia, en este artículo trataremos de solventar estas desventajas utilizando la prueba exacta de Fisher, para verificar si los participantes han respondido al azar, y el coeficiente kappa para comprobar si los participantes han respondido con criterios similares a los de un evaluador.

En cuanto al resumen automático, y en concreto al resumen de extracción, en palabras de Saggion y Poibeau (2013) la selección de oraciones de un resumen extractivo puede basarse en información estadística o en una teoría que tenga en cuenta la información lingüística y semántica. Por tanto, para realizar un resumen extractivo automático se utilizan técnicas superficiales y enfoques basados en el conocimiento. Molina-Villegas (2013a) por ejemplo, presenta un estudio sobre la comprensión de oraciones y propone un modelo de regresión que predice la eliminación de segmentos dentro de la oración con aplicación en la generación de resumen abstractivo. En cuanto a herramientas se refiere, a día de hoy es posible encontrar varias herramientas que permiten generar resúmenes de manera automática, a destacar MEAD (Radev et al., 2004) y SUMMA (Saggion, 2008). Estas herramientas, junto con otras, dan opción a trabajar con lenguas como el inglés o el chino, por ejemplo. Pero hasta donde nosotros sabemos, no hay ninguna herramienta de resumen automático que emplee información en euskera. Esto se debe a que estas herramientas no utilizan analizadores morfológicos, los cuales son necesarios para poder tratar con una lengua aglutinante como el euskera.

Así, Uno de los retos principales de este trabajo era reunir un corpus amplio de extracciones. Para lograrlo, se ha utilizado Compress-eus (Atutxa et al., 2017) que recopila el corpus de extracciones y abstracciones de los resúmenes realizados de textos en euskera. Además, esta herramienta ofrece información automática de las extracciones.

En cuanto a la evaluación humana, Atutxa (2018) propone criterios de evaluación para evaluar extracciones y abstracciones en euskera. Para ello se basan en la guía BABAR (Álvarez, 2004) y en la información que proporciona la estructura relacional de discurso (coherencia). La guía BABAR tiene como objetivo evaluar textos expositivos en inglés como lengua extranjera, y en ella se evalúan los siguientes apartados: contenido, organización, vocabulario, uso de la lengua y presentación.

Evaluar un resumen no es una tarea sencilla, ya que requiere evaluar un proceso complejo. Si la evaluación de resúmenes es compleja para los humanos, la evaluación automática lo es aún más, especialmente en lo que respecta a la fiabilidad. En la evaluación automática, Saggion y Poibeau (2013) diferencian dos tipos de evaluación:

i) Evaluación de un resumen automático comparándolo con *gold standards* creados manualmente: el resumen elaborado es evaluado comparándolo con un resumen de referencia creado por un humano. Por lo que nos consta, es este el método utilizado por Zipitria, Arruarte, y Elorriaga (2008) con LEA (una aplicación web para la evaluación de resúmenes en euskera). Zipitria et al. (2008) explica el método utilizado en su trabajo. Primero, observan cómo se realiza la evaluación humana y, luego, automatizan sus observaciones. 15 expertos tuvieron que resumir 5 textos para crear un modelo para la evaluación de los resúmenes. Los expertos estaban formados por cinco profesores de secundaria, cinco profesores de euskera de segunda lengua (L2) y cinco profesores universitarios.

ii) Evaluación de resúmenes automáticos sin *gold standard*: para evaluar se utiliza la información que contiene el propio resumen. Louis y Nenkova (2009) presentaron un marco para evaluar el contenido usando el input como referencia. Se basa en el hecho de que la distribución de palabras en el input (texto a resumir) y el resumen de ese *input* deben ser similares. En el caso del euskera, no se conoce

ningún trabajo relacionado con la evaluación de resúmenes automáticos sin ningún corpus *gold standard* para comparar.

3 Materiales y método

3.1 Compress-eus: una herramienta para recopilar resúmenes

Uno de los principales problemas para trabajar en una lengua como el euskera suele ser no contar con un corpus adecuado. Para ello, es muy útil contar con una herramienta de características similares a las de Compress-eus. Compress-eus es una interfaz que permite reunir resúmenes elaborados por alumnos y profesores. El texto original está segmentado en unidades elementales del discurso (EDU) y la idea más importante o la Unidad Central (UC) está anotada. El alumno realiza el resumen extractivo, eliminando del texto las EDUs que considere oportunas, es decir, los segmentos menos relevantes. Cuando finaliza la extracción, el usuario guarda el resumen. El uso de esta interfaz facilita la recopilación y el análisis del resúmenes. Además, facilita utilizar información jerárquica de la estructura discursiva de los textos originales, pero también de los resúmenes extractivos, lo que será de gran ayuda para este trabajo.

Compress-eus proporciona el seguimiento de todas las operaciones realizadas por el usuario (qué EDU se eliminó, si se eliminó la CU, por ejemplo) al realizar la extracción. Además, facilita la siguiente información sobre los textos a resumir: número de párrafos (en el texto); número de oraciones (en el texto y cada párrafo); número de EDUs (en el texto, cada párrafo y cada oración).

3.2 Corpus

Para este trabajo se ha reunido un corpus compuesto por 1036 resúmenes de extracción. 352 son de alumnos de quinto año de primaria (9-10 años), 88 estudiantes han resumido 4 textos cada uno. Los 684 resúmenes restantes han sido realizados por alumnos universitarios de la Facultad de Educación de Bilbao, en este caso han sido 171 alumnos quienes han resumido los 4 textos. Para tener un escenario real de enseñanza, fueron utilizados textos procedentes de libros de texto que se emplean en la escuela para llevar a cabo el programa curricular. De hecho, el alumno tuvo que resumir los textos seleccionados en las fechas reales programadas para ello. Los detalles de los textos utilizados para los experimentos se describen en la Tabla 1.

	Párrafos	Oraciones	EDUs	Pals.	UC
T1	5	11	17	121	1
T2	4	11	23	131	1
T3	10	17	37	218	2
T4	11	25	41	289	1

Tabla 1: Características del corpus.

3.3 Método

La metodología que se presenta para describir i) la evaluación entre etapas escolares, ii) la evaluación de la clase y iii) la evaluación individual se divide en tres pasos principales: 1) generación de meta-resúmenes (modelos de referencia), 2) armonización e inclusión de resúmenes (si fuese necesario) y 3) medidas de evaluación adecuadas.

Se propone el siguiente algoritmo para generar meta-resúmenes, lo cual permite comparar la diferencia que pueda existir entre los resúmenes elaborados por estudiantes de primaria y los universitarios.

Sea $E = \{\boldsymbol{e_1}, \ldots, \boldsymbol{e_n}\}$ el conjunto de n resúmenes de referencia (*e.g.* los resúmenes de los universitarios); tal que cada resumen está codificado por una tupla de m elementos binarios representando la presencia/ausencia de las m unidades discursivas del documento original: $\boldsymbol{e_i} = (e_{i1}, \ldots, e_{im})$. Es decir, para cada unidad discursiva del documento original hay una entrada en la tupla que tendrá asignado un 1 cuando la unidad se preservó en el resumen y 0 cuando la unidad se eliminó en el resumen. Note que el orden de las oraciones no cambia sino que algunas aparecerán en el resumen y algunas otras no. Siguiendo esta representación se han codificado tanto los resúmenes de estudiantes de primaria como los de estudiantes universitarios. La idea general es codificar un solo meta-resumen de referencia a partir de los n resúmenes de referencia, en principio distintos, pero hacerlo de manera tal que el meta-resumen generado maximice el acuerdo entre las referencias y por lo tanto se pueda considerar como un resumen modelo.

El **Algoritmo 1**, de complejidad $\mathcal{O}(n)$, resuelve el problema de encontrar un subconjunto $E^* \subseteq E$ tal que el acuerdo entre sus elementos sea máximo; donde el acuerdo está determinado por el coeficiente kappa κ de Fleiss (2013). Al inicializar el algoritmo se incluye solamente la primera tupla de

E. Luego, en cada iteración se calcula el valor de κ que se produciría al incluir una nueva tupla de E en el subconjunto óptimo E^*; si el nuevo elemento aumenta el acuerdo entre las referencias, entonces es incluido en el subconjunto óptimo. Una vez realizadas todas las comparaciones, el resumen modelo de referencia se construye a partir de la moda de las entradas de las tuplas que hayan sido incluidas en E^*.

Algorithm 1 Algoritmo de generación de resúmenes modelo

procedure CREATEMODEL(E)
  $E^* \leftarrow \{\boldsymbol{e_1}\}$
  $\kappa* \leftarrow 0{,}0$
  for $\boldsymbol{e_i}$ in $E - \{\boldsymbol{e_1}\}$ **do**
    $\kappa \leftarrow FleissKappa(E^* \cup \{\boldsymbol{e_i}\})$
    if $\kappa \geq \kappa*$ **then**
      $\kappa \leftarrow \kappa*$
      $E^* \leftarrow E^* \cup \{\boldsymbol{e_i}\}$
    end if
  end for
  $model \leftarrow (mode(e_{ij})$ for $\boldsymbol{e_i} \in M; \boldsymbol{e_i} = (e_{i1}, \ldots, e_{im}))$
  return $model$
end procedure

4 Resultados

4.1 Evaluación entre distintas etapas escolares

Una característica del algoritmo es que el meta-resumen generado puede ser diferente dependiendo del orden de lectura de las tuplas $\boldsymbol{e_i}$. A mayor variabilidad entre los criterios de los alumnos, habrá una mayor variedad de meta-resumenes posibles. La Tabla 2 muestra que el algoritmo ha creado 63 meta-resúmenes con los resúmenes de primaria y 60 con los de la universidad. Se han creado más meta-resúmenes en primaria teniendo una cantidad sensiblemente inferior de resúmenes (352 de primaria y 684 de la universidad). En cuanto a los textos, el algoritmo ha creado 85 meta-resúmenes que corresponden al Texto-4, y solamente 36 con el Texto-2. Esto refleja que la longitud y la estructura discursiva del texto tienen mayor impacto (en cuanto a cantidad se refiere) al crear meta-resúmenes. Además, los meta-resúmenes de primaria se distribuyen de manera más equilibrada. Se ha creado un mínimo de 14 meta-resúmenes (Texto-1 y Texto-2), y un máximo de 18 (Texto-3). Sin embargo, con los resúmenes de los universitarios, la cantidad de los meta-resúmenes creados tiene una variación sensiblemente superior. Con el Texto-1 (17 EDUs) y Texto-2 (23 EDUs) se han creado menos que con el Texto-3 (37 EDUs) y Texto-4 (41 EDUs). Esto refleja que cuantos más EDUs tenga el texto, más meta-resúmenes creará el algoritmo con los resúmenes de los universitarios.

	T1	**T2**	**T3**	**T4**	**Total**
Primaria	17	14	18	14	63
Universidad 2018-2019	10	5	19	25	59
Universidad 2019-2020	7	8	18	22	55
Universidad 2018-2020	10	9	17	24	60
Total	44	36	72	85	

Tabla 2: Cantidad de meta-resumenes de resúmenes extractivos creados por el Algoritmo-1.

Para analizar y comparar los meta-resúmenes de primaria y universidad, el método se basa en las siguientes tres variables: i) cantidad de resúmenes en un meta-resumen, ii) acuerdo kappa obtenido y iii) puntuación obtenida en la coherencia (acuerdo entre la estructura retórica descrita por el profesor y la estructura retórica que se obtiene de la extracción hecha por el estudiante).

La coherencia fue calculada con el siguiente procedimiento:

- Paso-1: el texto se segmentó en EDUs y la unidad que contiene la idea principal (UC) fue etiquetada según Iruskieta, Diaz de Ilarraza, y Lersundi (2014). A continuación, el texto fue anotado manualmente según la Teoría de la Estructura Retórica (RST) (Mann y Thompson, 1987).
- Paso-2: Cada texto resumido se dividió en 4 cuartiles (Q1 a Q4) siguiendo la estructura jerárquica del árbol RST. Se calcula la distancia de cada EDU contando cuántas relaciones se necesitan para llegar a la UC de los resúmenes. La EDU más cercana está a la distancia 0 (la misma UC) y la EDU más lejana está a 6 relaciones de distancia. Pero, en algunos textos, la distancia más lejana es de 4 relaciones. Una vez obtenida la distancia de cada EDU, se clasifican las EDUs en cuartiles. Por ejemplo, el árbol RST con una distancia máxima de 6 es el siguiente: *i*) Q1 representa la UC y todas

las EDUs a la distancia 1 (distancia 0-25 %). *ii*) Q2 representa todas las EDUs a 2 y 3 relaciones de distancia, lo que significa que se saltan 2 o 3 relaciones desde la EDU en que se sitúa la UC (distancia 26-50 %). *iii*) Q3 representa todas las EDUs a 4 y 5 relaciones de distancia (distancia 51-75 %). *iv*) Q4 representa el resto de las EDUs a 6 y 7 relaciones de distancia (distancia 76-100 %).

- Paso-3: Puntuación de la coherencia. Por un lado, para representar y comparar la calidad de los meta-resúmenes, todas las EDUs son clasificadas en grupos según su distancia respecto a la UC. Después, se ha calculado el porcentaje de EDUs mantenidas en cada grupo, para poder aplicar las siguientes reglas: *i*) regla aplicada a EDUs de nivel 1 (distancia 0): si el porcentaje de las ideas mantenidas es igual a 100 %, entonces 1 punto, si no 0. *ii*) Regla aplicada a EDUs de nivel 2 (distancia 1): si el porcentaje de las ideas mantenidas es igual o menor al porcentaje de las ideas del nivel previo (siendo este superior al 0 %), entonces 1 punto, si no 0. *iii*) Regla aplicada al resto de niveles: si el porcentaje de las ideas mantenidas es igual o menor al porcentaje de las ideas del nivel previo y menor al porcentaje de las ideas del nivel anterior al previo, entonces 1 punto, si no 0.

  Por otro lado, cada cuartil se ha ponderado en base a su relevancia: *i*) 0,4 para la puntuación obtenida en Q1. *ii*) 0,3 para el Q2. *iii*) 0,2 para el Q3. *iv*) 0,1 para el Q4.

De esta forma, es posible representar y visualizar los resultados empleando gráficos de burbuja. Estos gráficos ayudan a comparar las distintas etapas escolares.

La Figura 1 muestra la calidad de los meta-resúmenes creados en ambas etapas escolares. La gran mayoría tiene una puntuación superior al 0,6, lo cual es comprensible. Por una parte, es normal que los estudiantes universitarios no tengan grandes dificultades para trabajar textos que pertenecen a libros de texto de primaria. Por otra parte, los alumnos de primaria han trabajado con estos textos a lo largo de toda la unidad didáctica, lo cual ha podido facilitar la acción de resumir.

Las burbujas que representan a los alumnos de la universidad son más grandes, lo cual indica que la cantidad de resúmenes que contienen es superior. Se podría pensar que los estudiantes de la universidad resumen de manera mucho más similar entre ellos si se comparan con los estudiantes de primaria. Sin embargo, se han analizado 88 estudiantes de primaria y 171 universitarios; por lo tanto, era de esperar que se generasen burbujas más grandes en la universidad.

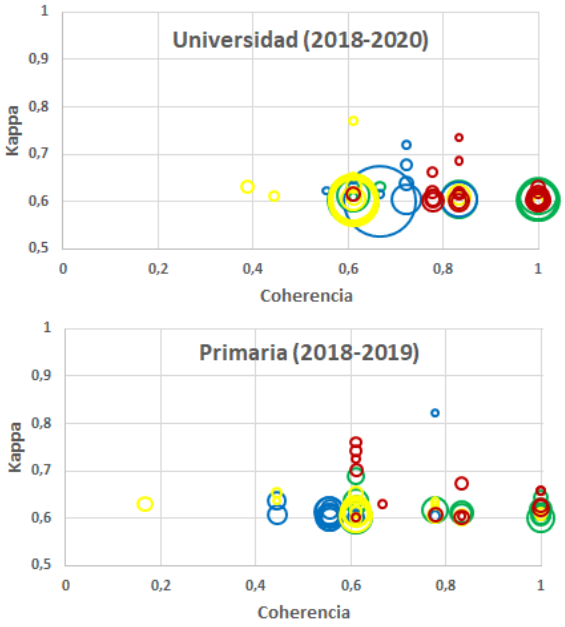

Figura 1: Cantidad y calidad de los meta-resúmenes creados por el algoritmo con estudiantes de universidad y primaria. El eje horizontal muestra la puntuación de coherencia (Coherencia) obtenida. El eje vertical muestra acuerdo entre los anotadores y resúmenes (Kappa). El tamaño de la burbuja refleja el número de resúmenes que contiene el meta-resumen. Las burbujas verdes representan los resultados del T1, azules T2, amarillas T3 y rojas T4.

Si se compara el tamaño de los meta-resúmenes creados en el Texto-2 (burbujas azules), la Figura 1 muestra que los universitarios tienen una manera muy concreta de resumir dicho texto. Este fenómeno no se da entre los alumnos de primaria. La gran cantidad de meta-resúmenes creados con el Texto-2 (burbujas azules), con un tamaño similar entre ellos, indica una gran diversidad a la hora de resumir este texto. Si se observa el eje de la coherencia, se puede intuir que los estudiantes (tanto en primaria como universidad)

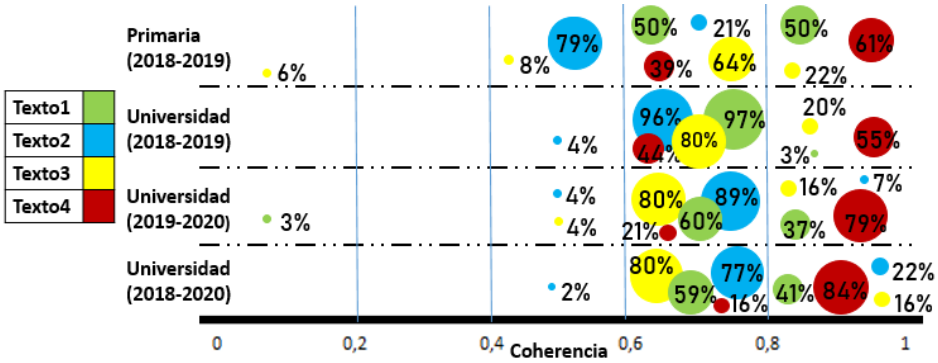

Figura 2: Calidad de los meta-resúmenes de los estudiantes de primaria y universidad. El eje horizontal muestra la puntuación de coherencia mientras que el eje vertical muestra los diferentes niveles escolares que se han analizado. Los tamaños de las burbujas muestran cuántos resúmenes hay en cada área de la cuadrícula.

han tenido mayor facilidad a la hora de resumir el Texto-1 y el Texto-4. Por el contrario, los meta-resúmenes creados con los textos 2 y 3 han obtenido una puntuación menor en general.

Para describir de manera más concisa esta diferencia, se utilizara el eje horizontal que muestra la Figura 2.

La Figura 2 indica que los estudiantes han tenido mayor facilidad con el Texto-1 y Texto-4, especialmente en este último; ya que todos los meta-resúmenes han logrado una puntuación de coherencia que se sitúa entre 0,6 y 0,8 o entre 0,8 y 1. En cambio, el algoritmo ha creado meta-resúmenes de menor calidad con los textos 2 y 3. En este caso, la diferencia entre los alumnos de universidad y primaria es considerable. En el Texto-3, la mayoría de los meta-resúmenes de los estudiantes tienen una puntuación de coherencia de 0,6 y 0,8, sin embargo, aparecen algunos con una puntuación más baja. Estos últimos pertenecen principalmente a los estudiantes de primaria, ya que como se muestra en la Figura 2 el 8 % de los estudiantes se ubica entre 0,4 y 0,6 y el 6 % entre 0 y 0,2. En cuanto al Texto-2, la mayoría de los estudiantes universitarios han obtenido una puntuación entre 0,6 y 0,8; sin embargo, la mayoría de los estudiantes de primaria se ubican entre 0,4 y 0,6. Esto podría deberse a que cuando los estudiantes resumen textos de poca complejidad, no existe una gran diferencia entre los resultados de los estudiantes de universidad y primaria. Pero cuando se trabaja con textos que son más difíciles de resumir, los estudiantes universitarios tienen más recursos para resumir el texto, a pesar de que los estudiantes de primaria hayan trabajado previamente estos textos.

Los estudiantes del curso 2019-2020 han logrado en general mejores resultados que los del curso 2018-2019, concretamente con los textos 1, 2 y 4. El ejemplo más evidente se da con el Texto-1. El 3 % de los meta-resúmenes pertenecientes al curso 2018-2019 han logrado una puntuación de coherencia entre 0,8 a 1; sin embargo, este porcentaje sube hasta el 37 % en el curso 2019-2020. Cuando se tienen en cuenta ambos cursos de la universidad, el porcentaje de estudiantes que se incluyen en un meta-resumen y obtienen una puntuación de coherencia de 0,8 a 1 sube hasta el 41 %. Este fenómeno se da en 3 de los 4 textos, lo que deja ver que cuanto mayor es el corpus mayor es la calidad de los meta-resúmenes en cuanto a la coherencia. Por lo cual, el sistema identifica más meta-resúmenes y meta-resúmenes mucho más fiables cuando hay más resúmenes. Sin embargo, la calidad es un factor fundamental. El sistema encontró mejores meta-resúmenes en el curso 2019-2020 con 82 resúmenes que en el 2018-2019 con 89 resúmenes. Esto parece indicar que ambos factores: el número de resúmenes y la calidad de los resúmenes son necesarios para crear meta-resúmenes fiables.

Aunque los resultados expuestos hasta el momento sirvan para detectar las diferencias principales entre alumnos de distintas etapas escolares, no se pueden utilizar estos datos para hacer una evaluación individual de los estudiantes y una evaluación colectiva de la clase. Las dos limitaciones que se exponen a continuación hacen que la evaluación deje a un lado algunos resúmenes (cobertura) y es

necesario que la precisión sea la mayor posible, para que la evaluación sirva en el proceso de aprendizaje en la escuela.

- Baja representación de los estudiantes: los meta-resúmenes tienen que tener un valor mínimo de 0,6 en kappa (valor establecido por los autores). Los resúmenes que no logren ese valor quedan fuera, y cabe la posibilidad de que un resumen no se incluya en ningún meta-resumen. Por lo tanto, si se quiere evaluar la clase en su conjunto con los meta-resúmenes, se deberá solventar esta limitación.

- El acuerdo entre la calidad del resumen y el meta-resúmenes: los meta-resúmenes fueron creados teniendo en cuenta primero el acuerdo de los segmentos del texto con kappa (medición cuantitativa) y después se calculó la puntuación de coherencia del meta-resumen (medición cualitativa). Por tanto, puede surgir el problema de que dos resúmenes distintos acaben estando en el mismo meta-resumen siendo cualitativamente muy diferentes entre sí (pudiendo ser que la única diferencia sea que uno contenga la idea más importante y el otro no). En ese caso, al menos uno de los dos resúmenes no está debidamente incluido en ese meta-resumen, en cuanto a la medida de coherencia. En consecuencia, es indispensable descartar los resúmenes que no se ajustan al meta-resumen.

4.2 Evaluación de la clase

Para la evaluación de toda una clase, el primer paso es cerciorarse de resolver las limitaciones mencionadas previamente. Para saber si un resumen está correctamente representado en un meta-resumen e incluir los resúmenes que habían quedado fuera de los meta-resúmenes, se utilizan la puntuación de coherencia (C) y la distancia de Hamming (H) para armonizar los meta-resúmenes: método C+H (criterios de coherencia y Hamming). Para incluir o mantener cualquier resumen en el meta-resumen, la distancia de coherencia (C) entre resumen y meta-resumen debe estar entre -0,1 y 0,1 (20 % de la puntuación) y la distancia de Hamming (H) debe ser inferior a 0,2 (20 % de la puntuación).

En la Figura 3 se presentan los meta-resúmenes y los resúmenes del Texto-1. El

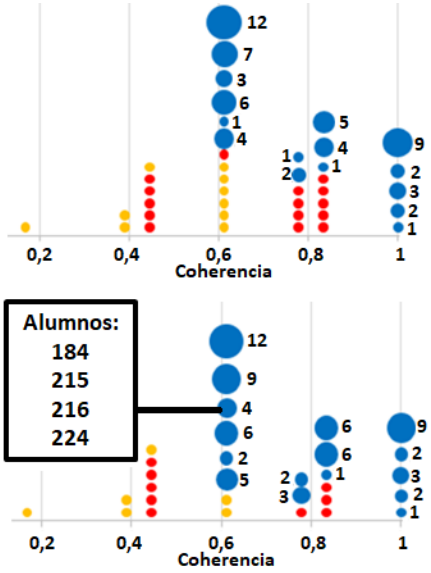

Figura 3: Evaluación del aula de primaria con el método de Coherencia y Hamming (C+H). Las burbujas azules muestran los meta-resúmenes creados por el algoritmo y la cantidad de resúmenes incluidos en ellos. Las burbujas amarillas representan cada resumen no incluido en ningún meta-resumen. Las burbujas rojas representan resúmenes incluidos en meta-resúmenes que no cumplen con los criterios del método C+H.

algoritmo ha creado 17 meta-resúmenes (burbujas azules), los cuales constituyen el 72 % de los resúmenes. El algoritmo no ha incluido 10 resúmenes (burbujas amarillas) para crear meta-resúmenes, es decir, el 11 % de los resúmenes había quedado fuera. Los últimos 15 resúmenes (burbujas rojas) (17 % de los resúmenes) según el método C + H, no estaban correctamente representados en los meta-resúmenes donde habían sido incluidos por el algoritmo.

Tras aplicar el método C+H, los resultados se muestran en la Figura 3. En esta figura se muestran los resúmenes no incluidos en los meta-resúmenes (burbujas amarillas), los resúmenes que se han mantenido en los meta-resúmenes después de aplicar el método C+H (burbujas azules) y los resúmenes que se han dejado fuera de los meta-resúmenes después de aplicar el método C+H (burbujas rojas). En el gráfico inferior se han añadido en los meta-resúmenes (burbujas azules), los resúmenes que anteriormente se habían dejado fuera (burbujas rojas y amarillas), pero que cumplen los criterios del método C+H. Al aplicarlo, el 40 % (4 de 10) de las burbujas

amarillas se han incluido en los meta-resúmenes. Por otra parte, el 60 % (9 de 15) de los resúmenes excluidos (burbujas rojas) se han incluido en otros meta-resúmenes.

La mayoría de los estudiantes está en torno al 0,6 en cuanto a la coherencia. Son 8 meta-resúmenes diferentes los que se sitúan en ese punto (0,6), conteniendo un total de 41 estudiantes (46 % de los estudiantes). Estos 41 estudiantes han mostrado una capacidad similar de resumir el Texto-1. Sin embargo, el docente tiene que tener en cuenta que aun habiendo logrado la misma puntuación, el sistema creó 8 meta-resúmenes diferentes. Esto significa que puede haber estudiantes que hayan obtenido la misma nota, pero que hayan utilizado estrategias totalmente diferentes para resumir. Por tanto, el docente deberá analizar cuáles son esas estrategias y cómo trabajarlas pudiendo dar un *feedback* personalizado.

Además del ya mencionado grupo de alumnos, los resultados muestran otros tres grupos: i) algunos de estos estudiantes (17 de 88) han alcanzado la máxima puntuación C. ii) Otros (16 de 88) están alrededor de 0,8 C. Se puede decir que todos estos estudiantes han resumido fácilmente el texto y han obtenido buenos resultados. iii) Sin embargo, hay otro grupo (8 de 88) que obtuvo una C más baja, alrededor de 0,4. Estos estudiantes han tenido más dificultades para resumir el Texto-1. iv) Para concluir, un estudiante ha obtenido una puntuación C (0,2) muy baja, por lo que el docente podría preocuparse.

Este tipo de gráficos ofrecen al docente una visión general para analizar en qué grupo se encuentra cada alumno y esto puede ser muy útil para programar actividades futuras. Por ejemplo, el profesor puede saber qué alumno puede ayudar a otros y quién necesita ayuda, o quién tiene que trabajar con textos más simples o más complejos. Además, puede indicarle qué resúmenes puede utilizar con los alumnos como modelo a seguir o cuales pueden servirle para emplearlos como *feedback*.

4.3 Evaluación automática de los estudiantes

El algoritmo creó 17 meta-resúmenes, de los cuales hay que decidir cuales utilizar para la evaluación de cada estudiante. Para decidirlo, se han seguido los dos siguientes criterios: i) el número de resúmenes contenidos en el meta-resumen es al menos el 10 % total de los resúmenes y ii) el modelo alcanza al menos una puntuación de 0,5 (C). Los meta-resúmenes A, B, C, D, E y F (ver figura 4) cumplen los criterios ya mencionados, por lo cual, se emplearan para evaluar a los estudiantes de manera automática.

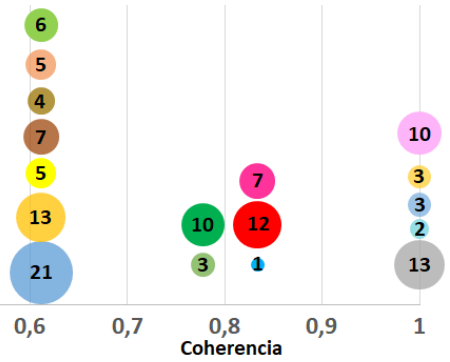

Figura 4: Meta-resúmenes utilizados para la evaluación de estudiantes de primaria. El eje horizontal muestra la puntuación de la coherencia. El tamaño de las burbujas muestra cuántos resúmenes hay en cada meta-resumen. Los números de cada burbuja muestran la cantidad de resúmenes que contiene el meta-resumen.

Cada resumen es evaluado frente a los 6 meta-resúmenes seleccionados previamente mediante kappa (para calcular el acuerdo) y Fisher (para calcular la probabilidad con la que el resumen haya sido hecho al azar). En cuanto al acuerdo, un valor Kappa superior a 0,6 se considera óptimo. Por otra parte, el resumen extractivo es una acción subjetiva, por tanto, se considerara que se ha realizado un resumen de forma aleatoria si obtiene un valor de Fisher de 0,8 o superior.

La Tabla 3 muestra los resultados obtenidos por los estudiantes 265, 206 y 214 después de evaluar automáticamente sus resúmenes con los seis meta-resúmenes seleccionados.

		Modelos					
Est.		**A**	**B**	**C**	**D**	**E**	**F**
265	K	0,26	0,33	0,10	0,46	0,84	0,72
	F	0,29	0,23	1,00	0,12	0,02	1,00
206	K	0,33	1,00	0,45	0,45	0,26	0,20
	F	0,23	0,59	0,17	0,17	0,29	0,51
214	K	-0,18	-0,08	-0,16	-0,16	-0,20	-0,21
	F	1,00	1,00	1,00	1,00	1,00	0,09

Tabla 3: Evaluación de los estudiantes 265, 206 y 214 mediante la comparación con los 6 meta-resúmenes utilizando Kappa y Fisher.

El resumen 265 ha logrado un gran acuerdo Kappa con los meta-resúmenes E (0,84) y

F (0,72). Es una buena señal, ya que el resumen 265 se ha incluido previamente en ambos meta-resúmenes. Además, si se calcula la coherencia del resumen 265, la puntuación es 1, igual que los dos meta-resúmenes (ver Figura 4). De ahí que, aunque el resumen sea diferente al de los dos meta-resúmenes, tienen la misma calidad. Sin embargo, se debe tener en cuenta la métrica estadística de Fisher, para saber si el resumen se ha hecho al azar. En el Meta-resumen-E el valor de Fisher es de 0,02 (inferior a 0,8), lo cual indica que el resumen no se hizo al azar.

El resumen 206 es idéntico al Meta-resumen-B, ya que han alcanzado la máxima puntuación en Kappa y, además, el valor de Fisher es bajo (0,59). El Meta-resumen-B tiene una puntuación de coherencia de 0,61 (consulte la Tabla 3); por lo tanto, el resumen 206 también debería alcanzar esa puntuación, ya que son iguales. Cabe señalar que el resumen 206 se ha incluido previamente en el Meta-resumen-B, pero no en el A. Los dos meta-resúmenes obtuvieron la misma puntuación de coherencia, esto indica que los meta-resúmenes A y B tienen la misma calidad pero son bastante diferentes.

Al evaluar el resumen del estudiante 214, (Tabla 3), el acuerdo alcanzado con los meta-resúmenes ha sido prácticamente nulo. Además, los valores de Fisher indican una alta probabilidad de haber hecho el resumen al azar. A pesar de todo, es importante recordar lo siguiente. Cuando un resumen logra un gran acuerdo con un meta-resumen, se deduce que el resumen es bueno. Sin embargo, cuando el acuerdo es bajo, no es posible garantizar que sea malo. En este caso, se calculó la puntuación del resumen 214 manualmente (0,16 C), para comprobar que el resumen es de baja calidad. En futuros trabajos será necesario incluir como modelos de referencia meta-resúmenes de baja calidad.

5 Conclusiones

En este artículo hemos presentado un método para la evaluación automática de resúmenes por extracción en euskera, aunque el sistema y él método podrían usarse en otras lenguas con una mínima adecuación. En el método se hace uso del crowd-sourcing para crear meta-resúmenes de referencia y se ha utilizado para realizar tres tipos de evaluación: i) diferencias entre distintas etapas escolares, ii) evaluación colectiva del aula y iii) evaluación automática individual de cada estudiante usando meta-resúmenes fiables. En los experimentos se ha trabajado con alumnos reales y utilizado sus textos escolares, es decir, textos que son utilizados para trabajar contenidos curriculares en la escuela en tiempo real. Se ha propuesto el método C+H, permitiendo evaluar el acuerdo cualitativo (coherencia) y cuantitativo (Hamming) entre resúmenes y meta-resúmenes.

Los resultados demuestran que la evaluación propuesta es viable y robusta para aplicarse a mayor escala; lo cual, nos permitirá seguir trabajando en un contexto real pero con intervenciones más largas y sistemáticas en el tiempo, además de incluir diferente tipología de textos para dar el salto a los resúmenes de abstracción.

En futuros trabajos sería interesante analizar las relaciones discursivas para después poder ponderarlas. De esta forma, se podría desarrollar un algoritmo que cree meta-resúmenes basándose en un acuerdo cuantitativo y cualitativo. También será importante trabajar con más documentos pues permitiría analizar qué tipos de textos son más fáciles de resumir para los estudiantes y qué texto es conveniente para desarrollar las habilidades que requiere el resumen. Para ello, puede resultar de gran utilidad analizar tanto la distribución de la información más importante en el texto como la distribución de las relaciones de coherencia.

Agradecimientos

El trabajo de Unai Atutxa está financiado por una beca de doctorado (PIF18/118) de la Universidad del País Vasco (UPV/EHU).

Bibliografía

Álvarez, I. A. 2004. Evaluación y calificación de resúmenes de textos expositivos en el aula de ile/ife: la guía"babar". *Ibérica: Revista de la Asociación Europea de Lenguas para Fines Específicos (AELFE)*, 1(8):81–99.

Atutxa, U. 2018. Ikasleen laburpen-corpusa eta laburpen-gaitasunaren ebaluazioa: oinarri metodologikoak. Master's thesis, University of the Basque Country (UPV/EHU), Donostia.

Atutxa, U., M. Iruskieta, O. Ansa, y A. Molina. 2017. Compress-eus: I (ra) kasleen laburpenak lortzeko tresna. *EU-*

DIA: Euskararen bariazioa eta bariazioaren irakaskuntza-III, páginas 87–98.

Cabrera-Diego, L. A. y J.-M. Torres-Moreno. 2018. Summtriver: A new trivergent model to evaluate summaries automatically without human references. *Data & Knowledge Engineering*, 113:184 – 197.

CY, L. 2004. Rouge: a package for automatic evaluation of summaries. En *Proceedings of the Workshop on Text Summarization Branches Out. Barcelona, Spain*, páginas 56–60.

Fleiss, J. L., B. Levin, y M. C. Paik. 2013. *Statistical methods for rates and proportions*. john wiley & sons.

Iruskieta, M., A. D. Diaz de Ilarraza, y M. Lersundi. 2014. The annotation of the central unit in rhetorical structure trees: A key step in annotating rhetorical relations. En *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, páginas 466–475.

Louis, A. y A. Nenkova. 2009. Automatically evaluating content selection in summarization without human models. En *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, páginas 306–314. Association for Computational Linguistics.

Mann, W. C. y S. A. Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

Molina, A. y J.-M. Torres. 2015. El test de turing para la evaluación de resumen automático de texto. *Linguamática*, 7(2):45–55.

Molina-Villegas, A. 2013a. Compresión automática de frases: un estudio hacia la generación de resúmenes en español. *Inteligencia Artificial*, 16(51):41–62.

Molina-Villegas, A. 2013b. Sistemas web colaborativos para la recopilación de datos bajo el paradigma de ciencia ciudadana. *Komputer Sapiens*, 1(5):6–18.

Nenkova, A. y R. J. Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. En *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004*, páginas 145–152.

Papineni, K., S. Roukos, T. Ward, y W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. En *Proceedings of the 40th annual meeting on association for computational linguistics*, páginas 311–318. Association for Computational Linguistics.

Radev, D. R., H. Jing, M. Styś, y D. Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938.

Radev, D. R. y D. Tam. 2003. Summarization evaluation using relative utility. En *Proceedings of the twelfth international conference on Information and knowledge management*, páginas 508–511.

Saggion, H. 2008. A robust and adaptable summarization tool. *Traitement Automatique des Langues*, 49(2).

Saggion, H. y T. Poibeau. 2013. Automatic text summarization: Past, present and future. En *Multi-source, multilingual information extraction and summarization*. Springer, páginas 3–21.

Saggion, H., J.-M. Torres-Moreno, I. d. Cunha, y E. SanJuan. 2010. Multilingual summarization evaluation without human models. En *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, páginas 1059–1067. Association for Computational Linguistics.

Sanz, A. 2005. Irakurmena lantzeko jarduerak nola prestatu: Lehen hezkuntzako 3. zikloa eta dbhko 1. zikloa. *Nafarroako Gobernua*.

Zipitria, I., A. Arruarte, y J. A. Elorriaga. 2008. Lea: A summarization web environment based on human instructors' behaviour. En *2008 Eighth IEEE International Conference on Advanced Learning Technologies*, páginas 564–568. IEEE.

Zipitria, I., P. Larrañaga, R. Armañanzas, A. Arruarte, y J. A. Elorriaga. 2008. What is behind a summary-evaluation decision? *Behavior Research Methods*, 40(2):597–612.

Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 177-190 recibido 13-12-2020 revisado 24-01-2021 aceptado 02-02-2021

Discovering topics in Twitter about the COVID-19 outbreak in Spain

Descubriendo temas en Twitter sobre el brote del COVID-19 en España

Marvin M. Agüero-Torales[1], David Vilares[2], Antonio G. López-Herrera[1]
[1]University of Granada, Spain
[2]Universidade da Coruña, CITIC, Spain
maguero@correo.ugr.es, david.vilares@udc.es, lopez-herrera@decsai.ugr.es

Abstract: In this work, we apply topic modeling to study what users have been discussing in Twitter during the beginning of the COVID-19 pandemic. More particularly, we explore the period of time that includes three differentiated phases of the COVID-19 crisis in Spain: the pre-crisis time, the outbreak, and the beginning of the lockdown. To do so, we first collect a large corpus of Spanish tweets and clean them. Then, we cluster the tweets into topics using a Latent Dirichlet Allocation model, and define generative and discriminative routes to later extract the most relevant keywords and sentences for each topic. Finally, we provide an exhaustive qualitative analysis about how such topics correspond to the situation in Spain at different stages of the crisis.
Keywords: COVID-19, Twitter, social networks, topic modeling.

Resumen: En este trabajo, analizamos lo que los usuarios han estado discutiendo en Twitter durante el comienzo de la pandemia causada por el COVID-19. Concretamente, analizamos tres fases diferenciadas de la crisis del COVID-19 en España: el propio tiempo de pre-crisis, el estallido de la enfermedad y el confinamiento. Para llevar esto a cabo, primero recolectamos una gran cantidad de tuits que son preprocesados. A continuación, agrupamos los tuits en distintas temáticas usando un modelo de Latent Dirichlet Allocation, y definimos estrategias generativas y discriminativas para extraer las palabras clave y oraciones más representativas para cada tema. Finalmente, incluimos un exhaustivo análisis cualitativo sobre dichos temas, y cómo estos se corresponden con distintas problemáticas surgidas en España en distintos momentos de la crisis.
Palabras clave: COVID-19, Twitter, redes sociales, modelado de temas.

1 Introduction

The outbreak of the SARS-CoV-2 virus and the global spread of the COVID-19 disease has encouraged people and organizations to express their opinion, discuss topics and warn about the evolution of the pandemic in social media platforms such as Twitter.

Unlike previous occasions, such as SARS-CoV in 2002 (World Health Organization (WHO), 2020b), where social media still were in an embryonic state and natural language processing (NLP) still had limited practical applications; we are now in a situation where users generate a vast amount of written content, that can be analyzed by automatic tools to discover the topics societies care about, and their sentiment. This has been already the case for some precedent events or catastrophes in recent years, such as the 2016 US political elections (Grover et al., 2019) or some natural disasters, such as the 2011 East Japan Earthquake (Neubig et al., 2011).

In relation to the COVID-19 pandemic, a few specific NLP workshops (Verspoor et al., 2020b; Verspoor et al., 2020a) have already attempted to highlight how NLP can be used to respond to situations like the current one; addressing a number of challenges that include mining scientific literature and social media analysis, among many others (Wang et al., 2020; Kleinberg, van der Vegt, and Mozes, 2020; Afzal et al., 2020). With research purposes, there has been also efforts on releasing NLP datasets discussing COVID-19 topics (Chen, Lerman, and Fer-

ISSN 1989-7553. DOI 10.26342/2021-66-15

rara, 2020; Banda et al., 2020; Kerchner and Wrubel, 2020). In this context, the area of topic modeling has not been a stranger to this problem, and a number of authors have showed the options that clustering online posts such as tweets or Facebook messages can offer to monitor and evaluate the evolution of the pandemic through time (Asgari-Chenaghlu, Nikzad-Khasmakhi, and Minaee, 2020; Yin, Yang, and Li, 2020; Amara, Taieb, and Aouicha, 2020).

Contribution In this work, we also focus on the possibilities of performing effective and representative topic modeling over a large set of Spanish tweets. More particularly, we first collect a few millions tweets about COVID-19, mostly between 1 January to 20 April of 2020. Then, we apply latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan, 2003) to compute relevant topics in an unsupervised way, and obtain meaningful keywords and sentences through generative and discriminative routes. Finally, we provide an analysis to shed some light about the quality of the extracted topics, and how faithfully they represent what was happening in the Spanish society at different moments of the pandemic.

2 Related work

In what follows, we review topic modeling and NLP research related to COVID-19.

2.1 Topic modeling

In topic modeling, a topic is often viewed as a pattern of co-occurring words that can be exploited to cluster together documents from a large collection (Barde and Bainwad, 2017). Among methods for topic modeling we can find approaches such as the Vector Space Model (VSM) (Salton, Wong, and Yang, 1975), Latent Semantic Indexing (LSI) (Deerwester, 1988), Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) or lda2vec (Moody, 2016). Related to this, one of the most well-known, standardized and widely-used methods is Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan, 2003). More particularly, LDA is an unsupervised clustering approach where documents can belong to multiple topics, and where each topic is a mix of words, which can be shared among topics too.

The applications of these topic modeling approaches are many and include areas such as tag recommendation (Tuarob, Pouchard, and Giles, 2013), text categorization (Zhou, Li, and Liu, 2009), keyword extraction (Yijun and Tian, 2014), information filtering (Gao, Xu, and Li, 2014), similarity search in the fields of text mining (Pham, Do, and Ta, 2018), and information retrieval (Andrzejewski and Buttler, 2011).

2.2 Text Mining on English COVID-19 related tweets

With the COVID-19 outbreak, different authors have tried to apply topic modeling and text mining techniques to help analyze and monitor the situation of the pandemic, with a great focus on English messages. For instance, Asgari-Chenaghlu, Nikzad-Khasmakhi, and Minaee (2020) analyzed English tweets and detected the trending topics and major concerns of people with respect to COVID-19, by proposing a model based on the Universal Sentence Encoder (Cer et al., 2018). The model first derives a semantic representation and similarity of tweets and, over those similar tweets, it applies text summarization techniques to provide a summary of different clusters. In a related line, Yin, Yang, and Li (2020) proposed a framework to analyze the topic and sentiment changes in society over time due to the COVID-19, using Twitter to collect the source data. More specifically, they used a dynamic LDA for topic modeling over fixed time intervals (Blei and Lafferty, 2006), and VADER for sentiment analysis (Hutto and Gilbert, 2014). Chandrasekaran et al. (2020) examined the key topics among 13.9M English tweets about COVID-19, dealing with areas such as economy and markets, spread and growth in cases, treatment and recovery, impact on the healthcare sector, and governments response. They explored the trends and variations, and how those key topics, and associated sentiments changed over a period of time of 17 weeks, between 1 January 2020 and 9 May 2020. More particularly, they used guided LDA for topic modeling (Jagarlamudi, Daumé III, and Udupa, 2012), an LDA-variant where the model is guided to learn topics that are of specific interest, using priors in the form of seed words, and again VADER for sentiment analysis.

Also, Abd-Alrazaq et al. (2020) use LDA to detect topics such as the origin of the virus and its impact on people and coun-

tries, analyzing 2.8M English tweets. In addition, they performed sentiment analysis with `TextBlob` (Loria, 2020) and extracted some social network statistics for each topic, such as the number of followers, the number of likes of tweets, the number of retweets, the user mentions, or the link sharing, calculating the interaction rate per topic. At a smaller scale (100K English tweets) and considering only the pre-crisis lockdown period (from 12 December 2019 to 9 March 2020); Boon-Itt (2020) presented a work to understand public perceptions of the trends of the COVID-19 pre-pandemic time. The analysis included time series, sentiment analysis and emotional tendency using the NRC sentiment lexicon (Mohammad and Turney, 2013), as well as topic modeling using LDA.

2.3 Text Mining on Spanish and Multilingual COVID-19 related tweets

As usual in NLP, most of early efforts to monitor COVID-19 user-generated texts have focused on English. However, some work is already available for the Spanish language. For instance, Yu, Lu, and Muñoz-Justicia (2020) compare the news updates of two of the main Spanish newspapers Twitter accounts, *El País* and *El Mundo*, during the pandemic; applying topic modeling and network analysis methods. They identified eight news frames for each newspaper and split it in three clusters: the pre-crisis period (from 19 February to 14 March of 2020), the lockdown period (from 14 March to 11 May of 2020) and the recovery period (from 11 May to 3 June of 2020). Their goal was to understand how the Spanish news media covered the public health crisis in Twitter.

Besides, Carbonell Gironés (2020) proposed a geographical analysis of the opinion and influence of users in Twitter during the covid health crisis, considering tweets written in English and Spanish, and using LDA topic modeling. The first part of the study was a general approach to the analysis of the topics of US and UK users. The second part was an analysis of the interests of Twitter users in Spain during the confinement period (from 14 March to 22 July of 2020). To geolocate the tweets, they performed a country-level search for the English dataset, and a city or province-level search for the Spanish dataset; looking in both cases for any geographic references, both on the Twitter user location field and their biography.

Ordun, Purushotham, and Raff (2020) studied techniques to assess the distinctiveness of topics, key terms and features, as well as the speed of dissemination of retweets over time. They used pattern matching and topic modeling with LDA on a set of 5.5M of tweets written in multiple languages, resulting in 16 topics for English and one for Spanish, Italian, French and Portuguese, respectively. They also applied Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy, and Melville, 2018) to identify clusters of distinct topics, which discuss case spread, healthcare workers, and personal protective equipment issues.

Beyond Twitter, Amara, Taieb, and Aouicha (2020) have exploited 22K Facebook posts to track the evolution of COVID-19 related trends, with a multilingual dataset that covers seven languages (English, Arabic, Spanish, Italian, German, French and Japanese). They applied an end-to-end analytic process for discovering language-dependent topics covering the duration of the pre-crisis period and part of lockdown (from 1 January to 15 May of 2020). The experiments showed that the extracted topics corresponded to the chronological development of what has been happening, and the measures that were taken in various countries.

3 Methods

In what follows, we describe the methodology of our work, decomposed into four steps: (i) the collection of the corpus, (ii) the language identification and geolocation of the tweets, (iii) the preprocessing, and (iv) the topic modeling approach and its analysis, clustering tweets into topics and extracting representative keywords and sentences.

3.1 Collection of tweets

We first defined a set of keywords to download relevant tweets: *coronavirus*, *COVID-19*, *COVID19*, *2019-nCoV*, *2019nCoV*. Further, as of March 3th, 2020 we added more keys: *SARS-CoV-2*, *SARSCoV2*, *CoV-19*, *CoV19*, *COVD19*, *COVD 19*, *corona virus*, *corona outbreak*.

More particularly, we collected a multilingual corpus of 32.68M tweets, including Twitter posts from 1 January of 2019[1] to 20

[1]In order to have some preceding context, but ex-

April of 2020; from all over the world. We scraped the tweets using the GetOldTweets-python3 (GOT3) library.[2] The reason to use this tool was that it allowed to retrieve old tweets without time limitation. However, the tool did not permit us to filter the retrieval by language. Besides, the Twitter Official API cannot retrieve tweets more than a week ago with a free subscription mode.[3]

3.2 Language identification and geolocation

The next step is language identification to keep only the Spanish tweets. We used four tools for detecting languages, since with GOT3 we could not obtain the language attribute. Those four tools were: `polyglot`,[4] `langdetect`,[5] `langid.py`,[6] and `fastText`.[7] The language is assigned based on majority voting. In case of a tie, we consider the tweet to be Spanish, except if all tools predicted a different language.

In total, we identified 5.35M Spanish tweets. In this work, we try to restrict the analysis to the content generated in Spain. For this purpose, we proceeded to filter the tweets in Spanish using the location attribute of the user profile, and look for the name of Spanish cities with more than 50K inhabitants, province names, autonomous regions names, and also any location specified as simply 'Spain'.[8]

After the cleaning process, we obtained ∼1.85M tweets for our topic modeling analysis. It is fair to point out that there is a percentage of tweets with a risk of not being correctly filtered, since the same place name might exist in more than one Spanish speaking country (e.g., 'Guadalajara' for Spain vs. 'Guadalajara' for Mexico). This is a common limitation on Twitter analyses, when it comes to analyze geolocated tweets (see for instance (Vilares and Gómez-Rodríguez, 2018)).

3.3 Preprocessing

We first proceed to lowercase the tweets and remove retweets. We also delete the keywords that were used to collect the tweets (see again §3.1) and other Twitter reserved words such as 'rt', 'fav', 'vía', 'nofollow', 'twitter', 'href' or 'rel'. Moreover, we removed stopwords, non-words (i.e., words compounded with characters that are not alphabet letters), URLs, numbers and punctuation marks. To do this, we used spaCy[9] to tokenize the words, and the Spanish and English stopwords lists from three libraries: NLTK,[10] stop-words,[11] and stopwordsiso.[12] Besides, in order to remove extra noise and cluster more clean topics, we only kept content words (i.e., nouns, verbs, adjectives, and adverbs).

Finally, to reduce word sparsity we used a custom lemmatizer[13] for Spanish, which applies a rule-based lemmatization with spaCy, and relies on Wiktionary,[14] which is a collaborative free-content multilingual dictionary. After the lemmatization step, the tweets whose length is less than three characters were removed. As traditional topic modeling approaches such as LDA, based on bag-of-words, suffer if many outliers are present (which happens in NLP due to the Zipf's law), we ignore terms that have a corpus frequency strictly less than three.

3.4 Topic modeling

For a more clear and comprehensive topic modeling analysis, we cluster the tweets in four weeks per month, except for the year 2019 (for which we collect the few tweets discussing coronavirus topics at that time), and the month of January 2020, which covers the first fortnight and not a week.

More particularly, we cluster the time of analysis into three phases. First, a pre-crisis phase, which includes tweets up to 24 January of 2020; when there was still few cases

pecting just to be able to retrieve a small number of tweets.

[2] `https://github.com/Jefferson-Henrique/GetOldTweets-python`

[3] However, we noted that GOT3 as of 18 September 2020 has been suspended due to the new Twitter policies on tweet payload

[4] `https://polyglot.readthedocs.io/en/latest/Detection.html`

[5] `https://pypi.org/project/langdetect/`

[6] `https://pypi.org/project/langid/`

[7] `https://fasttext.cc/docs/en/language-identification.html`. We used the large model

[8] We obtained the list of place names from the Instituto Nacional de Estadística (INE) `https://www.ine.es/dynt3/inebase/es/index.htm?padre=517&capsel=525`

[9] `https://spacy.io/usage/v2-2` and the es_core_news_md language model

[10] `http://www.nltk.org/nltk_data/`

[11] `https://pypi.org/project/stop-words/`

[12] `https://pypi.org/project/stopwordsiso/`

[13] `https://github.com/pablodms/spacy-spanish-lemmatizer`

[14] `https://www.wiktionary.org/`

reported outside China. Second, we consider the outbreak phase, that we will consider to range from 25 January to 14 March of 2020; when the disease started to widely spread across Europe and the rest of the world, but Spain still was not under confinement. This is the period of time where the pandemic information, epidemic back then, was reported but was still not formally considered an alarm by the Spain government. Third, we cover about a month of the official lockdown period of the first wave (from 14 March to 20 April of 2020), when the Spanish government approved a strict social confinement.

As introduced previously, for topic modeling, we will be using *Latent Dirichlet Allocation* (LDA)[15] with collapsed Gibbs sampling inference (Griffiths and Steyvers, 2004); which processes raw text data in an unsupervised fashion to cluster documents that discuss the same topic. We chose LDA because it is standard and has probed robust for many tasks (see also §2.2 and §2.3). For each phase, we will mostly group tweets into weeks,[16] and for each week we will be extracting 10 topics. On the one hand, our goal was to facilitate the comprehension and interpretability. On the other hand, it is worth to note that selecting too few topics would make the clusters very generic and unspecific, while choosing too many could make them too sparse, not representative, and hard to analyze qualitative (Steinskog, Therkelsen, and Gambäck, 2017). Yet, we explored what would be in theory an optimal number of topics for different weeks using three methods: (i) the KL divergence (Arun et al., 2010), (ii) the pairwise cosine distance (Cao et al., 2009), (iii) and the loglikelihood. In all cases the results returned that the ideal number was between 5 and 20 in most of cases.

LDA setup We sampled up to 1500 epochs, and we kept the rest of parameters to the default value in the LDA library we used, i.e., $\alpha : 0.1$, $\eta : 0.01$, where the first corresponds to the Dirichlet parameter for the distribution over topics, and the second to the Dirichlet parameter for the distribution over words.

[15]In particular, we rely on the `https://github.com/lda-project/lda` implementation

[16]As introduced before, we use week here in an informal sense, referring to periods of time of 7 days, but not necessarily from Monday to Sunday.

3.5 Extracting top topic keywords and sentences

To extract the most representative keywords for each topic, we considered both generative (GS) (Equation 1) and discriminative (DS) (Equation 2) approaches:

$$\mathrm{GS(w,z)} = P(w|z) \quad (1)$$

$$\mathrm{DS}(w, z) = P(w|z)/[\max_{z' \neq z} P(w|z')] \quad (2)$$

where w represents a given word and z the topic at hand. In essence, the generative score allows to extract the words that are most representative for each topic independently, in a way that a given word could be relevant for one or more topics, potentially making such topics harder to differentiate among them. On the contrary, the discriminative score allows to represent a topic by a set of keywords that are very representative for such topic, but have little relevance for the remaining ones.

Although the top keywords for each topic are useful, they might provide a limited view of what is actually being discussed. To counteract this, we also defined a generative (Equation 3) and discriminative (Equation 4) routes to extract the most representative sentences (tweets) for each topic, ideally being able to determine the topic by simply reading a few documents. The motivation to define these two different routes is the same than the one we made to extract the top keywords.

$$\mathrm{GS}_{\mathrm{sent}}(s, z) = \sum_{w \in s} \mathrm{GS}(w, z)/\mathrm{Length}(s) \quad (3)$$

$$\mathrm{DS}_{\mathrm{sent}}(s, z) = \sum_{w \in s} \mathrm{DS}(w, z)/\mathrm{Length}(s) \quad (4)$$

where s is the input document, for which we consider its length, in order not to only select the longest documents; although in the case of Twitter this is less of an issue than in other topic modeling approaches that must deal with actual long documents.

The full code is available.[17]

Limitations Sociolinguistic studies that collect data from social media such as Twitter can suffer from biases that can be hard to measure, identify or correct. For instance, it is well-known that a small percentage of

[17]`https://github.com/mmaguero/twitter-analysis`

Topic	Discriminative Keywords	Generative Keywords
'W1' (from January to December of 2019)		
2	respiratorio, enfermedad, gripe	respiratorio, gripe, enfermedad
Magnifica guia para diferenciar los sintomas que causa la gripe y otros virus respiratorios. Junto con la gripe siguen circulando rinovirus, virus respiratorio sincitial y coronavirus, entre otros. <URL>		
1	enfermedad, gripe, respiratorio	enfermedad, respiratorio, gripe
@user informa de 27 casos de neumonia atipica, probablemente virica, en Wuhan (Hubei, China) en fecha 31/12/2019. El SARS (coronavirus) se inicio asi en 2003. Habra que seguir evolucion y esperar el diagnostico. <URL>		
W2-3 (from 1 to 16 January of 2020)		
8	alerta, hospital, poner, red, oms, china, mundial, mundo	china, oms, alerta, hospital, poner, mundial, mundo, red
UN NUEVO CORONAVIRUS PONE EN ALERTA A CHINA <URL>vía @user		
5	confirmar, japón, chino, infección, caso, china, animal, aparición	caso, confirmar, japón, china, infección, chino, ciudad, identificar
Japón confirma el primer caso de coronavirus vía @user <URL>		
W4 (from 17 to 24 January of 2020)		
9	emergencia, declaración, declarar, organización, reunión, convocar, decisión, determinar	oms, emergencia, internacional, declarar, mundial, alerta, salud, china
La OMS no declaró la emergencia por el coronavirus <URL>		
1	millón, cuarentena, habitante, frenar, ampliar, pekín, transporte, aislar	china, ciudad, wuhan, millón, cuarentena, persona, cerrar, brote
Más de once millones de chinos, en cuarentena por el coronavirus <URL>		

Table 1: Some representative topics for the weeks corresponding to the **pre-crisis** period of the COVID-19 pandemic in Spain. For each example topic, we include the top representative sentence according to its discriminative score.

Twitter users generate the majority of content (Wojcik and Hughes, 2019). In this line, we believe that many of the collected tweets have its origin in newspapers and journalists accounts, that condition how other users tweet about this topic on Twitter, and therefore the detected topics can be heavily dependent on how national media decide to spread the news. Yet, this is the natural behaviour of this network, and in this particular work we decided not to control for this variable.

4 Results

We consider sixteen sets of tweets (mostly grouped in a weekly basis), extracting the ten most representative topics for each one according to LDA. To refer the topics, we will represent them with the top eight keywords and the most salient tweets. For clarity, and due to the large amount of weeks and topics, we will just illustrate and analyze some relevant topics extracted by our approach for different weeks, and try not to repeat common topics that span through the whole period. Usernames and urls are cut due to anonymity and space reasons, respectively.

4.1 Pre-crisis time

During this pre-crisis time, it is possible to see how the model captures that the COVID-19 was still not a concern for the Spanish society, which perceived the disease as an external problem, as reflected in many of the extracted topics. For clarity, Table 1 illustrates some relevant topics with top keywords and tweets, but we briefly discuss the content of the table below. To assess the relevance of the topics, we will be matching those against news from the newspapers that were published at the time in different Spanish media.

'W1' (from January to December of 2019) For the year 2019, we only could extract a total of seven topics, since the corresponding subset of tweets related to COVID-19 or coronavirus was still tiny (a total of 43 tweets after preprocessing). Still, we believe the results are interesting, since we observed that at this time most of Spanish tweets dealing with coronavirus still had to do with veterinarian diseases or even the zoonosis of coronavirus (i.e., how it is transmitted between animals and humans through the air), Yet, we found a few relevant tweets about COVID-19 that started to show up. We illustrate this as part of Table 1.

W2-3 (from 1 to 16 January of 2020) This time can be considered as the start of the emergency (Agencia EFE, 2020). In this line, we observed how our model started to identify this situation as well, clustering tweets about the World Health Organization (WHO) alerts to hospitals about symptoms, procedures, etc., and also about the increase

in the number of cases in China.

W4 (from 17 to 24 January 2020) The crisis started to expand and from our model we see how the topics differ from previous weeks (see the third group of rows in Table 1). For instance, it shows how China started to apply restrictions in many locations of its territory (e.g., Wuhan) (El Boletín, 2020).

4.2 Outbreak time

In this phase, we see how the LDA approach reflects emergency declarations, the first cancellations of massive events in Spain, as well as the first suspicious cases; causing in consequence an increase of the concern among the Spanish society, which started to look and ask for sanitary products. This is also the phase where the approach captures a transition from international to national concerns. We will breakdown this more in detail in the next paragraphs, matching again the topics against news from the newspapers to qualitatively verify the quality of the extracted topics. Table 2 illustrates such topics with the top keywords and tweets from the model.

W5 (from 25 to 31 January of 2020) During this week, the approach kept identifying online discussions about the WHO emergency declarations, considering COVID-19 as a global coronavirus threat (Pérez, 2020). Also, the approach extracted topics related to international restrictions, such as the airplane company Iberia suspending flights to Shanghai (CatalunyaPress.es, 2020), at the same time that Russia closed its frontiers with China (Ellyatt, 2020).

W6 (from 1 to 7 February of 2020) Following the trend of announcing emergency declarations, the model started to identify international issues, such as the infection and posterior death of Li Wenliang (BBC News, 2020), a Chinese doctor that alerted about the first cases of COVID-19 in December 2019, but also national ones; such as the confirmation of the first case of coronavirus in Spain, in the Canary Island of La Gomera (Linde, 2020). This matches the time where the number of cases seemed to start to spread (still slowly) all around the world.

W7 (from 8 to 14 February of 2020) During this week, the coronavirus started to have an important economic effect in Spain, which is reflected by the model, discovering topics that showed how users discussed the potential (finally confirmed during this week too) cancellation of the 2020 Mobile World Congress (MWC 2020), which usually takes place in Barcelona (Pardeiro, 2020). On the healthcare side, additional (few) cases started to be reported in Spain, such as in Mallorca, where it was reported the second Spanish case of COVID-19 (Bohórquez and Güell, 2020). During this and next weeks, we started to observe how there is a slow transition from international to national topics.

W8 (from 15 to 21 February of 2020) During this week, the topics where in line with those discussed in the previous weeks, such as the cancellation of the MWC 2020 (see Table 2) and its repercussion. This 'last-long' topics made sense at the time, since the cancellation of the MWC 2020 was the first massive event cancelled in Spain, with important economic consequences. Other international issues such as the sustained increase of cases in China or in the cruise ship Diamond Princess (Almoguera, 2020) seemed to occupy Twitter users during this time, too.

W9 (from 22 to 29 February of 2020) These are the final days before the lockdown period, and in retrospective, it is easy to see how some of the topics extracted reflected the immediate seriousness of the situation. We see how the model captures that the WHO advised to the public (World Health Organization (WHO), 2020a) to wash hands frequently. It is interesting to see in Table 2 how 'farmacia' (pharmacy) appears together with 'gel' (gel), 'lavarse' (to wash), 'mano' (hand) and 'alcohol' (alcohol), 'agotar' (to run out of) among the top keywords for the corresponding topic. In this context, it is well-known that these products were scarce in pharmacies and stores, and actually this problem lasted for long during the lockdown period. Also, related to the immediate seriousness of the situation, the model captured how despite of not being confined, the world economy started to suffer with the stocks set for the worst week since 2008 (Sano, 2020).

W10 (from 1 to 8 March of 2020) Just before the lockdown, we observe how among the topics extracted there are topics that we see everyday in the current pandemic life. For instance, as shown in Table 2, we kept seeing the importance of washing hands and keep a good hygiene with the use of soap (World Health Organization (WHO), 2020a). Also,

Topic	Discriminative Keywords	Generative Keywords
W5 (from 25 to 31 January of 2020)		
9	oms, emergencia, declarar, declaración, sanitaria, organización, comité, convocar	oms, internacional, emergencia, salud, declarar, alerta, mundial, china
Declara OMS emergencia por coronavirus - Vía @user <URL>		
1	vuelo, cerrar, suspender, frontera, kong, hong, rusia, aerolínea	china, vuelo, cerrar, suspender, brote, frontera, evitar, kong
Iberia suspende los vuelos a Shanghái por el coronavirus <URL>...		
W6 (from 1 to 7 February of 2020)		
4	alertar, acusar, news, silenciar, intentar, bbc, difundir, confusión	médico, china, chino, morir, alertar, wuhan, muerte, wenliang
Por favor lean. Porque esto no lo va a contar ningún medio que alerte sobre el coronavirus . <URL>...		
1	gomera, alemán, ingresado, contacto, jalisco, vitoria, ecuador, isla	caso, españa, gomera, paciente, sospechoso, hospital, salud, síntoma
En España ya tenemos un caso de coronavirus ,en La Gomera ,un alemán.		
W7 (from 8 to 14 February of 2020)		
6	mallorca, negativo, británico, ingresado, palma, princess, diamond, gomera	caso, mallorca, españa, crucero, paciente, confirmar, sospechoso, salud
Confirman un caso de coronavirus en Palma de Mallorca <URL>... <URL>		
3	sony, amazon, gsma, bajas, lg, nvidia, ericsson, intel	mobile, congress, barcelona, mwc, empresa, cancela, cancelar, sony
Tras las bajas de LG, Ericsson, NVidia, Amazon y Sony #coronavirus #MWC2020 <URL>...		
W8 (from 15 to 21 February of 2020)		
5	crucero, diamond, princess, pasajero, colombiano, camboya, evacuado, ucrania	crucero, cuarentena, japón, caso, diamond, princess, pasajero, wuhan
NUEVOS CASOS DE CORONAVIRUS EN CRUCERO DIAMOND <URL>... <URL>		
6	mobile, barcelona, cancelación, cancelar, maratón, evento, mwc, congress	mobile, china, tokio, barcelona, cancelar, cancelación, maratón, guerra
Suspenden el Mobile World Congress de Barcelona por el coronavirus <URL>... <URL>		
W9 (from 22 to 29 February of 2020)		
6	mano, farmacia, lavarse, gel, desinfectante, alcohol, agotar, carne	mascarillas, mano, gente, mascarilla, evitar, comprar, miedo, hospital
Cómo prevenir el #coronavirus . Lávate las manos, lávate las manos, lávate las manos..... lávate las manos. <URL>... <URL>		
1	bolsa, economía, mercado, caída, ibex, pérdida, crecimiento, wall	china, bolsa, economía, mercado, crisis, mundial, impacto, económico
'Esto es mercado. Esto me pone' @user #bolsa #COVID19 <URL>...		
W10 (from 1 to 8 March of 2020)		
7	mano, metro, lavarse, higiene, agua, gel, jabón, lavar	mano, evitar, medido, contagio, mascarillas, covid, persona, tomar
me voy a lavar las manos que no quiero el coronavirus		
4	patología, contagioso, anciano, letalidad, diferencia, estacional, comparación, hambre	gripe, persona, año, morir, mortalidad, gente, matar, enfermedad
Se llama Virus Corona Patologías Previas		

Table 2: Some representative topics for the weeks corresponding to the **outbreak** period of the COVID-19 pandemic in Spain. For each example topic, we include the top representative sentence according to its discriminative score.

'metro' (underground) is a top keyword of such topic, since at that time there was a discussion about the chances of getting infected (e.g., in the public transport) (CNN, 2020). In a different topic, we see what it seems to be a discussion comparing the flu and covid, and how they affect to the population, which was a popular comparison at the time.

4.3 Lockdown time

During the lockdown phase (until April), we can observe in Table 3 how the topics discussed mostly focused on the worst consequences of the pandemic, such as the big economic crisis, the large number of deaths per day, and also some collective actions such as thanking the healthcare workers. Again, we give a brief explanation below these lines, and match the topics against news in the media.

W11 (from 9 to 16 March of 2020) Here we consider the week where the Spanish society stopped to have free movement. More particularly, the government approved strict social confinement on 14 March of 2020 (Cué, 2020). Besides, the model found topics about the acknowledgement to the healthcare workers and the solidarity applause (La Razón, 2020), which was very popular in Spain dur-

Topic	Discriminative Keywords	Generative Keywords
W11 (from 9 to 16 March of 2020)		
3	aplauso, frenalacurva, aplausosanitario, cuarentenaya, yoelijoserresponsable, felizlunes, arena, agradecimiento	covid, yomequedoencasa, casa, quedateencasa, cuarentena, coronavirusesp, cuarentenacoronavirus, responsabilidad
Aplausos para que suenen más que los truenos que hoy hay en Madrid. Hoy mis aplausos para todos. Para \Saldremos de esta / #COVID19 <URL>		
8	ocasionado, aprobar, pymes, paliar, erte, fiscal, boe, hipoteca	covid, medido, crisis, gobierno, alarma, situación, sanitaria, empresa
#RealDecreto 463/2020 #estadodealarma #COVID19 <URL>#pymes #Autonomo #Cordoba @user <URL>		
W12 (from 17 to 24 March of 2020)		
6	respirador, fabricar, ifema, impresora, envío, coronavirius, epis, todosobremovil	covid, hospital, sanitario, mascarillas, madrid, personal, estevirusloparamosunidos, quedateencasa
#ElonMusk puede que empiece a fabricar respiradores #COVID19 <URL>		
1	higiene, jabón, distanciamiento, acatar, lavado, fanb, geacam, comerciales	covid, medido, evitar, contagio, prevención, propagación, salud, tomar
Entre más higiene se tenga, mayor es la protección ante los patógenos como el #COVID19 <URL>...		
W13 (from 25 to 31 March of 2020)		
1	erte, pago, despido, prestación, contrato, fiscal, ertes, alquiler	crisis, medido, gobierno, empresa, económico, trabajador, autónomo, sanitaria
Información para los afectados por ERTE debido al COVID19 . #ERTE #Coronavirus <URL ><URL>		
3	civil, guardia, desinfección, desinfectar, higiene, cumplimiento, jabón, estación	medido, persona, evitar, contagio, salud, seguridad, prevención, casa
Unos 400 guardias civiles con coronavirus en #CLM , según la @user @user -. Vía @user <URL>... <URL>		
W14 (from 1 to 7 April of 2020)		
1	animal, respiratorio, tigre, gato, mascota, zoo, bronx, contaminación	persona, paciente, enfermedad, síntoma, casa, contagio, evitar, matar
Si los tigres se contagian de coronavirus , ¡ojito los que tenéis gato!		
8	confirmado, cifra, elevar, defunción, diarios, ascender, activos, descender	caso, fallecido, españa, muerte, muerto, dato, número, país
637 muertes por coronavirus en un día, la cifra más baja en 13 días <URL>		
W15 (from 8 to 14 April of 2020)		
2	cifra, curado, reino, récord, ascender, contabilizar, diagnosticado, acumular	caso, fallecido, españa, muerto, muerte, dato, número, persona
Las 510 muertes por COVID-19 en un día, la cifra más baja desde el 23 de marzo <URL>		
5	johnson, intensivo, boris, testimonio, alta, universitario, clmpressdigital, sosprisiones	hospital, médico, paciente, sanitario, madrid, personal, persona, profesional
Coronavirus : Boris Johnson fue dado de alta. <URL>		
W16 (from 15 to 20 April of 2020)		
5	luis, sepúlveda, escritor, homenaje, chileno, fútbol, club, dep	año, morir, hospital, luis, fallecer, quedateencasa, yomequedoencasa, historia
Luis Sepúlveda muere por coronavirus <URL>... <URL>		
2	distanciamiento, prohibidorendirse, enestafamilianadieluchasolo, yonosoyungastosuperfluo, bicicleta, espandemia, comunidadvalenciana, saltarse	confinamiento, quedateencasa, yomequedoencasa, cuarentena, medido, casa, evitar, alarma
¿El distanciamiento social podría ir incluso más allá de 2021? #COVID19 #coronavirus <URL>		

Table 3: Some representative topics for the weeks corresponding to the **lockdown** period of the COVID-19 pandemic in Spain. For each example topic, we include the top representative sentence according to its discriminative score.

ing the lockdown period. In a related line, topics like this one also captured the feeling of the importance of staying at home to prevent becoming infected and reduce the workload of these workers.

W12 (from 17 to 24 March of 2020) For this week, the model extracted topics discussing the personal hygiene measures to combat the COVID-19. The topics also reflect the lack of equipment in the hospitals, which was a problem at the beginning of the pandemic. More particularly, the model was able to identify as a topic the lack of ventilators in Spain, and also the rest of the world, as reflected by the most salient discriminative tweet. This matches the news at the time, which discussed the use of 3D printers to provide such ventilators (Polo, 2020), or hacking some objects to adapt them for medical use (Cristian Fracassi, 2020).

W13 (from 25 to 31 March of 2020) This week covers the last days of March 2020.

Due to the strict confinement, topics concerning job losses and the measures taken by the government to counteract the situation (e.g. the so-called ERTEs) started to arise (RTVE.es, 2020; Gestiona.es, 2020). Among the rest of the topics of this week, we also would like to remark the massive infection of public workers, such as the Guardia Civil officers in Castilla La-Mancha (EFE/CMM, 2020). The infection of public workers during this time of the pandemic was also widely discussed in the news (Requeijo, 2020).

W14 (from 1 to 7 April of 2020) On the national side, some topics reflected the number of casualties per day. More particularly, the beginning of April corresponded to the peak of the first wave, and the beginning of the decreasing trend in the number of infections and deaths per day (Justo, 2020). A bit on a different line, we found topics discussing more diverse aspects of COVID-19, such as the infection in the Zoo of Bronx (New York, USA) of tigers and lions (M.R.M., 2020).

W15 (from 8 to 14 April of 2020) Here, we would like to remark a topic related to an international breaking news, and more particularly about Boris Johnson (the UK Prime Minister) being infected by the coronavirus, together with his evolution, when he even entered the ICU (La Vanguardia, 2020). On the national side, the models kept detecting topics related to the number of deaths in Spain, which was still high and dynamic during that time, but reached some local minima these days (Soteras, 2020).

W16 (from 15 to 20 April of 2020) For the last days of our study, the model found relevant topics too, such as the death of the Chilean writer Luis Sepúlveda (Safont Plumed, 2020) due to COVID-19, or topics related to the need of keeping social distancing, maybe even for months (elEconomista.es, 2020), as reflected by some of the most representative tweets.

4.4 Quantitative evaluation

We performed a small human evaluation to quantitatively estimate the quality of the extracted topics. We took 20 topics randomly from all periods. Then, two annotators were in charge of: (i) determining if given the top 8 keywords and 3 top sentences made possible to infer a topic, (ii) determining if for each top topic word (according to the discriminative score) they belonged to the inferred topic, and (iii) the same as in (ii), but for the 3 most representative sentences. We calculated the percentage of times both annotators positively labelled a sample, obtaining scores of 80%, 56.88% and 71.66% for (i), (ii), and (iii), respectively. In addition, we calculated (ii) but taking into account only the first 3 top keywords of the topic, yielding a score of 75% of positive samples.

5 Conclusion

This paper used a topic modeling approach to shed some light about the topics discussed in Spain during the early stages of the COVID-19 pandemic, including a period of pre-crisis, the outbreak of the disease, and the beginning of the confinement. We collected a large amount of tweets using keywords and cleaned them to keep only Spanish tweets that were written in Spain. After that, we used a Latent Dirichlet Allocation model that learned to cluster such tweets according to the topic they discuss. To represent the topics, we used generative and discriminative routes to extract the most salient keywords and sentences. To verify the quality of the extracted topics, we performed a qualitative analysis matching the topics against relevant news in the newspapers at the same period of time, and a small quantitative evaluation. Overall, the topics show that during the pre-crisis period, users focused on the international panorama than the local situation, while during the outbreak and lockdown phases they focused the most on the Spanish emergency, considering health and economic problems.

Acknowledgements

MMAT has been partially funded by Barcelona Supercomputing Center (BSC) through the Spanish Plan for advancement of Language Technologies 'Plan TL' and the Secretaría de Estado de Digitalización e Inteligencia Artificial (SEDIA). DV is supported by MINECO (TIN2017-85160-C2-1-R), by Xunta de Galicia (ED431C 2020/11), by Centro de Investigación de Galicia 'CITIC' (European Regional Development Fund-Galicia 2014-2020 Program, ED431G 2019/01), and by a 2020 Leonardo Grant for Researchers and Cultural Creators from the BBVA Foundation.

References

Abd-Alrazaq, A., D. Alhuwail, M. Househ, M. Hamdi, and Z. Shah. 2020. Top concerns of tweeters during the covid-19 pandemic: infoveillance study. *Journal of medical Internet research*, 22(4):e19016.

Afzal, Z., V. Yadav, O. Fedorova, V. Kandala, J. van de Loo, S. A. Akhondi, P. Coupet, and G. Tsatsaronis. 2020. CORA: A deep active learning covid-19 relevancy algorithm to identify core scientific articles. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December. Association for Computational Linguistics.

Agencia EFE. 2020. La OMS pone en alerta a la red mundial de hospitales por un nuevo coronavirus en China. *www.efe.com*, January.

Almoguera, P. 2020. El coronavirus pone en jaque ahora a Japón y Corea del Sur. *El País*, February.

Amara, A., M. A. H. Taieb, and M. B. Aouicha. 2020. Multilingual topic modelling for tracking covid-19 trends based on facebook data analysis.

Andrzejewski, D. and D. Buttler. 2011. Latent topic feedback for information retrieval. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 600–608.

Arun, R., V. Suresh, C. V. Madhavan, and M. N. Murthy. 2010. On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 391–402. Springer.

Asgari-Chenaghlu, M., N. Nikzad-Khasmakhi, and S. Minaee. 2020. Covid-transformer: Detecting trending topics on twitter using universal sentence encoder. *arXiv preprint arXiv:2009.03947*.

Banda, J. M., R. Tekumalla, G. Wang, J. Yu, T. Liu, Y. Ding, K. Artemova, E. Tutubalina, and G. Chowell. 2020. A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration, August.

Barde, B. V. and A. M. Bainwad. 2017. An overview of topic modeling methods and tools. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 745–750.

BBC News. 2020. Li Wenliang: Coronavirus kills Chinese whistleblower doctor. *BBC News*, February.

Blei, D. M. and J. D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 113–120, New York, NY, USA. Association for Computing Machinery.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bohórquez, L. and O. Güell. 2020. El segundo caso de coronavirus en España es un británico que se contagió en los Alpes. *El País*, February.

Boon-Itt, S. 2020. A text-mining analysis of public perceptions and topic modeling during the covid-19 pandemic using twitter data. *JMIR public health and surveillance, JMIR Preprints. 30/06/2020:21978*.

Cao, J., T. Xia, J. Li, Y. Zhang, and S. Tang. 2009. A density-based method for adaptive lda model selection. *Neurocomputing*, 72(7-9):1775–1781.

Carbonell Gironés, L. 2020. Geographical analysis of the opinion and influence of users on twitter during the coronavirus health crisis. Final project/degree, Escola Tècnica Superior d'Enginyeria Informàtica, Universitat Politècnica de València.

CatalunyaPress.es. 2020. Iberia suspende los vuelos a Shanghái por el coronavirus.

Cer, D., Y. Yang, S. yi Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, and R. Kurzweil. 2018. Universal sentence encoder.

Chandrasekaran, R., V. Mehta, T. Valkunde, and E. Moustakas. 2020. Topics, trends, and sentiments of tweets about the covid-19 pandemic: Temporal infoveillance study. *Journal of Medical Internet Research*, 22(10):e22624.

Chen, E., K. Lerman, and E. Ferrara. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 6(2):e19273.

CNN. 2020. Medidas globales por el coronavirus: mantener distancia de un metro, cierre de escuelas y museos, evitar los besos y otras, March.

Cristian Fracassi. 2020. Charlotte valve, March.

Cué, C. E. 2020. El Gobierno informa de que es la única autoridad en toda España, limita los desplazamientos y cierra comercios, March.

Deerwester, S. 1988. Improving information retrieval with latent semantic indexing.

EFE/CMM. 2020. 400 guardias civiles de Castilla-La Mancha tienen Covid-19, según la AUGC.

El Boletín. 2020. China pone en cuarentena a más de 30 millones de personas por el coronavirus. January.

elEconomista.es. 2020. Las medidas de distanciamiento social podrían extenderse hasta 2022 de manera intermitente - elEconomista.es.

Ellyatt, H. 2020. Russia closes border with China to prevent spread of the coronavirus, January.

Gao, Y., Y. Xu, and Y. Li. 2014. Pattern-based topics for document modelling in information filtering. *IEEE Transactions on Knowledge and Data Engineering*, 27(6):1629–1642.

Gestiona.es. 2020. Información para los afectados por ERTE debido al COVID19, March.

Griffiths, T. L. and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.

Grover, P., A. K. Kar, Y. K. Dwivedi, and M. Janssen. 2019. Polarization and acculturation in us election 2016 outcomes–can twitter analytics predict changes in voting preferences. *Technological Forecasting and Social Change*, 145:438–460.

Hofmann, T. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, page 289–296, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Hutto, C. and E. Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, volume 81, page 82.

Jagarlamudi, J., H. Daumé III, and R. Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213.

Justo, D. 2020. España sigue la tendencia a la baja: 4.273 nuevos contagios por coronavirus y 637 muertes, April.

Kerchner, D. and L. Wrubel. 2020. Coronavirus Tweet Ids.

Kleinberg, B., I. van der Vegt, and M. Mozes. 2020. Measuring Emotions in the COVID-19 Real World Worry Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July. Association for Computational Linguistics.

La Razón. 2020. Emotivo reconocimiento a los sanitarios en forma de aplausos desde los balcones, March.

La Vanguardia. 2020. Boris Johnson recibe el alta y continuará recuperándose de la Covid-19 en su casa, April.

Linde, P. 2020. Sanidad confirma en La Gomera el primer caso de coronavirus en España. *El País*, February.

Loria, S. 2020. textblob documentation. *Release 0.16*, 2.

McInnes, L., J. Healy, and J. Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, February.

Mohammad, S. M. and P. D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Moody, C. E. 2016. Mixing dirichlet topic models and word embeddings to make lda2vec.

M.R.M. 2020. Un tigre del zoo de Nueva York tiene coronavirus, April.

Neubig, G., Y. Matsubayashi, M. Hagiwara, and K. Murakami. 2011. Safety information mining—what can nlp do in a disaster—. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 965–973.

Ordun, C., S. Purushotham, and E. Raff. 2020. Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs. *arXiv preprint arXiv:2005.03082*.

Pardeiro, M. 2020. El fracaso político del MWC: "No se va a suspender". "No cuelga de un hilo".

Pham, P., P. Do, and C. D. Ta. 2018. W-pathsim: novel approach of weighted similarity measure in content-based heterogeneous information networks by applying lda topic modeling. In *Asian conference on intelligent information and database systems*, pages 539–549. Springer.

Polo, J. 2020. Coronavirus: La Zona Franca fabricará 100 respiradores diarios con impresoras 3D, March.

Pérez, B. 2020. La OMS rectifica y declara la emergencia global por el coronavirus, January.

Requeijo, A. 2020. La Policía y la Guardia Civil suman ya más de 400 positivos por coronavirus, March.

RTVE.es. 2020. Los ERTE por la crisis del coronavirus suman más de 240.000, March.

Safont Plumed, J. 2020. Muere el escritor chileno Luis Sepúlveda, a causa del coronavirus.

Salton, G., A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, November.

Sano, H. 2020. GLOBAL MARKETS-World stocks set for worst week since 2008 as virus fears grip markets. *Reuters*, February.

Soteras, A. 2020. COVID-19: 510 muertes en un día, la cifra más baja desde el 23 de marzo.

Steinskog, A., J. Therkelsen, and B. Gambäck. 2017. Twitter topic modeling by tweet aggregation. In *Proceedings of the 21st nordic conference on computational linguistics*, pages 77–86.

Tuarob, S., L. C. Pouchard, and C. L. Giles. 2013. Automatic tag recommendation for metadata annotation using probabilistic topic modeling. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pages 239–248.

Verspoor, K., K. B. Cohen, M. Conway, B. de Bruijn, M. Dredze, R. Mihalcea, and B. Wallace, editors. 2020a. *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December. Association for Computational Linguistics.

Verspoor, K., K. B. Cohen, M. Dredze, E. Ferrara, J. May, R. Munro, C. Paris, and B. Wallace, editors. 2020b. *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July. Association for Computational Linguistics.

Vilares, D. and C. Gómez-Rodríguez. 2018. Grounding the semantics of part-of-day nouns worldwide using twitter. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 123–128.

Wang, L. L., K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. M. Kinney, Y. Li, Z. Liu, W. Merrill, P. Mooney, D. A. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, N. X. R. Wang, C. Wilhelm, B. Xie, D. M. Raymond, D. S. Weld, O. Etzioni, and S. Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July. Association for Computational Linguistics.

Wojcik, S. and A. Hughes. 2019. Sizing up twitter users. *PEW research center*, 24.

World Health Organization (WHO). 2020a. Advice for the public on COVID-19 – World Health Organization.

World Health Organization (WHO). 2020b. WHO statement regarding cluster of

pneumonia cases in Wuhan, China. January. Accessed: 2020-08-28.

Yijun, G. and X. Tian. 2014. Study on keyword extraction with lda and textrank combination. *Data Analysis and Knowledge Discovery*, 30(7):41–47.

Yin, H., S. Yang, and J. Li. 2020. Detecting topic and sentiment dynamics due to covid-19 pandemic using social media. *arXiv preprint arXiv:2007.02304*.

Yu, J., Y. Lu, and J. Muñoz-Justicia. 2020. Analyzing spanish news frames on twitter during covid-19—a network study of el país and el mundo. *International Journal of Environmental Research and Public Health*, 17(15):5414.

Zhou, S., K. Li, and Y. Liu. 2009. Text categorization based on topic model. *International Journal of Computational Intelligence Systems*, 2(4):398–409.

Tesis

Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 193-196 recibido 13-12-2020 revisado 24-01-2021 aceptado 03-02-2021

Negation Processing in Spanish and its Application to Sentiment Analysis

Tratamiento de la Negación en Español y su Aplicación al Análisis de Sentimientos

Salud María Jiménez-Zafra
SINAI, Department of Computer Science, CEATIC, Universidad de Jaén,
Campus Las Lagunillas s/n, 23071, Jaén (Spain)
sjzafra@ujaen.es

Abstract: This is a summary of the Ph.D. thesis written by Salud María Jiménez Zafra at Universidad de Jaén under the supervision of Ph.D. María Teresa Martín Valdivia and Ph.D. L. Alfonso Ureña López. The author was examined on Friday, September 13th, 2019 by a committee formed by Ph.D. Ruslan Mitkov from the University of Wolverhampton, Ph.D. Miguel Ángel García Cumbreras from Universidad de Jaén and Ph.D. Eugenio Martínez Cámara from Universidad de Granada. The Ph.D. thesis obtained Summa Cum Laude and the international mention. Moreover, it was awarded as the best thesis in Natural Language Processing at the 36th International Conference of the Spanish Society for Natural Language Processing (SEPLN 2020).
Keywords: Negation processing, sentiment analysis, machine learning, natural language processing.

Resumen: Este es un resumen de la tesis doctoral realizada por Salud María Jiménez Zafra en la Universidad de Jaén bajo la dirección de los doctores Dña. María Teresa Martín Valdivia y D. L. Alfonso Ureña López. El acto de defensa tuvo lugar el viernes 13 de septiembre de 2019 ante el tribunal formado por los doctores D. Ruslan Mitkov de la Universidad de Wolverhampton, D. Miguel Ángel García Cumbreras de la Universidad de Jaén y D. Eugenio Martínez Cámara de la Universidad de Granada. La tesis obtuvo la calificación de Sobresaliente Cum Laude por unanimidad y mención de doctorado internacional. Además, recibió el premio a la mejor tesis doctoral en Procesamiento del Lenguaje Natural en el XXXVI Congreso Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN 2020).
Palabras clave: Tratamiento de la negación, análisis de sentimientos, aprendizaje automático, procesamiento del lenguaje natural.

1 Introduction

Natural Language Processing (NLP) is the field of Artificial Intelligence that aims to provide mechanisms to facilitate the communication between humans and machines through natural language (Indurkhya y Damerau, 2010). This is a challenge because computers must be able to process, understand and generate language. If we want to develop systems that approach human understanding, we must incorporate in them the processing of a diversity of linguistic phenomena, such as negation, irony or sarcasm, which are used to give words a different meaning.

This doctoral thesis focuses on the study of one of the main linguistic phenomena that we use in our daily communication, *negation* (Horn, 1989; Morante y Sporleder, 2012). Four tasks are usually performed in relation to processing negation: (i) negation cue detection, in order to find the words that express negation; (ii) scope identification, in order to find which parts of the sentence are affected by the negation cues; (iii) negated event recognition, to determine which events are affected by the negation cues; and (iv) focus detection, in order to find the part of the scope that is most prominently negated. Example (1) shows a sentence in which the negation cue appears in bold, the event in italics, the focus underlined and the scope between brackets.

ISSN 1989-7553. DOI 10.26342/2021-66-16

1. Es una persona que [**no** *tiene* límites], aunque a veces puede controlarse.
   He is a person who has no limits, although sometimes he can control himself.

This thesis addresses negation cue detection and scope identification tasks. In contrast to most of the studies existing so far that are on English language, it is carried out on Spanish texts, since not even Google has been able to process Spanish negation adequately. For example, if we perform in Google the search -*películas que no sean de aventuras*-, we can see that it returns adventure films when it should return films of other themes. Negation processing is not only important for information retrieval systems. In other systems, such as those of sentiment analysis, not processing negation can lead to the extraction of a completely different opinion than the one expressed by the user. For example, the polarity of the sentence "Una película fascinante, repetiría" *[A fascinating film, I would repeat]* should be the opposite of its negation "Una película nada fascinante, no repetiría" *[A not at all fascinating film, I would not repeat]*.

The objective of this dissertation is to advance in the processing of negation in Spanish and to show the importance of the computational treatment of negation for NLP systems. For this, an exhaustive study of negation is carried out, incorporating negation processing systems, corpora and sentiment analysis systems in which negation has been taken into account. In addition, a typology of negation patterns in Spanish is defined, which is applied for the annotation of a corpus with negation, the SFU Review$_{SP}$-NEG corpus. This corpus is used to develop a Spanish negation processing system which is applied to sentiment analysis in order to improve the predictive capacity of opinion classification systems that are so in demand today. Finally, NEGES has been launched, the first initiative promoting negation research in Spanish for which three editions have already been held in the context of the International Conference of the Spanish Society for Natural Language Processing (SEPLN).

2 Structure

This thesis is organized in eight chapters and one appendix, which are described hereafter.

Chapter 1 presents the motivation, objective and difficulty of the research addressed.

Chapter 2 introduces the concepts of negation and sentiment analysis, and presents the state-of-the-art for negation processing systems, the corpora annotated with negation, and sentiment analysis systems that take into account negation.

Chapter 3 shows the preliminary research, which reveals the importance of a correct processing of negation and the need to annotate a corpus with sentiment and negation. Spanish sentiment analysis systems existing up to now take negation into account as one more feature, but its effect on the classification is not evaluated.

Chapter 4 presents the SFU Review$_{SP}$-NEG corpus and the process followed for its annotation. In this chapter the components of negation are defined and delimited and it is proposed a typology of negation patterns in Spanish, which is applied for the annotation of the corpus. Moreover, it includes the annotation scheme used, the annotation process followed, the main sources of disagreement and the statistics and description of the corpus.

Chapter 5 includes all the details of the Spanish negation processing system developed. It contains an exhaustive analysis of the existing corpora in order to select the set of data for training the system. In addition, it presents the architecture of the proposed system, the experiments carried out, the results obtained and an analysis of errors aimed at understanding the limitations of the system.

Chapter 6 corresponds to the integration of the Spanish negation processing system developed into a sentiment analysis system. It presents the methodology followed to study the effect of negation, the experiments carried out and the results obtained, as well as an error analysis. It shows the importance of the development of accurate negation processing systems for NLP tasks.

Chapter 7 presents NEGES: Workshop on Negation in Spanish, the first initiative promoting negation research in Spanish. It contains the details of the origin of the workshop, its objective, the editions held, the tasks proposed, the datasets provided and the participants and results obtained.

Chapter 8 summarizes the conclusions, the main contributions, the research awards

and distinctions obtained, and the future lines of work.

Finally, **Appendix A** contains the tables summarizing the corpora analysis carried out in Chapter 5.

3 Contributions

Its main contributions can be grouped into 5 categories: state-of-the-art, resources, systems, analysis, and workshops.

State-of-the-art. We provided a thorough review of the work developed so far on the following topics (Jiménez-Zafra et al., 2018a; Jiménez-Zafra et al., 2019c): (i) English and Spanish negation processing systems; (ii) English and Spanish sentiment analysis systems that take into account negation; and (iii) Corpora annotated with negation.

Resources. Until now, there was no typology in Spanish to characterize and classify negation. Therefore, we defined our own (i) typology of negation patterns (Martí et al., 2016) taking into account their syntactic structure and their semantic interpretation. In addition, we defined an (ii) annotation scheme for negation and how it affects the sentiment of the sentence (Jiménez-Zafra et al., 2018b). We also generated (iii) the SFU Review$_{SP}$-NEG corpus (Jiménez-Zafra et al., 2018b) [1], the first corpus annotated with negation in the review domain for Spanish in which it is annotated how negation affects the words that are within its scope, that is, whether there is a change in the polarity or an increase or decrease of its value. Moreover, we presented (iv) a compilation of the corpora annotated with negation for all languages (Jiménez-Zafra et al., 2019c).

Systems. We developed (i) a polarity classification system for Spanish tweets that incorporates a set of syntactic rules for determining the scope of negation (Jimenez-Zafra et al., 2017). This rule-based approach has been proved to be better than the method most used to determine the scope of negation in English tweets. Furthermore, we implemented (ii) a machine learning system to process negation in Spanish (Jiménez-Zafra et al., 2020a). This system outperforms state-of-the-art results for negation cue detection, whereas for scope identification it is the first system that performs the task for Spanish.

[1]First Online: 22 May 2017 https://doi.org/10.1007/s10579-017-9391-x

Analysis. We reported (i) the problematic cases found during the annotation of the SFU Review$_{SP}$-NEG corpus in order to facilitate the annotation of this phenomenon for other researchers (Jiménez-Zafra et al., 2016). We also conducted (ii) an analysis of the corpora annotated with negation discussing the possibility of merging the corpora to create a larger data set to train a negation processing system. Moreover, we showed overall negation processing tasks for which the corpora could be used, and specific tasks for which the corpora could be used to evaluate the impact of processing negation. In addition, we provided (iii) a qualitative error analysis showing which negation cues and scopes are straightforward to predict automatically, and which ones are challenging (Jiménez-Zafra et al., 2020a). Furthermore, we studied (iv) the effect of the negation processing system developed on the sentiment analysis task (Jiménez-Zafra et al., 2020b).

Workshops. Finally, we created NEGES group and NEGES workshop, the first initiative promoting negation research in Spanish (Jiménez-Zafra et al., 2018a; Jiménez-Zafra et al., 2018b; Jiménez-Zafra et al., 2019a; Jiménez-Zafra et al., 2019b). NEGES is the acronym for "NEGación en ESpañol" (Negation in Spanish). It provides a means of exchanging news of recent research developments and other matters of interest as well as it makes available resources relevant to negation detection in Spanish, including corpora, annotation guidelines, evaluation scripts, etc. Up to now, three editions of NEGES have been held in the context of the SEPLN International Conference.

4 Conclusions

Negation is a complex linguistic phenomenon and the issue of its computational treatment has not been resolved yet due to its complexity, the multiple linguistic forms in which it can appear and the different ways it can act on the words within its scope. All languages possess different types of resources (morphological, lexical, syntactic) that allow speakers to speak about properties that people or things do not hold or events that do not happen. The presence of a negation in a sentence can have enormous consequences in many real world situations, for example, when processing clinical records. One might think that, given the fact that negations are

so crucial in language, most NLP pipelines incorporate negation modules and that the computational linguistics community has already addressed this phenomenon. However, this is not the case. Work on processing negation has started relatively late as compared to work on processing other linguistic phenomena. This doctoral thesis aims to advance the study of negation processing in Spanish.

Acknowledgements

This Ph.D. thesis has been partially supported by a grant from the Ministerio de Educación Cultura y Deporte (MECD - scholarship FPU014/00983), Fondo Europeo de Desarrollo Regional, LIVING-LANG project (RTI2018-094653-B-C21), REDES project (TIN2015-65136-C2-1-R) and ATTOS project (TIN2012-38536-C03-0) from the Spanish Government.

References

Horn, L. R. 1989. *A natural history of negation*. CSLI Publications.

Indurkhya, N. y F. J. Damerau. 2010. *Handbook of Natural Language Processing*, volumen 2. CRC Press.

Jiménez-Zafra, S. M., N. P. Cruz Díaz, R. Morante, y M. T. Martín-Valdivia. 2018a. Tarea 1 del Taller NEGES 2018: Guías de Anotación. En *Proceedings of NEGES 2018: Workshop on Negation in Spanish*, volumen 2174, páginas 15–21, Seville, Spain. CEUR-WS.

Jiménez-Zafra, S. M., N. P. Cruz Díaz, R. Morante, y M. T. Martín-Valdivia. 2018b. Tarea 2 del Taller NEGES 2018: Detección de Claves de Negación. En *Proceedings of NEGES 2018: Workshop on Negation in Spanish*, volumen 2174, páginas 35–41, Seville, Spain. CEUR-WS.

Jiménez-Zafra, S. M., N. P. Cruz Díaz, R. Morante, y M. T. Martín-Valdivia. 2019a. NEGES 2018: Workshop on Negation in Spanish. *Procesamiento del Lenguaje Natural*, (62):21–28.

Jiménez-Zafra, S. M., N. P. Cruz Díaz, R. Morante, y M. T. Martín-Valdivia. 2019b. NEGES 2019 Task: Negation in Spanish. En *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, CEUR Workshop Proceedings, Bilbao, Spain. CEUR-WS.

Jiménez-Zafra, S. M., N. P. Cruz-Díaz, M. Taboada, y M. T. Martín-Valdivia. 2020b. Negation detection for sentiment analysis: A case study in spanish. *Natural Language Engineering*, 1(1):1–30.

Jiménez-Zafra, S. M., M. T. Martín-Valdivia, L. A. U. Lopez, M. A. Martí, y M. Taulé. 2016. Problematic cases in the annotation of negation in spanish. En *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM)*, páginas 42–48.

Jiménez-Zafra, S. M., R. Morante, E. Blanco, M. T. M. Valdivia, y L. A. U. Lopez. 2020a. Detecting negation cues and scopes in spanish. En *Proceedings of The 12th Language Resources and Evaluation Conference*, páginas 6902–6911.

Jiménez-Zafra, S. M., R. Morante, M. T. Martín-Valdivia, y L. A. U. Lopez. 2018a. A review of spanish corpora annotated with negation. En *Proceedings of the 27th International Conference on Computational Linguistics*, páginas 915–924.

Jiménez-Zafra, S., M. Taulé, M. Martín-Valdivia, L. A. Ureña-López, y M. A. Martí. 2018b. SFU ReviewSP-NEG: a Spanish corpus annotated with negation for sentiment analysis. A typology of negation patterns. *Language Resources and Evaluation*, 52(2):533–569.

Jimenez-Zafra, S. M., M. T. M. Valdivia, E. M. Camara, y L. A. Urena-Lopez. 2017. Studying the scope of negation for spanish sentiment analysis on twitter. *IEEE Transactions on Affective Computing*, 10(1):129–141.

Jiménez-Zafra, S. M., R. Morante, M. T. Martín-Valdivia, y L. A. Ureña-López. 2019c. Corpora Annotated with Negation: An Overview. *Computational Linguistics (Under review - Second round)*.

Martí, M. A., M. Taulé, M. Nofre, L. Marsó, M. T. Martín-Valdivia, y 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. y C. Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational Linguistics*, 38(2):223–260.

Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 197-200 recibido 13-12-2020 revisado 24-01-2021 aceptado 03-02-2021

Document-Level Machine Translation – Ensuring Translational Consistency of Non-Local Phenomena

Traducción Automática a Nivel de Documento – Asegurando la Consistencia en la Traducción de Fenómenos no Locales

Eva Martínez Garcia
TALP Research Center
Univesitat Politècnica de Catalunya
Jordi Girona, 1-3, 08034 Barcelona, Spain
martinezgarcia.eva@gmail.com

Abstract: PhD Thesis written by Eva Martínez Garcia under the supervision of Dr. Cristina España-Bonet and Dr. Lluís Màrquez. The thesis was defended at the Universitat Politècnica de Catalunya in Barcelona on the 19th of December, 2019. The doctoral committee comprised of Dr. Kepa Sarasola (President, University of Basque Country (UPV/EHU)), Dr. Marta Ruiz Costa-Jussà (Universitat Politècnica de Catalunya (UPC)) and Dr. Sara Stymne (Uppsala Universitet). The thesis was awarded an excellent grade and international mention.
Keywords: Machine Translation, Document-level, Context-aware translation.

Resumen: Tesis doctoral elaborada por Eva Martínez Garcia bajo la supervisión de los doctores Cristina España-Bonet y Lluís Màrquez. La defensa de la tesis tuvo lugar en la Universitat Politècnica de Catalunya en Barcelona el 19 de diciembre de 2019. El tribunal estuvo compuesto por los doctores Kepa Sarasola (Presidente, Universidad del País Vasco (UPV/EHU)), Marta Ruiz Costa-Jussà (Universitat Politècnica de Catalunya (UPC)) y Sara Stymne (Uppsala Universitet). La tesis obtuvo la calificación de sobresaliente y la mención internacional.
Palabras clave: Traducción Automática, Nivel de documento, traducción co contexto.

1 Introduction

Machine Translation (MT) is very present in our daily lives. We use it to access information in other languages on the Internet or to figure out how to say something in languages we do not master for interaction and communication purposes. We are frequent users of the most popular online translation services (e.g., Google Translate, Bing, or DeepL) and we are also used to consuming the MT services provided by social networks (e.g., Facebook or Twitter), which allow us to access the published information in our preferred language. MT is present even in telecommunication applications like Skype, which offers video chats with real-time speech-to-speech translation services. This extended use of MT technology makes us familiarized with its advantages and drawbacks.

Although current MT systems have achieved good translation quality, even comparable with human translation quality in some cases (Wu et al., 2016; Hassan et al., 2018), they still hold a known limitation: they work at sentence level. MT systems translate a document sentence by sentence, taking into account a short context and ignoring document-level information. For all kinds of systems, ignoring extra-sentential information is required due to performance concerns and to the difficulty of properly representing long-distance dependencies. Statistical Machine Translation (SMT) systems (Koehn et al., 2007) rely on local n-gram information, and for Neural Machine Translation (NMT) systems (Bahdanau, Cho, and Bengio, 2015; Vaswani et al., 2017) it is still an open problem how to represent long sequences of words with fixed-length vectors. Thus, state-of-the-art systems perform translation assuming that every sentence can be translated in an isolated way.

ISSN 1989-7553. DOI 10.26342/2021-66-17

However, texts contain relationships among words that hold their coherence, cohesion, and consistency across sentences (Dijk, 1977; Sanders and Pander Maat, 2006).We consider that a good translation should reflect and maintain these qualities at the same degree as they appear in the source text. This is the motivation for our work, which explores techniques to improve the coherence and cohesion levels of the translations generated by state-of-the-art MT systems. Some of the typical mistakes of current MT systems can be linked to the lack of contextual coherence present in the followed translation approaches. We take as inspiration how human translators can resolve these phenomena naturally, by using the entire document's context information.

As an example to illustrate this phenomenon, consider using an MT system to translate a news item in English about a claim process in some office. The word "desk" can appear several times and it can be translated into Spanish as "mostrador", "ventanilla", "escritorio", or "mesa". These Spanish words are not synonyms. Where "mostrador" and "ventanilla" can both be a counter where a service is offered, "mesa" and "escritorio" refer to a piece of furniture. So, "desk" is a word with ambiguous translation into Spanish. Within the context of our example, "mesa" and "escritorio" are not correct translations for "desk". We address this as a contextual coherence problem. Our work aim is to use inter-sentence context to help the system choose a more adequate translation without any knowledge from the domain.

Another typical issue is word agreement across translation segments. Coreference chains confer cohesion to a document, and it is desirable to see this property projected into the produced translations. Unfortunately, this is a property that is typically difficult to maintain for MT systems. Also, gender and number agreement between words is sometimes challenging for current MT approaches. For example, consider the following set of sentences in a source document in English: "She studied civil engineering. [...] The civil engineer was the youngest in the company." These sentences can be translated into Spanish as "*Ella* estudió ingeniería civil. [...] *El ingeniero* era *el* más joven de la empresa." This translation is correct in Spanish if we look at it sentence by sentence.However, it is incorrect if we consider it in its entirety as part of the same document, since there is no gender agreement between the translations of "the engineer" and "she". Taking document context into account, the correct translation would be "*Ella* estudió ingeniería civil. [...] *La ingeniera* era *la* más joven de la empresa."

Our work is motivated by the idea that exploiting discourse information would help to improve the quality of the resulting machine translations at document level. All the techniques we explore in this thesis attempt to find the best way to exploit such kind of information within the current MT frameworks.

2 Research Goals

The general goal of the thesis' work is to improve MT quality by exploring the use of document-level information at different steps of the translation process in order to fix or prevent some of the errors made by sentence-level MT systems. The main goal is to improve machine translation coherence and cohesion by leveraging the information given by the relations of the words along a document.

We define a research strategy with the following steps:

1. *Analyzing translation errors related to document-level phenomena and designing simple methods to tackle them.* A first step towards improving document-level machine translation is to identify those phenomena that confer coherence and cohesion to documents and are susceptible to be lost in the MT process. Before solving such mistakes during the MT process, it is interesting to implement a set of simple post-processing techniques and evaluate their impact.
2. *Capturing the semantic information of a document in a useful manner to aid the MT decoding process.* Leveraging a document's semantic context should help improve the coherence and cohesion levels of its translation. It is necessary to explore ways to introduce contextual semantic information into the MT process. Our final intention is to *extend a document-oriented decoder to incorporate document context semantics.*
3. *Enhancing an NMT framework using context-aware techniques.* To finalize, one of our goals is to integrate the explored ideas into the NMT paradigm.

3 Thesis Overview

The thesis is organized in 7 chapters, followed by an appendix.

Chapter 1 introduces and motivates the work, highlighting the importance of using the context information to improve translationś quality. Chapter 2 revisits the state-of-the-art of the MT research area, focusing on the main technologies of the SMT and NMT paradigms, both at sentence and document level. Chapters 3 to 6 present our results.

In Chapter 3, we analyze some of the translation errors related to document-level phenomena and present a set of post-processing strategies to handle them.

Chapter 4 describes how to use word embeddings for decoding. First, we study the applicability of word embeddings to enhance the MT process. Then, we explain a method to enhance a document-oriented SMT decoder with word embeddings working as Semantic Space Language Models.

Next, Chapter 5 describes our extension of a document-oriented SMT decoder to handle the phenomenon of lexical choice consistency. We present a new feature function that guides the decoder towards more lexically consistent translation candidates, as well as a new strategy to shortcut the exploration of the search space.

Chapter 6 presents our approach to extend the NMT decoding process to take into account contextual semantics. In particular, we extend the beam search decoding algorithm by fusing the discourse information captured by the models described in Chapter 4 to work in tandem with the NMT model.

Finally, Chapter 7 draws the conclusions and describes possible avenues of future work.

Additionally, Appendix *A* describes a new document-level decoding strategy based on a swarm optimization algorithm, integrated into the decoder used in Chapter 4 and 5 as an alternative to its default hill climbing.

4 Main Contributions

The set of contributions is as follows:

- *Analysis of translation errors related to document-level phenomena and the development of a set of simple, yet effective, post-processing techniques to handle them.* Since the particular document-level phenomena they handle are sparse, we need a manual evaluation to assess their effectiveness because the automatic evaluation metrics do not capture their improvements. These findings were published as a technical report (Martínez Garcia, España-Bonet, and Màrquez, 2014b) and presented in the SEPLN2014 conference (Martínez Garcia, España-Bonet, and Màrquez, 2014a).

- *Demonstrating that bilingual word embeddings are capable of modeling semantic relations that help the SMT process.* We observe that the quality of the translation and alignments previous to building the semantic models are crucial for the final performance of the embeddings. Word embeddings prove to be helpful in the task of lexical substitution for words that are ambiguously translated within a document. This work resulted in a publication in the SSST-8 conference (Martínez Garcia et al., 2014).

- *Showing that the introduction of bilingual word embeddings guides document-oriented SMT decoders towards more coherent and cohesive translations.* Although we only observe a slight improvement in the results of automatic evaluation metrics, the improvement is consistent among metrics and is larger as we introduce more semantic information, getting the best results when using the models with bilingual information. This approach was presented in the EAMT2015 conference (Martínez Garcia, España-Bonet, and Màrquez, 2015).

- *Designing new strategies that guide document-oriented decoders through the translation search space towards more consistent, coherent, and cohesive translations, focusing on maintaining lexical consistency.* Our strategies based on word embeddings aid the decoder to assess the compatibility of the possible translations for ambiguous words with their context. This extension led to participating in the EAMT2017 conference (Martínez Garcia et al., 2017).

- *Enhancing the NMT decoding algorithm to include contextual semantics captured by a language model based on word embeddings.* We show how the semantic language models can help NMT systems to produce better translations. Our ap-

proach does not need to modify the training process, so we do not need to increase the training time or document-level annotated data. This work was presented in the DiscoMT2019 (Martínez Garcia, Creus, and España-Bonet, 2019).

Acknowledgements

The thesis work was partially supported by an FPI 2010 grant from the Spanish Ministry of Science and Innovation (MICINN) within the OpenMT-2 project (ref. TIN2009-14675-C03-03) of MICINN, a mobility EEBB 2013 grant from the Spanish Ministry of Economy and Competitiveness (MINECO) for a stay at the Department of Linguistics and Philology at the Uppsala University, and by the TACARDI project (ref. TIN2012-38523-C02-02) of the MINECO.

References

Bahdanau, D., K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*.

Dijk, T. A. v. 1977. *Text and context: Explorations in the semantics and pragmatics of discourse*. Number 21 in Longman Linguistics Library. Longman.

Hassan, H., A. Aue, C. Chen, V. Chowdhary, J. Clark, C. Federmann, X. Huang, M. Junczys-Dowmunt, W. Lewis, M. Li, S. Liu, T. Liu, R. Luo, A. Menezes, T. Qin, F. Seide, X. Tan, F. Tian, L. Wu, S. Wu, Y. Xia, D. Zhang, Z. Zhang, and M. Zhou. 2018. Achieving human parity on automatic Chinese to English news translation. *CoRR*, abs/1803.05567.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics on Interactive Poster and Demonstration Sessions*.

Martínez Garcia, E., C. Creus, C. España-Bonet, and L. Màrquez. 2017. Using word embeddings to enforce document-level lexical consistency in machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108.

Martínez Garcia, E., C. Creus, and C. España-Bonet. 2019. Context-aware neural machine translation decoding. In *Proceedings of the 4th Workshop on Discourse in Machine Translation*.

Martínez Garcia, E., C. España-Bonet, and L. Màrquez. 2014a. Document-level machine translation as a re-translation process. *Procesamiento del Lenguaje Natural*, 53.

Martínez Garcia, E., C. España-Bonet, and L. Màrquez. 2014b. Experiments on document level machine translation. Technical Report LSI-14-11-R, Universitat Politècnica de Catalunya, Spain.

Martínez Garcia, E., C. España-Bonet, and L. Màrquez. 2015. Document-level machine translation with word vector models. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT)*.

Martínez Garcia, E., C. España-Bonet, J. Tiedemann, and L. Màrquez. 2014. Word's vector representations meet machine translation. In *Proceedings of the 8th Workshop on Syntax, Semantics and Structure in Statistical Translation*.

Sanders, T. J. M. and H. L. W. Pander Maat. 2006. Cohesion and coherence: Linguistic approaches. In *Encyclopedia of Language & Linguistics*. Elsevier.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*.

Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Ł. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 201-204 recibido 24-11-2020 revisado 11-01-2021 aceptado 30-01-2021

Hacia el análisis de sentimientos en euskera

Towards sentiment analysis in Basque

Jon Alkorta Agirrezabala
HiTZ Center - Ixa, Universidad del País Vasco (UPV/EHU)
Manuel Lardizabal 1, 20018 Donostia
jon.alkorta@ehu.eus

Resumen: Este es un resumen de la tesis escrita por Jon Alkorta bajo la supervisión del Dr. Koldo Gojenola (Departamento de Lenguajes y Sistemas Informáticos) y el Dr. Mikel Iruskieta (Departamento de Didáctica de la Lengua y la Literatura) y presentada en la Universidad del País Vasco (UPV/EHU). El título completo de la tesis es *Sentimenduen analisi automatikorantz: oinarrizko baliabideen sorkuntza eta hizkuntza maila ezberdinetako balentzia-aldatzaileen identifikazioa* y la defensa de la tesis se celebró en la Facultad de Informática de Donostia-San Sebastián el 4 de diciembre de 2019, ante el tribunal formado por Juliano Desiderato Antonio (Universidade Estadual de Maringa), Iria da Cunha (investigadora Ramón y Cajal de la Universidad Nacional de Educación a Distancia (UNED)) y Arantza Diaz de Ilarraza (Universidad del País Vasco (UPV/EHU)). La tesis obtuvo la calificación de sobresaliente Cum Laude otorgada por unanimidad y mención internacional.
Palabras clave: análisis de sentimientos, corpus, lexicón de sentimientos, discurso, clasificador, euskera.

Abstract: This is a summary of the thesis written by Jon Alkorta under the supervision of Dr. Koldo Gojenola (Department of Computer Languages and Systems) and Dr. Mikel Iruskieta (Department of Didactis of Language and Literature) and presented at the University of the Basque Country (UPV / EHU). The full title of the thesis is *Towards the automatic analysis of sentiments in Basque: the creation of basic resources and the identification of valence shifters in different language levels* and the defense of the thesis was held on the 4th December 2019 in the Computer Science Faculty in Donostia-San Sebastián, and the members of the commission were Juliano Desiderato Antonio (State University of Maringa), Iria da Cunha (Ramón y Cajal researcher at the National Distance Education University (UNED)) and Arantza Diaz de Ilarraza (University of the Basque Country (UPV/EHU)). The thesis was awarded an excellent grade and Cum Laude honours and the international mention.
Keywords: sentiment analysis, corpus, sentiment lexicon, discourse, classifier, Basque.

1 Introducción de la tesis

El análisis de sentimientos tiene como objetivo analizar distintos aspectos relacionados con la información subjetiva. En las últimas decadas, su importancia ha ido en aumento porque hoy en día se puede encontrar un gran volúmen de información subjetiva en Internet. El objetivo de la tesis es crear recursos y herramientas para el procesamiento de la información subjetiva en euskera. Para ello, se han definido los siguientes objetivos:

- Crear recursos y herramientas básicas para el procesamiento de la información subjetiva. El objetivo es crear un corpus, un lexicón de sentimientos y un clasificador de textos de opinión.
- Identificar cambiadores de valencia de diferentes niveles lingüísticos en euskera para mejorar la precisión de las herramientas. Dentro de los niveles lingüísticos, se quiere poner énfasis en el rol de diferentes estructuras de discurso.

2 Estructura de la tesis

Esta tesis consta de dos volúmenes. El principal volumen está escrito en euskera y

ISSN 1989-7553. DOI 10.26342/2021-66-18

se titula *Sentimenduen analisi automatikorantz: oinarrizko baliabideen sorkuntza eta hizkuntza maila ezberdinetako balentzia-aldatzaileen identifikazioa*. Por otra parte, el segundo volumen está escito en inglés y su título es *Towards the automatic analysis of sentiments in Basque: the creation of basic resources and the identification of valence shifters in different language levels*. Los dos volúmenes no tienen la misma estructura pero comparten algunas secciones. La estructura del volumen en euskera es la siguiente:

1. En el capítulo de la introducción, se presenta la motivación, los hipótesis generales, los objetivos, las publicaciones relacionadas con la tesis y la estructura de la tesis.
2. El segundo capítulo presenta los trabajos realizados previamente que tienen relación con esta tesis.
3. En el tercer capítulo, se explica la metodología de esta tesis. Por un lado, se explica cómo se han creado y evaluado las herramientas para el procesamiento de análisis de sentimientos. Por otra parte, también se explican los pasos realizados para identificar los cambiadores de valencia en euskera.
4. El cuarto capítulo trata sobre el desarrollo de las herramientas para el procesamiento de la información subjetiva.
   - El corpus de opiniones en euskera. Este corpus se ha diseñado tomando en cuenta la estructura que tiene el corpus llamado *SFU Review Corpus* (Taboada, 2008). Además, el corpus se ha anotado con información subjetiva e información de discurso. basando en la teoría RST (Mann y Thompson, 1987).
   - El lexicón de sentimientos en euskera. En este apartado se explica cómo se ha desarrollado la traducción del lexicón de sentimientos en inglés de la herramienta SO-CAL (Taboada et al., 2011) al euskera utilizando los diccionarios *Elhuyar* (Zerbitzuak, 2013) y *Zehazki* (Sarasola, 2005).
   - El clasificador de sentimientos en euskera. Por último, se da a conocer i) la estructura de la herramienta SO-CAL, ii) la adaptación de la herramienta al euskera y iii) la evaluación de la herramienta.
5. En el quinto capítulo, se explican los resultados en la identificación de cambiadores de valencia en euskera.
   - Nivel fonológico y morfológico. Esta sección clasifica la palatalización expresiva y los morfemas según su influencia en la orientación semántica de las palabras.
   - Nivel sintáctico. Se enumeran los marcadores de negación que son cambiadores de valencia extraídos del corpus. También se explica cómo se han desarrollado las reglas en formato de Gramática de restricciones (Karlsson et al., 2011) para la identificación de los marcadores de negación.
   - Nivel de discurso. En esta sección, basando en los resultados de la investigación, se explica la relación que puede haber entre la unidad central, el núcleo o la primera parte de la relación y los cambiadores de valencia.
6. En el sexto capítulo se presentan las contribuciones, los límites del trabajo y trabajos futuros.

3 Contribución de la tesis

Las contribuciones se pueden clasificar en dos grupos: i) contribuciones teóricas relacionadas con aspectos lingüísticos y su aplicación en el análisis de sentimientos y ii) creación de recursos lingüísticos para el procesamiento de la información subjetiva en euskera. Las contribuciones teóricas son las siguientes:

- En **fonología** se ha observado que la palatalización expresiva refuerza la valencia de sentimientos de las palabras. Por otra parte, en **morfología** se ha visto que los morfemas pueden reforzar o debilitar la valencia de sentimientos.
- En **sintaxis** (Alkorta, Gojenola, y Iruskieta, 2018a), se ha observado que las partículas de negación generalmente debilitan (y a veces invierten) la orientación semántica y la valencia de sentimientos del conjunto de palabras de

afectadas por la negación. Sin embargo, en algunas casos, la partícula de negación no tiene ningún efecto y en un sólo caso (la partícula *ez* "no" + adjectivo/adverbio), la partícula de negación intensifica la orientación semántica y la valencia de sentimientos.

- En **discurso**, basado en la teoría de la estructura retórica (RST) , las contribuciones han sido las siguientes:
  - En relaciones de discurso (Alkorta et al., 2015; Alkorta, Gojenola, y Iruskieta, 2016b; Alkorta et al., 2017):
    - El núcleo coincide en más ocasiones con la orientación semántica de la relación retórica que el satélite.
    - La última parte de la relación coincide en más ocasiones con la orientación semántica de la relación que la primera parte.
  - En textos de opinión:
    - Cuando una relación de discurso está más cerca de la unidad central, hay más coincidencia de la orientación semántica entre esa relación y el texto de opinión.
    - Las relaciones de discurso suelen aparecer más en un lugar concreto dentro de la estructura de discurso del texto. Así, CAUSA/ANTÍTESIS → Unidad central → EVALUACIÓN → EVIDENCIA parece ser la estructura más común.

Las contribuciones en cuanto a recursos lingüísticos de esta tesis son las siguientes:

- Creación de un **corpus de textos de opinión en euskera** (Alkorta, Gojenola, y Iruskieta, 2016a). Este corpus contiene 240 textos de opinión recolectados de distintos medios de comunicación en euskera, así como de blogs especializados. Además, 39 textos de opinión relacionados con la literatura están anotados con la teoría RST[1] y la orientación semántica de las relaciones de discurso de estos textos también está etiquetada (Alkorta, Gojenola, y Iruskieta, 2019).
- Creación de un **léxicon de sentimientos**[2] (Alkorta, Gojenola, y Iruskieta, 2018b). Este lexicón ha sido creado a partir de los lexicones de sentimientos en inglés y castellano de la herramienta SO-CAL (Taboada et al., 2011) para la clasificación de sentimientos.

  El lexicón consta de dos versiones. La primera versión no está adaptada a dominios concretos y consta de 8.140 entradas. La entradas pueden ser palabras que aparecen en el diccionario, unidades lexicales de varias palabras, palabras con el sufijo de genitivo, etc. 2.282 entradas son nombres (28,06%), 3.162 son adjetivos (38,85%), 652 son adverbios (7,98%), 1.657 son verbos (20,36%) y 384 entradas son intensificadores (4,75%). Por el contrario, las entradas de la segunda versión del lexicón están adaptadas a los dominios concretos que corresponden con los del corpus. En este caso, el lexicón contiene 1.237 entradas. 461 entradas son nombres (%37,27), 446 son adjetivos (%36,06), 54 son adverbios (4,36%) y por último, 276 entradas son verbos (22,32%). En esta versión, los intensificadores no se han incluido.

Asimismo, se ha creado una herramienta para el procesamiento de la información subjetiva o sentimientos en euskera:

- **La versión en euskera de la herramienta SO-CAL**. Esta herramienta indica si una oración, parágro o texto tiene una valoración positiva o negativa. La herramienta consta de tres modulos.
  - El primer módulo se basa en el léxicon creado en esta tesis. En este caso, hemos cambiado el lexicón en inglés por el lexicón en euskera.
  - En el segundo módulo, hemos integrado el lematizador *Eustagger* (Ezeiza et al., 1998) en la herramienta para lematizar el texto y asignar la valencia de sentimientos a las palabras del texto, si la palabra aparece en el lexicón. En este caso, la versión en inglés y

[1] Los textos anotados se encuentran disponibles en el corpus RST Basque TreeBank: http://ixa2.si.ehu.es/diskurtsoa/index.php

[2] http://ixa.si.ehu.es/node/11438

en euskera de la herramienta varían por la tipología morfológica.

- La tercera sección contiene varias reglas que modifican la valencia de sentimientos de las palabras en el texto para que la clasificación de la subjetividad del texto sea más precisa. Si una palabra tiene una valencia de sentimientos negativa, estas reglas asignan peso a este tipo de palabras. Si una palabra con valencia de sentimientos se repite muchas veces en un mismo texto, las reglas restan el valor a las valencias de estas palabras repetidas.

En resumen: primeramente, la herramienta lematiza el texto y asigna la valencia de sentimientos a las palabras del texto que aparecen en el lexicón. Después, varias reglas modifican la valencia de sentimientos de estas palabras y por último, la herramienta calcula la subjetividad del texto (y si el texto tiene una valoración positiva o negativa).

Agradecimientos

La tesis se ha desarrollado gracias a las becas PRE_2015_1_0121, PRE_2016_2_0153, PRE_2017_2_0041 y PRE_2018_2_0033 del Gobierno Vasco y ha sido financiada por el proyecto *Ixa Taldea: Financiación UPV/EHU para grupos de investigación* (GIU16/16) de la UPV/EHU.

Bibliografía

Alkorta, J., K. Gojenola, y M. Iruskieta. 2016a. Creating and evaluating a polarity-balanced corpus for basque sentiment analysis. En *IWoDA16 Fourth International Workshop on Discourse Analysis. Santiago de Compostela, September*, volumen 29.

Alkorta, J., K. Gojenola, y M. Iruskieta. 2016b. Sentimenduen analisia euskaraz: lexiko-mailatik erlaziozko diskurtso-egiturarako proposamena. *Gogoa 14*.

Alkorta, J., K. Gojenola, y M. Iruskieta. 2018a. Saying no but meaning yes: negation and sentiment analysis in basque. En *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, páginas 85–90.

Alkorta, J., K. Gojenola, y M. Iruskieta. 2018b. Sentitegi: Semi-manually created semantic oriented basque lexicon for sentiment analysis. *Computación y Sistemas*, 22(4).

Alkorta, J., K. Gojenola, y M. Iruskieta. 2019. Towards discourse annotation and sentiment analysis of the basque opinion corpus. En *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, páginas 144–152.

Alkorta, J., K. Gojenola, M. Iruskieta, y A. Prez. 2015. Using relational discourse structure information in basque sentiment analysis. En *SEPLN 5th Workshop RST and Discourse Studies*.

Alkorta, J., K. Gojenola, M. Iruskieta, y M. Taboada. 2017. Using lexical level information in discourse structures for basque sentiment analysis. En *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, páginas 39–47.

Ezeiza, N., I. Alegria, J. M. Arriola, R. Urizar, y I. Aduriz. 1998. Combining stochastic and rule-based methods for disambiguation in agglutinative languages. En *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Karlsson, F., A. Voutilainen, J. Heikkilae, y A. Anttila. 2011. *Constraint Grammar: a language-independent system for parsing unrestricted text*, volumen 4. Walter de Gruyter.

Mann, W. C. y S. A. Thompson. 1987. Rhetorical structure theory: Description and construction of text structures. En *Natural language generation*. Springer, páginas 85–95.

Sarasola, I. 2005. *Zehazki: gaztelania-euskara hiztegia*. Alberdania.

Taboada, M. 2008. Sfu review corpus [corpus]. vancouver: Simon fraser university.

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

Zerbitzuak, E. H. 2013. Elhuyar hiztegia: euskara-gaztelania, castellano-vasco. *Elhuyar*.

Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 205-208 recibido 26-11-2020 revisado 20-01-2021 aceptado 02-02-2021

Language and Structure in Polarized Communities

Lenguaje y estructura en comunidades polarizadas

Mirko Lai
PRHLT Research Center, Universitat Politècnica de València, Spain
Dipartimento di Informatica, Università degli Studi di Torino, Italy
Computer Science Departmen, University of Turin
Corso Svizzera 185, Turin, Italy
mirko.lai@unito.it

Abstract: PhD thesis in Computer Science has been written by Mirko Lai under the supervision of Prof. Paolo Rosso (Universitat Politècnica de València), Dr. Giancarlo Ruffo (University of Turin) and Dr. 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 Turin, Italy on February 11, 2019. The doctoral committee was integrated by: Leo Ferres (Universidad del Desarrollo, Chile), Delfina Malandrina (Università degli Studi di Salerno, Italy) and Sara Tonelli (Fondazione Bruno Kessler, Italy).
Keywords: stance detection, communities detection, online social network, polarized political debates.

Resumen: Tesis doctoral en Informática ha sido realizada por Mirko Lai y dirigida por el Prof. Paolo Rosso (Universitat Politècnica de València), Dr. Giancarlo Ruffo (University of Turin) 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 Turin, Italia el 11 de febrero de 2019 ante un tribunal compuesto por: Leo Ferres (Universidad del Desarrollo, Chile), Delfina Malandrina (Università degli Studi di Salerno, Italia) y Sara Tonelli (Fondazione Bruno Kessler, Italia).
Palabras clave: detección de las opiniones, detección de comunidades, red social, debates políticos polarizados.

1 Introduction

Nowadays, social media are gaining a very important role in public debates and a significant part of the population is exposed to information though them. Furthermore, political leaders use social media directly to communicate with their citizens. On the other hand, citizens take part in the discussion, by supporting or criticizing their political opinions. For these reasons, social media provide a powerful experimental tool to deduce the mood of the public opinion and investigate how individuals are exposed to diverse viewpoints. The large amount of users' generated data motivated the need for new automated forms of textual content analysis. Research on this topic could have a positive impact on different aspects such as public administration, policy-making, and security. In fact, through the constant monitoring of people's opinion, desires, complaints and beliefs on political agenda or public services, administrators could better meet population's needs and prevent extremely marked ideological polarization and extremist tendencies.

In this thesis, we address the problem of stance detection in social media focusing on polarized political debates in Twitter. Stance detection consists in automatically determine whether the author of a post is in favor or against a target of interest, or whether the opinion toward the given target can not be inferred. We deal with political topics such as electoral events (e.g., political elections or referendums) and consequently the targets of interest are both politicians and referendums. We also explore the communications which take place in these polarized debates shedding some light on dynamics of communications among people having concordant or contrasting opinions, particularly focusing on

ISSN 1989-7553. DOI 10.26342/2021-66-19

observing opinions' shifting. We propose machine learning models for addressing stance detection as a classification problem. We also explore features based on the textual content of the tweet, but also features based on contextual information that do no emerge directly from the text.

2 Thesis Overview

This thesis consists in a collection of our most relevant publications about the research project I was involved to during my Ph.D. It consist of 7 chapters that are briefly introduced below.

Chapter 2 (Lai et al., 2016) contains the first result of our research on political debates in social media that investigates stance detection. The paper has been published in the proceedings of the *15th Mexican International Conference on Artificial Intelligence*. Starting from a benchmark dataset of English tweets released at the first shared task on stance detection (SemEval-2016 Task 6), we propose a feature based on the context surrounding the targets of interest. In particular, we define the two concepts "enemies."and "friends" for denoting the possible relations among the target and the entities related to the target. Namely, we try to model that when a tweeter is against an "enemy"/"friend."of the target, then the tweeter is in favor/against the target, and vice versa. Since our particular interests in political debates, we focus on the two targets related to the political campaign for the 2016 U.S. presidential elections: Hillary Clinton and Donald Trump. Our results, that take advantage from the proposed feature, outperform the best ones obtained by the teams participating in the task. We show that the information about "enemy" and "friend" of politicians helps in detecting the stance towards them.

Chapter 3 (Lai, Cignarella, and Hernández Farías, 2017) provides a technical report including a brief description of our approach, an illustration of our experiments, and an analysis of our results for our submission for the *Stance and Gender Detection in Tweets on Catalan Independence* shared task held at IberEval-2017. The released dataset consists in Catalan and Spanish tweets about the regional elections in Catalonia (Spain) held in September 2015. The election has been explained as a de facto referendum on the possible independence of Catalonia from Spain. For this reason, the organizers chose "independence of Catalonia" as target for the stance detecting task. Our system (iTACOS) ranked in as the first position among ten participating teams for both languages at the stance detection sub-task. Our approach, based on *context* and *structural* features, shows that contextual features helps in stance detection even when the target of interest is not a person.

Chapter 4 (Lai et al., 2017) contains the paper included in the proceedings of the international conference *Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2017)*. In this paper, we explore in depth opinion shifting applying the 2016 United Kingdom European Union membership referendum as case of study. We created the TW-ChronosBrexit corpus for stance detection that we used for training a model for automatically estimate the stance of all users of our dataset. We shown that users having the same stance towards this topic tend to belong to the same social network community. Moreover, we found evidences that the neighbours are more likely to have similar opinions. The extension of this work was afterwards published in the Journal of Intelligent & Fuzzy Systems (Lai et al., 2020).

Chapter 5 (Lai et al., 2018) has been published in the proceedings of the *23rd International Conference on Natural Language & Information Systems (NLDB 2018)*. We created the ConRef-STANCE-ita corpus for stance detection for inspecting stance detection at user level and in a diachronic perspective applying the 2016 referendum on the reform of the Italian Constitution as case of study. Here, we investigate in depth social network exploiting different types of relations such as retweets, quotes, and replies. The analysis shows that users with the same stance towards a particular issue tend to belong to the same social network community. For this reason, we propose three new features for stance detection based on the online social community the user belongs. The performed experiments show that the accuracy of stance detection prediction is considerably improved adding features derived from communities extracted from retweets-

based and quotes-based networks to content-based ones. This does not happen using the feature based on the communities extracted from the replies-based network. Indeed, the users mainly reply to other users with a similar opinion and we observe about 20 % of cross-stance edges among them. We also shed some light on users' opinion shift dynamics observing that in this debate, users tend to be less explicit on their stance as the outcome of the vote approaches. The research has been expanded and published in the Journal Data & Knowledge Engineering after the thesis defense (Lai et al., 2019).

Chapter 6 summarizes the obtained results and presents extended experiments we carried out. First, we deeply analyze our system (iTACOS) ranked in as the first position in the *Stance and Gender Detection in Tweets on Catalan Independence* shared task held at IberEval-2017. Then, we propose an extended version of iTACOS for classifying stance in a multilingual scenario (MultiTACOS). We also carry out a qualitative analysis of the features used for addressing stance detection in the debate about the BREXIT referendum, and after, we analyze the communication among users with similar and divergent viewpoints in the Italian Constitutional referendum case of study. Finally, we explore the features extracted from a network structure in a task different from stance detection e.g. talent identification in sport particularly focusing on the case of study of table tennis. The work described in this chapter was published in the Journal Computer Speech & Language (Lai et al., 2020) after the defense of the thesis.

Chapter 7 finally draws conclusions from the results presented in this thesis. Furthermore, the chapter outlines our publications during the Ph.D. Here we also propose some future research lines for this work.

3 Contributions

Stance detection has been identified as a not trivial task independent from sentiment analysis. Indeed, if on the one hand, sentiment analysis aims to detect the sentiment expressed in a piece of text, on the other, stance detection seeks to identify the user's opinion toward a defined target of interest (not necessarily mentioned in the text). In this thesis we concentrated our attention on online political debated and we faced stance detection as a classification task proposing different type of features, in particular, increasingly focusing on contextual ones. The achievements of our research could be summarized as following:

- We presented a brief description of the approaches proposed in the literature particularly focusing on the two shared tasks on Stance Detection held at SemEval 2016 and IberEval 2017. Our method, obtaining the highest result at IberEval 2017 and amounting the state of the art achieved at Semeval 2016, validates the assumption that contextual features could be useful for the task of stance detection.
- We created four new annotated corpora of tweets for stance detection: the English TW-ChronosBrexit, the Italian ConRef-STANCE-ita, and the E-FRA and R-ITA corpora respectively in French and Italian.
- Facing stance detection in a multilingual perspective, we detected linguistic characteristics peculiar of each language. Furthermore, we showed that results are affected by the different styles used by users for communicating stance towards target entities of different types (persons or referendum).
- We observed, on two different political debates, that users tend to aggregate themselves in like-minded groups. For this reason, we proposed a contextual feature based on the community the users belong for detection their stance. The results outperform those obtained by using only features based on the content of the post.
- We show how, representing a complex problem with a network, could be useful for extracting features from the network structure for dealing with other classification task suck as talent prediction.
- Users use different type of communication depending on the level of agreement with the interlocutor's opinion. Friendship, retweets, and quote relations are more common among like-minded users, while replies are often used for interacting with users having different stances.

- Approaching on stance detection in a diachronic perspective, we observed both opinion shifting and a mitigation of the debate towards an unaligned position after the outcome of the vote. In a deeper analysis, results tend to shows that users having heterogeneous relations tend, approaching the end of the debate, to more likely keep their opinions unclear than user having homogeneous links.

References

Lai, M., A. T. Cignarella, and D. I. Hernández Farías. 2017. iTACOS at IberEval2017: Detecting stance in Catalan and Spanish tweets. In *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017)*, volume 1881 of *CEUR Workshop Proceedings*, pages 185–192, Aachen, Germany, September. CEUR-WS.org.

Lai, M., A. T. Cignarella, D. I. Hernández Farías, C. Bosco, V. Patti, and P. Rosso. 2020. Multilingual stance detection in social media political debates. *Computer Speech & Language*, 63:101075.

Lai, M., D. I. Hernández Farías, V. Patti, and P. Rosso. 2016. Friends and enemies of Clinton and Trump: Using context for detecting stance in political tweets. In *Proceedings of the 15th Mexican International Conference on Artificial Intelligence*, MICAI, pages 155–168, Cham, Germany, October. Springer International Publishing.

Lai, M., V. Patti, G. Ruffo, and P. Rosso. 2018. Stance evolution and twitter interactions in an italian political debate. In M. Silberztein, F. Atigui, E. Kornyshova, E. Métais, and F. Meziane, editors, *Natural Language Processing and Information Systems*, pages 15–27, Cham. Springer International Publishing.

Lai, M., V. Patti, G. Ruffo, and P. Rosso. 2020. #Brexit: Leave or remain? the role of user's community and diachronic evolution on stance detection. *Journal of Intelligent & Fuzzy Systems*, 39:2341–2352.

Lai, M., M. Tambuscio, V. Patti, G. Ruffo, and P. Rosso. 2017. Extracting graph topological information and users' opinion. In *Proceeding of the 8th International Conference of the CLEF Association*, CLEF, pages 112–118, Cham, Germany, September. Springer International Publishing.

Lai, M., M. Tambuscio, V. Patti, G. Ruffo, and P. Rosso. 2019. Stance polarity in political debates: A diachronic perspective of network homophily and conversations on twitter. *Data & Knowledge Engineering*, 124:101738.

Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 209-212 recibido 28-11-2020 revisado 20-01-2021 aceptado 05-02-2021

Buscando robustez en un mundo multilingüe: de pipelines a embeddings

Seeking robustness in a multilingual world: from pipelines to embeddings

Yerai Doval
Grupo COLE, Escola Superior de Enxeñaría Informática,
Universidade de Vigo, España
yerai.doval@uvigo.es

Resumen: Tesis elaborada por Yerai Doval Mosquera bajo la supervisión de los profesores Jesús Vilares Ferro (Universidade da Coruña), Manuel Vilares Ferro (Universidade de Vigo) y la colaboración de Carlos Gómez Rodríguez (Universidade da Coruña). Su defensa tuvo lugar el día 17 de diciembre de 2019 en la Universidade da Coruña, con un tribunal compuesto por los profesores Miguel Ángel Alonso Pardo (Universidade da Coruña), Pavel Bernard Brazdil (Universidade do Porto) y María Lourdes Araújo Serna (Universidad Nacional de Educación a Distancia). Mereció la calificación de Sobresaliente *cum laude* con Mención Doctor Internacional.
Palabras clave: texto ruidoso, pipelines, word embeddings, multilingüismo, sistemas robustos

Abstract: Thesis prepared by Yerai Doval Mosquera under the supervision of professors Jesús Vilares Ferro (Universidade da Coruña), Manuel Vilares Ferro (Universidade de Vigo) and with the collaboration of Carlos Gómez Rodríguez (Universidade da Coruña). Its defense took place on December 17, 2019 at the University of Coruña, with a panel composed of professors Miguel Ángel Alonso Pardo (Universidade da Coruña), Pavel Bernard Brazdil (Universidade do Porto) and María Lourdes Araújo Serna (Universidad Nacional de Educación a Distancia). It obtained the highest qualification with *cum laude* honours and International Mention.
Keywords: noisy text, pipelines, word embeddings, multilingualism, robust systems

1 Introducción

Los usuarios de Internet producen y comparten todo tipo de contenido escrito en una amplia variedad de servicios y plataformas: páginas Web, correos electrónicos, mensajes de chat, publicaciones en redes sociales, etc. El tipo de textos producido y compartido por estos usuarios tiene dos rasgos específicos que lo diferencian de la mayoría de textos escritos, a la vez que lo acercan al lenguaje hablado: su *espontaneidad* e *informalidad*. Esto da como resultado lo que se denominan *textos ruidosos*, con un estilo de escritura fuertemente influenciado por hábitos del habla.

Asimismo, aún cuando el inglés es el idioma predominante, Internet demuestra un claro y creciente *multilingüismo* al acomodar contenidos en prácticamente cualquier idioma. No solo eso, es también habitual el *code-switching* o combinación de palabras en distintos idiomas en una misma frase u oración.

En este trabajo de tesis (Doval, 2019) estudiamos dos enfoques para abordar los desafíos en el procesamiento de contenidos textuales no estándar y multilingües generados por los usuarios, tal y como se pueden encontrar en la Web a día de hoy. Este tipo de textos son denominados a menudo *textos cortos* o *microtextos*.

En primer lugar, presentamos un enfoque tradicional basado en *pipelines* discretos en el que el texto de entrada es preprocesado para facilitar su tratamiento por otros sistemas. Esto implica abordar el problema del multilingüismo identificando el idioma de la entrada para, seguidamente, tratar los fenómenos de escritura no estándar específicos de dicho idioma presentes en dicha entrada. Para ello se aplicarán técnicas de normalización del texto y (re-)segmentación de palabras.

En segundo lugar, analizamos las limitaciones inherentes a este tipo de modelos dis-

ISSN 1989-7553. DOI 10.26342/2021-66-20

cretos, lo cual nos conduce a un enfoque centrado en el empleo de modelos continuos basados en *word embeddings* (i.e., representaciones vectoriales). En este caso, el preprocesamiento explícito de la entrada es sustituido por la codificación de las características lingüísticas y demás matices propios de los textos no estándar en el propio espacio de *embedding* (un espacio vectorial). Nuestro objetivo es obtener modelos continuos que no sólo superen las limitaciones de los modelos discretos, sino que también se alineen con el actual estado de la cuestión del Procesamiento de Lenguaje Natural (PLN), dominado por sistemas basados en redes neuronales.

2 Estructura de la tesis

La memoria, escrita en inglés, está organizada en cinco partes, más apéndices:

Parte 1

- El **Capítulo 1** describe los fenómenos de *texting* que caracterizan el uso del lenguaje en Internet, los cuales constituyen la motivación de este trabajo, y las tareas de preprocesamiento consideradas, a la vez que introduce los enfoques que se estudian en los siguientes capítulos.
- El **Capítulo 2** recoge la terminología relevante para el dominio de nuestro trabajo y, a continuación, introduce dos importantes recursos ampliamente utilizados no sólo aquí, sino en muchos otros sistemas de PLN: los modelos de lenguaje y las *word embeddings*.

Parte 2

- El **Capítulo 3** presenta la tarea de identificación del idioma en el contexto del taller TweetLID (Zubiaga et al., 2014) de identificación del idioma de tuits en el contexto ibérico, y analiza el rendimiento de las herramientas comunes de identificación del idioma para dicha tarea.
- El **Capítulo 4** propone un enfoque sencillo para la normalización de microtextos de cara a la Tarea 2 del W-NUT 2015 (Baldwin et al., 2015), con una estructura clásica en dos pasos (generación y selección de candidatos de normalización), y centrándose en la modularidad y la adaptabilidad de la aproximación.
- El **Capítulo 5** presenta un enfoque de segmentación de palabras basado en un algoritmo de búsqueda y un modelo de lenguaje, además de estudiar su rendimiento cuando este último componente se implementa como una red neuronal recurrente o bien un modelo de n-gramas.

Parte 3

- El **Capítulo 6** analiza, desde un punto de vista teórico, las limitaciones inherentes a los *pipelines* discretos y otros enfoques similares, y cómo el uso directo de *word embeddings* resuelve o evita los problemas resultantes.

Parte 4

- El **Capítulo 7** describe cómo mejorar la integración de espacios de *word embeddings* multilingües obtenidos mediante la alineación de espacios monolingües.
- El **Capítulo 8** describe una técnica de adaptación que mejora el rendimiento de los modelos de *word embeddings* monolingües existentes en el caso de textos ruidosos. También presenta un breve estudio sobre el efecto de la mala segmentación de palabras en el rendimiento de las *word embeddings*.

Parte 5

- El **Capítulo 9** cierra el trabajo de tesis presentando las conclusiones más relevantes y las futuras líneas de trabajo.

Apéndices La memoria de tesis incluye a mayores una serie de apéndices que, si bien aportan conclusiones significativas, no son necesarias para seguir la línea argumental principal de la disertación.

- El **Apéndice A** analiza el rendimiento de una amplia gama de algoritmos fonéticos en el proceso de generación de candidatos para normalización (Doval, Vilares, y Vilares, 2018), proceso descrito en el Capítulo 4.
- El **Apéndice B** presenta un amplio análisis de los factores que suelen intervenir en la alineación bilingüe de los espacios de *embedding* monolingües descritos en el Capítulo 7.

3 Contribuciones

Resumimos a continuación las contribuciones más relevantes de la tesis, correspondientes a las Partes 2, 3 y 4 de la misma.

3.1 Enfoque discreto: el *pipeline*

Nuestro *pipeline* de preprocesamiento está formado por las siguientes etapas o módulos:

Identificación del idioma Nuestra solución propuesta (Doval, Vilares, y Vilares, 2014) pasa por adaptar y reentrenar las herramientas de identificación automática del idioma existentes utilizando varios corpus de nuestra elección, de modo que todos compartan el mismo punto de partida.

Normalización de microtexto Nuestro enfoque (Doval, Vilares, y Gómez-Rodríguez, 2015) se basa en un proceso en dos fases: (1) generación de normalizaciones candidatas empleando correctores ortográficos y diccionarios; (2) selección de candidatas, implementada a través de un modelo de lenguaje a nivel de palabra y un algoritmo de búsqueda.

Segmentación de palabras Nuestra solución para corregir las segmentaciones incorrectas (Doval y Gómez-Rodríguez, 2019; Doval, Gómez-Rodríguez, y Vilares, 2016) consta de dos componentes: (1) un algoritmo de *beam search*, que genera y elige entre los posibles candidatos de segmentación de forma incremental; y (2) un modelo de lenguaje a nivel de byte o carácter que permite que el algoritmo pueda clasificar los candidatos, e implementado como una red neuronal recurrente o modelo de n-gramas.

3.2 Limitaciones del enfoque discreto y transición a un modelo continuo

Tras un estudio de las limitaciones inherentes al enfoque discreto tradicional, hemos comprobado que podemos resolver los problemas derivados de la propagación de errores y la fragmentación del contexto cambiando a un modelo continuo centrado en *word embeddings*. Estas representaciones vectoriales permiten mejorar la integración de nuestro sistema a la vez que constituyen un *lenguaje intermedio* para soportar entornos multilingües. Asimismo, se pueden utilizar para codificar las particularidades derivadas de los fenómenos propios de los textos generados por los usuarios, haciendo así innecesario su procesamiento explícito.

3.3 *Embeddings* multilingüe

Hemos desarrollado un método de postprocesamiento (Doval et al., 2018; Doval et al., 2019), que hemos denominado Meemi (por "*Meeting in the Middle*"), que mejora la integración de espacios monolingües inicialmente aislados y posteriormente alineados mediante herramientas como VecMap (Artetxe, Labaka, y Agirre, 2018) y MUSE (Conneau et al., 2018). Para mejorar dicha integración, aplicamos sobre estos alineamientos una transformación lineal no restringida que se aprende haciendo corresponder las traducciones de palabras con sus representaciones promedio.

De manera notable, hemos ido más allá de la configuración bilingüe habitual en este tipo de herramientas y hemos mostrado también cómo Meemi puede extenderse, de forma natural, a un número arbitrario de idiomas, los cuales acaban integrados en un único espacio vectorial compartido. En este caso, utilizamos métodos ortogonales en el primer paso de alineación que solo transforman el espacio de *embedding* de uno de los lenguajes (origen) mientras deja intacto el otro espacio (destino), que se convierte en el espacio vectorial multilingüe. Este proceso se repite para los espacios de origen correspondientes a cada idioma restante.

3.4 *Embeddings* robustas para microtextos

Los modelos de *word embeddings* como word2vec (Mikolov et al., 2013) o fastText (Bojanowski et al., 2016), son de por sí capaces de agrupar variantes estándar y no estándar de palabras si se les proporciona un corpus de entrenamiento lo suficientemente grande que incluya tales variantes (Sumbler et al., 2018). Sin embargo, nosotros proponemos ir un paso más allá con una modificación del modelo *skipgram* de fastText que permite no solo mejorar el rendimiento de las *word embeddings* resultantes en textos ruidosos, sino que además permite preservar su rendimiento en los textos estándar (Doval, Vilares, y Gómez-Rodríguez, 2020).

Para ello, introducimos un nuevo conjunto de palabras en el proceso de entrenamiento, que denominamos *bridge-words* (*palabras puente*), y cuyo objetivo es actuar a modo de guía a la hora de enlazar una palabra estándar con sus contrapartidas con ruido.

La robustez de las *embeddings* resultantes frente al ruido presente en el texto, se hace especialmente patente en el caso de modelos entrenados de extremo a extremo. El uso de estas *embeddings* nos evita tener que preprocesar la entrada original para modificarla, co-

mo venía ocurriendo hasta ahora, lo que solía llevar a introducir errores que luego se propagarían a otras partes de nuestros sistemas.

Bibliografía

Artetxe, M., G. Labaka, y E. Agirre. 2018. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. En *Proc. of the 32th AAAI Conf. on Artificial Intelligence, AAAI 2018*, páginas 5012–5019.

Baldwin, T., M.-C. de Marneffe, B. Han, Y.-B. Kim, A. Ritter, y W. Xu. 2015. Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition. En *Proc. of the 1st Workshop on Noisy User-generated Text, W-NUT 2015*, páginas 126–135.

Bojanowski, P., E. Grave, A. Joulin, y T. Mikolov. 2016. Enriching Word Vectors with Subword Information. *Transactions of the Association of Computational Linguistics*, 5(1):135–146.

Conneau, A., G. Lample, M. Ranzato, L. Denoyer, y H. Jégou. 2018. Word translation without parallel data. En *Proc. of the 6th Int. Conf. on Learning Representations, ICLR 2018*.

Doval, Y. 2019. *Seeking robustness in a multilingual world: from pipelines to embeddings*. Ph.D. tesis, Universidade da Coruña, 12.

Doval, Y., J. Camacho-Collados, L. E. Anke, y S. Schockaert. 2019. Meemi: A simple method for post-processing cross-lingual word embeddings. *arXiv preprint arXiv:1910.07221*.

Doval, Y., J. Camacho-Collados, L. Espinosa-Anke, y S. Schockaert. 2018. Improving cross-lingual word embeddings by meeting in the middle. En *Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing, EMNLP 2018*, páginas 294–304.

Doval, Y. y C. Gómez-Rodríguez. 2019. Comparing neural-and n-gram-based language models for word segmentation. *Journal of the Association for Information Science and Technology*, 70(2):187–197.

Doval, Y., C. Gómez-Rodríguez, y J. Vilares. 2016. Spanish word segmentation through neural language models. *Procesamiento del Lenguaje Natural*, 57:75–82.

Doval, Y., D. Vilares, y J. Vilares. 2014. Identificación automática del idioma en Twitter: adaptación de identificadores del estado del arte al contexto ibérico. En *Proc. of the Tweet Language Identification Workshop co-located with the 30th Conf. of the Spanish Society for Natural Language Processing, TweetLID@SEPLN 2014*, páginas 39–43.

Doval, Y., J. Vilares, y C. Gómez-Rodríguez. 2015. LYSGROUP: Adapting a Spanish microtext normalization system to English. En *Proc. of the 1st Workshop on Noisy User-generated Text, W-NUT 2015*, páginas 99–105, Beijing, China.

Doval, Y., J. Vilares, y C. Gómez-Rodríguez. 2020. Towards robust word embeddings for noisy texts. *Applied Sciences*, 10(19):6893.

Doval, Y., M. Vilares, y J. Vilares. 2018. On the performance of phonetic algorithms in microtext normalization. *Expert Systems with Applications*, 113:213–222.

Mikolov, T., G. Corrado, K. Chen, y J. Dean. 2013. Efficient estimation of word representations in vector space. *Proc. of the International Conference on Learning Representations, ICLR 2013*, páginas 1–12.

Sumbler, P., N. Viereckel, N. Afsarmanesh, y J. Karlgren. 2018. Handling Noise in Distributional Semantic Models for Large Scale Text Analytics and Media Monitoring. *Proc. of the 4th Workshop on Noisy User-generated Text, W-NUT 2018*.

Zubiaga, A., I. S. Vicente, P. Gamallo, J. R. Pichel, I. Alegría, N. Aranberri, A. Ezeiza, y V. Fresno. 2014. Overview of TweetLID: Tweet Language Identification at SEPLN 2014. En *Proc. of the Tweet Language Identification Workshop co-located with the 30th Conf. of the Spanish Society for Natural Language Processing, TweetLID@SEPLN 2014*, volumen 1228, páginas 1–11. CEUR-WS.org.

Información General

SEPLN 2021

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

VI Congreso Español de Informática
22-24 de septiembre 2021
https://congresocedi.es, http://www.hitz.eus/sepln2021/

1 Presentación

La XXXVII edición del Congreso Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) se celebrará los días 22, 23 y 24 de septiembre de 2021 de forma virtual como parte del VI Congreso Español de Informática (CEDI).

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 esta en entornos multilingües. Muchas de las bases formales para abordar adecuadamente estas necesidades han sido y siguen siendo establecidas en el marco del procesamiento del lenguaje natural y de sus múltiples vertientes: Extracción y recuperación de información, Sistemas de búsqueda de respuestas, Traducción automática, Análisis automático del contenido textual, Resumen automático, Generación textual y Reconocimiento y síntesis de voz.

2 Objetivos

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

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

3 Áreas Temáticas

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

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

ISSN 1989-7553

- 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, proyectos de investigación en marcha y demostraciones de aplicaciones. Además, tendrá lugar la tercera edición de IberLEF el día 22 de septiembre.

5 Comité ejecutivo SEPLN 2021

Presidenta del Comité Organizador

- Aitziber Atutxa Salazar (Universidad del País Vasco).

Colaboradores

- Eugenio Martínez Cámara (Universidad de Granada).
- Paloma Martínez Fernández (Universidad Carlos III).
- Álvaro Rodrigo Yuste (Universidad Nacional de Educación a Distancia).
- Koldo Gojenola (Universidad del País Vasco).
- Itziar González Dios (Universidad del País Vasco).
- Jon Alkorta (Universidad del País Vasco).

6 Consejo Asesor

Miembros:

- Manuel de Buenaga Rodríguez (Universidad de Alcalá, 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).
- José Camacho Collados (Cardiff University, Reino Unido).
- Arantza Díaz de Ilarraza (Universidad del País Vasco, España).
- Antonio Ferrández Rodríguez (Universidad de Alicante, España).
- Koldo Gojenola Galletebeitia (Universidad del País Vasco, España).
- Xavier Gómez Guinovart (Universidad de Vigo, España).
- José Miguel Goñi Menoyo (Universidad Politécnica de Madrid, España).
- Mariana Lara Neves (German Federal Institute for Risk Assessment, Alemania).
- Elena Lloret (Universidad de Alicante, España).
- Bernardo Magnini (Fondazione Bruno Kessler, Italia).
- Nuno J. Mamede (Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa, Portugal).
- 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).
- Leonel Ruiz Miyares (Centro de Lingüística Aplicada de Santiago de Cuba, Cuba).
- Manuel Montes y Gómez (Instituto Nacional de Astrofísica, Óptica y Electrónica, México).
- Lluis Padró Cirera (Universidad Politécnica de Cataluña, España).
- Manuel Palomar Sanz (Universidad de Alicante, España).
- Ferrán Pla (Universidad Politécnica de Valencia, España).
- Germán Rigau Claramunt (Universidad del País Vasco, España).
- Horacio Saggion (Universidad Pompeu Fabra, España).
- Emilio Sanchís (Universidad Politécnica de Valencia, España).
- Kepa Sarasola Gabiola (Universidad del País Vasco, España).
- Thamar Solorio (University of Houston, Estados Unidos de América).
- Maite Taboada (Simon Fraser University, Canadá).
- Mariona Taulé (Universidad de Barcelona, España).

- Juan-Manuel Torres-Moreno (Laboratoire Informatique d'Avignon / Université d'Avignon, Francia).
- José Antonio Troyano Jiménez (Universidad de Sevilla, España).
- L. Alfonso Ureña López (Universidad de Jaén, España).
- Rafael Valencia García (Universidad de Murcia, España).
- René Venegas Velásques (Pontificia Universidad Católica de Valparaíso, Chile).
- M. Felisa Verdejo Maíllo (Universidad Nacional de Educación a Distancia, España).
- Karin Vespoor (University of Melbourne, Australia).
- 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: 30 de abril de 2021.
- Notificación de aceptación: 4 de junio de 2021.
- Fecha límite para entrega de la versión definitiva: 18 de junio de 2021.

Información para los Autores

Formato de los Trabajos

- La longitud máxima admitida para las contribuciones será de 10 páginas DIN A4 (210 x 297 mm.), incluidas referencias y figuras.
- Los artículos pueden estar escritos en inglés o español. El título, resumen y palabras clave deben escribirse en ambas lenguas.
- El formato será en Word o LaTeX

Envío de los Trabajos

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

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 Maíllo
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 de Alcalá (España)
Sylviane Cardey-Greenfield	Centre de recherche en linguistique et traitement automatique des langues (Francia)
Irene Castellón	Universidad de Barcelona (España)
José Camacho Collados	Cardiff University (Reino Unido)
Arantza Díaz de Ilarraza	Universidad del País Vasco (España)
Antonio Ferrández	Universidad de Alicante (España)
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)
Mariana Lara Neves	German Federal Institute for Risk Assessment (Alemania)
Elena Lloret	Universidad de Alicante (España)
Bernardo Magnini	Fondazione Bruno Kessler (Italia)
Nuno J. Mamede	Instituto de Engenharia de Sistemas e Computadores (Portugal)
M. Teresa Martín Valdivia	Universidad de Jaén (España)

Cartas al director

Sociedad Española para el Procesamiento del Lenguaje Natural
Departamento de Informática. Universidad de Jaén
Campus Las Lagunillas, Edificio A3. Despacho 127. 23071 Jaén
secretaria.sepln@ujaen.es

Más información

Para más información sobre la Sociedad Española del Procesamiento del Lenguaje Natural puede consultar la página web http://www.sepln.org.

Si desea inscribirse como socio de la Sociedad Española del Procesamiento del Lenguaje Natural puede realizarlo a través del formulario web que se encuentra en esta dirección http://www.sepln.org/sepln/inscripcion-para-nuevos-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/la-revista/consejo-de-redaccion.

Las funciones del Consejo Asesor están disponibles Internet a través de la página http://www.sepln.org/la-revista/consejo-asesor.

La inscripción como nuevo socio de la SEPLN se puede realizar a través de la página http://www.sepln.org/sepln/inscripcion-para-nuevos-socios

Información General

XXXVII Congreso Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural 219
Información para los autores 219
Información adicional 221