Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1135-5948

Artículos

Tesis

Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1135-5948

Comité Editorial

Consejo de redacción

ISSN: 1135-5948
ISSN electrónico: 1989-7553
Depósito Legal: B:3941-91
Editado en: Universidad de Jaén
Año de edición: 2023

Editores:	Eugenio Martínez Cámara	Universidad de Jaén	emcamara@ujaen.es
	Álvaro Rodrigo Yuste	UNED	alvarory@lsi.uned.es
	Aitziber Atutxa Salazar	Universidad del País Vasco	aitziber.atucha@ehu.eus

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: 1135-5948

Preámbulo

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

Las áreas temáticas tratadas son las siguientes:

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

El ejemplar número 70 de la revista *Procesamiento del Lenguaje Natural* contiene trabajos correspondientes a comunicaciones científicas y resúmenes de tesis doctorales. 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 41 trabajos para este número, de los cuales 37 eran artículos científicos y 4 resúmenes de tesis doctorales. De entre los 37 artículos recibidos, 17 han sido finalmente seleccionados para su publicación, lo cual fija una tasa de aceptación del 45,94%.

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 2023
Los editores.

Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1135-5948

Preamble

The *Natural Language Processing* journal aims to be a forum for the publication of high-quality unpublished scientific and technical papers on Natural Language Processing (NLP) for both the national and international scientific community and companies. Furthermore, we want to strengthen the development of different areas related to NLP, widening the dissemination of research carried out, identifying the future directions of basic research and demonstrating the possibilities of its application in this field. Every year, the Spanish Society for Natural Language Processing (SEPLN) publishes two issues of the journal that include original articles, ongoing projects, book reviews and summaries of doctoral theses. All issues published are freely distributed to all members, and contents are freely available online.

The subject areas addressed are the following:

- Linguistic, Mathematical and Psychological models to language
- Grammars and Formalisms for Morphological and Syntactic Analysis
- Semantics, Pragmatics and Discourse
- Computational Lexicography and Terminology
- Linguistic resources and tools
- Corpus Linguistics
- Speech Recognition and Synthesis
- Dialogue Systems
- Machine Translation
- Word Sense Disambiguation
- Machine Learning in NLP
- Monolingual and multilingual Text Generation
- Information Extraction and Information Retrieval
- Question Answering
- Automatic Text Analysis
- Automatic Summarization
- NLP Resources for Learning
- NLP for languages with limited resources
- Business Applications of NLP
- Sentiment Analysis
- Opinion Mining
- Text Mining
- Evaluation of NLP systems
- Textual Entailment and Paraphrases

The 70th 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.

Forty-one papers were submitted for this issue, from which thirty-seven were scientific papers and four doctoral dissertation summaries. From these thirty-seven papers, we selected seventeen papers (45.94%) 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 2023
Editorial board.

Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1135-5948

Artículos

Tesis

Información General

Artículos

Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 15-26 recibido 12-12-2022 revisado 09-02-2023 aceptado 15-02-2023

Annotating reliability to enhance disinformation detection: annotation scheme, resource and evaluation

Anotando la confiabilidad para mejorar la tarea de detección de desinformación: esquema de anotación, recurso y evaluación

Alba Bonet-Jover, Robiert Sepúlveda-Torres, Estela Saquete, Patricio Martínez Barco
Department of Software and Computing Systems, University of Alicante, Spain
{alba.bonet, rsepulveda, stela, patricio}@dlsi.ua.es

Abstract: Disinformation is a critical problem in our society. The COVID-19 pandemic and the Russia-Ukraine war have been key events for the spreading of fake news. Assuming that fake news mixes reliable and unreliable information, we propose RUN-AS (Reliable and Unreliable Annotation Scheme), a fine-grained annotation scheme that labels the structural parts and essential content elements of a news item to enable their classification into Reliable and Unreliable. This type of annotation will be used for training systems to automatically classify the reliability of a news item. To this end, RUN dataset in Spanish was built and annotated with RUN-AS. A set of experiments were conducted to validate the annotation scheme. The experiments evidence the validity of the annotation scheme proposed, obtaining the best F_1m, i.e., 0.948.
Keywords: Natural Language Processing, Annotation Guideline, Dataset Annotation, Reliability Detection, Disinformation Detection.

Resumen: La desinformación es un problema crítico en nuestra sociedad. La pandemia de covid-19 y la guerra entre Rusia y Ucrania han sido escenarios clave para la difusión de noticias falsas. Partiendo de la base de que las noticias falsas mezclan información confiable y no confiable, proponemos RUN-AS (*Reliable and Unreliable Annotation Scheme*), un esquema de anotación de grano fino que etiqueta las partes estructurales y los elementos de contenido esenciales de una noticia y permite clasificarlos en Confiable y No confiable. Esta anotación será usada en el entrenamiento de sistemas para la clasificación automática de la confiabilidad de una noticia. Para ello, se construyó el corpus RUN en español y se anotó con RUN-AS. Se llevó a cabo un conjunto de experimentos para validar el esquema de anotación. Los experimentos evidencian la validez del esquema de anotación propuesto, obteniendo el mejor F_1m 0,948.
Palabras clave: Procesamiento Lenguaje Natural, Guía Anotación, Anotación Corpus, Detección Confiabilidad, Detección Desinformación.

1 Introduction

The disinformation problem is critical for today's society. Disinformation is fake or inaccurate information that is intentionally spread to mislead or deceive (Shu et al., 2020). Fake news is one of the most widespread phenomena of disinformation and, as defined by Zhou and Zafarani (2020), fake news is intentionally false information created by journalists and non-journalists that broadly includes articles, claims, statements, speeches, and posts, among other types of information, related to public figures and organizations.

The Internet has made it possible to be continuously informed, driving an almost instant dissemination of unverified news, as anyone can share and access information at no cost. A complex mix of cognitive, social and algorithmic biases makes us more vulnerable to believing and being manipulated by online disinformation (Shao et al., 2017). Algorithms make possible the exponential spread of fake news, but they can

ISSN 1135-5948 DOI 10.26342/2023-70-1

also be deployed to mitigate their propagation (Giansiracusa, 2021). Therefore, the facilitator of the disinformation problem, the algorithm, can be used to combat the problem. However, these algorithms are not yet robust enough to perform a verification of which information is false or true (Figueira and Oliveira, 2017). The disinformation phenomenon has become a challenge for many researchers from different research areas. In Natural Language Processing (NLP), several approaches are used to tackle this problem, such as automated fact-checking, sentiment analysis, deception and stance detection, contradiction detection, credibility, among others (Saquete et al., 2020).

The concepts of reliability and veracity are closely related, as fake news includes both reliable and unreliable information. In the literature, the term veracity is usually used in tasks where information is contrasted and verified (Vosoughi, Roy, and Aral, 2018), whereas the concept of reliability is often used in methods that investigate the credibility of the source of the news item (Zhou and Zafarani, 2020). In this research, as we tackle the problem by using the news item and not external knowledge, an absolute judgment on the veracity of a text is not possible. Instead of focusing on the veracity concept, we deal with the concept of reliability by following a style-based method that enables the detection of unreliable elements in a text through linguistic indicators as they mark the inaccuracy or subjectivity of the information provided (Zhang et al., 2018). To address this task computationally, annotated datasets are required (Stenetorp et al., 2012). This is a costly, slow and time-consuming task and therefore, labelled corpora are scarce, especially in languages other than English, such as Spanish.

The novelty of our proposal is the design of an innovative semantic annotation scheme that focuses on classifying news as Reliable or Unreliable from a linguistic perspective and without external knowledge. This annotation scheme will be beneficial to future disinformation detection tasks. The annotation proposal, hereafter referred to as RUN-AS (Reliable and Unreliable News Annotation Scheme), enables the essential parts of a news item to be detected, namely the structure (Inverted Pyramid) and the content (5W1H) along with their reliability. Furthermore, reliability criteria followed in the annotation process is clearly defined in Section 3.2. Following the proposed annotation guideline, a new dataset (RUN dataset) is created and used to validate the RUN-AS scheme under an evaluation framework. Furthermore, the language used for the annotation scheme and the dataset is Spanish due to the lack of resources in languages other than English.

This paper is structured as follows: Section 2 presents the background; Section 3 describes the annotation scheme proposed; Section 4 introduces the dataset created to test our proposal and two inter-annotator agreements to avoid bias in assessing news; Section 5 presents several experiments that validate our annotation scheme; Section 6 summarises the results and discussion; and finally, Section 7 presents the conclusions of this research and future work.

2 Related Work

This section presents relevant literature regarding state-of-the-art (SOTA) disinformation datasets, work regarding journalistic techniques applied in our proposal and finally, literature regarding research about linguistic characteristics of news in order to detect disinformation.

2.1 Annotated corpora for disinformation detection

Several datasets have been released for disinformation detection. LIAR dataset (Wang, 2017) comprises 12,836 real-world short statements classified in a scale of six fine-grained labels (pants-fire, false, barely-true, half-true, mostly-true and true). EMERGENT dataset (Ferreira and Vlachos, 2016) contains 300 claims and 2,595 associated news articles. This dataset classifies news into three veracity values (true, false and unverified) and assigns a stance label to the headline with respect to the claim (for, against and observing). Ferreira and Vlachos (2014) also released a fake news detection dataset comprising 221 statements annotated with a five-label-tag classification: true, mostlytrue, halftrue, mostlyfalse and false. Pérez-Rosas et al. (2017) introduced two new datasets for fake news detection covering several domains and linguistic differences between legitimate and fake news articles. The CLEF-2021 CheckThat! Lab: Task 3 on Fake News Detection (Shahi, Struß, and Mandl,

2021) is a lab that focuses on evaluating automatic detection of the news story's veracity, classified as true, partially true, false, or other. The dataset consists of 900 news articles, leaving 354 articles for testing.

As our dataset is also focused on health and COVID-19, it is relevant to mention two recent corpora addressing this domain: a fake news dataset consisting of 10,700 fake and real news (Patwa et al., 2021) and a large COVID-19 Twitter Fake News dataset (CTF) (Paka et al., 2021), which works with labelled and unlabelled tweets using two-scale labels (fake and genuine).

Concerning corpora in other languages, Spanish resources are scarce, creating a need for proposals that focus on the Spanish language. A fake news dataset in Spanish was released by Posadas-Durán et al. (2019), consisting of 491 true news and 480 fake news annotated with two labels (real and fake). In Portuguese, a dataset of labeled true and fake news called the Fake.Br corpus was presented (Silva et al., 2020). It is composed of 7,200 news (fake and legitimate). Assaf and Saheb (2021) present a novel dataset of Arabic fake news containing 323 articles (100 reliable news and 223 unreliable news) and focused on traditional linguistic features. Regarding datasets that annotate reliability, Gruppi et al. (2018) constructed two datasets of political news articles from United States sources (1,997 reliable, 794 unreliable and 50 satire) and Brazilian sources (4,698 reliable, 755 unreliable and 58 satire). For each article, they assigned a class reliable (R), unreliable (U) or satire (S) based on the source from which the article was collected.

To the authors' knowledge, most current datasets classify and annotate news with a single global veracity value. Many datasets created for disinformation detection have so far focused on fact-checking techniques, veracity classification (true/false) and global news annotation.

2.2 Corpora based on the journalistic techniques

Considering that our proposal uses two well-known journalistic concepts such as the Inverted Pyramid and the 5W1H[1], this subsection focuses on presenting some corpora that also use them. Norambuena et al. (2020) propose the Inverted Pyramid Scoring method to evaluate how well a news article follows the Inverted Pyramid structure using main event descriptors (5W1H) extraction and news summarisation. Their proposal, which was evaluated in a dataset consisting of 65,535 articles from the Associated Press News (AP News), shows that the method adopted helps to distinguish structural differences between breaking and non-breaking news, reaching the conclusion that breaking news articles are more likely to follow the Inverted Pyramid structure. Another interesting work related to the 5W1H journalistic concept is that of Chakma and Das (2018), in which an annotation approach to assign semantic roles is described. This proposal is applied to a corpus of 3,000 tweets related to the US elections of 2016. Khodra (2015) introduces a new 5W1H corpus of 90 Indonesian news articles to train event extraction. They were obtained from popular news websites and annotated following the 5W1H concept and extracting the event information of the news item.

The novelty of our annotation compared to the state of the art lies in the annotation of the 5W1H of all parts of a news item, permitting more in-depth analysis of the whole news article.

2.3 Research focused on linguistic features to detect disinformation

This subsection presents the research relevant to analysing linguistic features in news to determine reliability.

Zhang et al. (2018) present a set of content and context indicators for article reliability. Regarding the content indicators, which are the ones that are of interest to our research, the following are considered: title representativeness; clickbait title; quotes from outside experts; citation of organizations and studies; calibration of confidence; logical fallacies; and, tone and inference. Their dataset consists of 40 articles annotated with both content and context indicators. Furthermore, Horne and Adali (2017) state that the style and the language of articles allows differentiation of fake from real news. In this study, three content-based features categories are analysed: stylistic, complexity, and psychological. Horne and Adali (2017) conclude that there is a notable difference in titles

[1]Referring to: who, what, where, when, why, how.

and content between fake and real news in terms of length, punctuation, quotations, lexical features or capitalised words. Another study showing that linguistic characteristics can help determine the truthfulness of text is that of Rashkin et al. (2017). This work compares the language of real news with that of satire, hoaxes and propaganda. To analyse the linguistic patterns, they sampled standard trusted news articles from the English Gigaword corpus and crawled articles from seven different unreliable news sites. Mottola (2020) also carries out a comparative study between Italian and Spanish in order to identify the common textual characteristics of digital disinformation. Through this linguistic analysis, it is shown that there are several characteristics that fake news share related to headlines, punctuation, capital letters, lack of data or emotional aspects.

Our proposal makes a threefold contribution to disinformation detection. Firstly, a proposal of reliability classification instead of veracity, considering linguistic features, without external knowledge. Secondly, instead of exclusively annotating the entire article with a single global classification value, we also annotate all the structural parts and essential content of a news item in line with the 5W1H and Inverted Pyramid. Thirdly, this fine-grained annotation produces a quality resource in Spanish.

3 RUN-AS annotation scheme

3.1 Annotation labels

The goal of this annotation proposal is to support disinformation detection by analysing news on the basis of a purely textual and linguistic analysis and thereby explore how a news item's structure and wording influence its reliability. RUN-AS (Reliable and Unreliable News Annotation Scheme)[2] is a fine-grained annotation scheme based on two well-known journalistic techniques: the Inverted Pyramid and the 5W1H. To find out whether a news item presents objective information and follows journalistic standards, this proposal enables a three-level annotation: Structure labels (Inverted Pyramid), Content labels (5W1H) and Elements of Interest labels (EoI). Structure labels contain content and elements of interest labels within them. Content and EoI labels can be overlapped.

3.1.1 Structure labels

The Inverted Pyramid structure is one of the techniques used by journalists to reflect objectivity in a news item (Thomson, White, and Kitley, 2008). It consists of presenting the information in order of relevance, placing the most relevant information at the beginning and the least important at the end (DeAngelo and Yegiyan, 2019). The five structure labels of our proposal are TITLE, SUBTITLE, LEAD, BODY and CONCLUSION. Depending on the source, not all parts have to be present (such as the SUBTITLE or the CONCLUSION). However, the lack of essential parts of a news item (such as the TITLE, the LEAD or the BODY) strongly suggests that a news item is poorly structured. The definition of the structure labels is:

TITLE: headline of the news item. This label has two possible attributes. The attribute **title_stance** serves to indicate the relation and level of consistency between the TITLE and the BODY of a news item by means of the following values: Agree (information is consistent); Disagree (information is inconsistent); or, Unrelated (information has no relation). The attribute **style** is an attribute, which as with the title_stance is only used in the TITLE, but in this case marks the values Objective or Subjective of the information provided in the TITLE.

SUBTITLE: sentence completing the information of the TITLE.

LEAD: first paragraph presenting the essential information of the news item. It develops and usually repeats the idea presented in the title.

BODY: set of paragraphs developing the story and presenting in detail all the information of the news.

CONCLUSION: last sentence or paragraph summarising the content of the news article. It is not always present.

3.1.2 Content labels

The other technique used is the 5W1H which consists of answering six key questions. These questions describe the main event of a news story (Hamborg et al., 2018) and are usually found at the beginning of the news item, such as the TITLE or the LEAD. As stated by Chakma et al. (2020), "the 5W1H represents the semantic constituents of a sentence which are comparatively simpler to un-

[2] Available at `http://bit.ly/3T4XMzn`

derstand and identify". If a news item answers all these questions, it will mean that the information is communicated in a complete way and, therefore, the news item will have a higher degree of reliability than a news item that does not communicate the information in such a precise way. All the 5W1H elements are annotated as Reliable or Unreliable depending on their level of accuracy and objectivity (reliability attribute explained next).

WHAT: facts, circumstances, actions. Example: *los contagios de coronavirus se disparan* (coronavirus infections skyrocket).

WHO: subject, entity. Example: *la Agencia Europea del Medicamento* (European Medicines Agency).

WHEN: time, moment. Example: *el 20 de diciembre* (on 20 December).

WHERE: place, location. Example: *en España* (in Spain).

WHY: cause, reason. Example: *a causa de la muerte* (due to the death).

HOW: manner, method. Example: *con abundante agua* (with abundant water).

The 5W1H labels have the following attributes:

reliability is the main attribute of our annotation and allows to classify each element as well as the global news item with the values R (Reliable) or U (Unreliable), depending on the level of accuracy, objectivity, and the linguistic characteristics.

lack_of_information is used to indicate evidence is missing. This attribute has a single value (Yes). It is not indicate otherwise.

role is the attribute used with the WHO label only. It indicates the role played by the WHO entity in the event. It presents 3 values: Subject (if the entity causes the event), Target (if the entity receives the effects of the event) and Both (if the entity performs both functions).

main_event is only used with the WHAT label when the WHAT indicates the main event(s) of the story. It is possible to find several events (each one with its own 5W1H), but one is considered the main event.

3.1.3 Elements of Interest labels

The following Elements of Interest labels enable the annotation of textual information that could distinguish Unreliable from Reliable news:

QUOTE: label that marks the presence of quotes in the news item. It has the attribute **author_stance** that serves to annotate the author's stance regarding the QUOTE content. It has three values: Disagree (to express its disagreement towards the idea), Agree (to share its agreement) or Unknown (neutral stance). For example: *el experto niega que "el limón cura el cáncer"* (the expert denies that "lemon cures cancer") is a QUOTE with Disagree author_stance.

KEY_EXPRESSIONS: label containing phraseology that urges readers to share the information or that expresses emotions or economic purposes. For example: v*amos a salvar vidas compartiendo esta gran información* (let's save lives by sharing this important information)

FIGURE: numerical values in a news item.

ORTHOTYPOGRAPHY: label annotating poor writing and text with grammatical, spelling or formatting mistakes.

Figure 1 presents the specification of the three types of levels of the RUN-AS annotation scheme together with the attributes for each label, and the possible values for each attribute.

3.2 Reliability criteria

This work focuses on assigning a reliability value to the essential content labels described in our annotation scheme.

There are textual and linguistic features that enable the detection of the reliability of a news item and of each part of the news item, permitting an assessment of the news item's overall reliability. The criteria used when classifying the reliability consider accuracy and neutrality of the content relies on the state-of-the-art research presented in Section 2.3.

3.2.1 Accuracy

Accuracy is one of the key factors in determining the reliability of information. In our reliability modeling we have considered the following clues:

Vagueness and ambiguity. Evasive or vague expressions indicate that something is being concealed or that a fact cannot be justified, which makes the information provided Unreliable. For example, it is more reliable to give an exact date or precise details on a scientist (name, institution, degree) than to generalise or to provide inaccurate data. For example, a reliable WHEN is: *el*

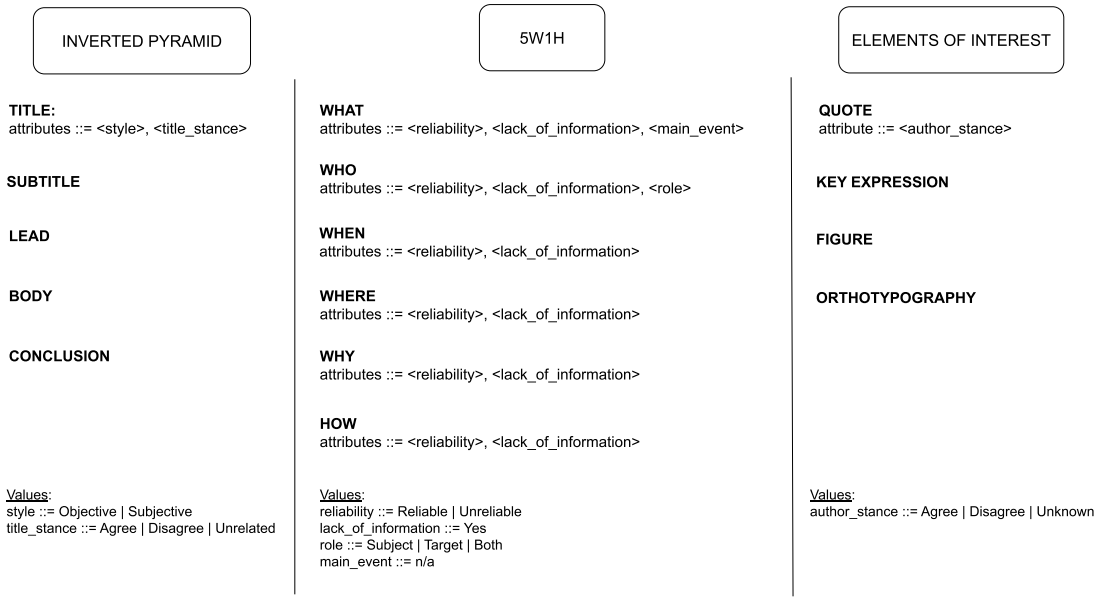


Figure 1: RUN-AS annotation scheme.

viernes 19 de marzo (on Friday 19 March) whereas *hace mucho tiempo* (a long time ago) lacks of accuracy. The existence of vagueness or ambiguity will be annotated with an **U**nreliable value associated with the corresponding 5W1H label.

On the contrary, the presence of figures, annotated with the **FIGURE** EoI label indicates accurate information that can be easily fact-checked with external sources, thus denoting reliability, for instance se han administrado {FIGURE: *6.000.000*} *dosis de vacunas* (6,000,000 doses of vaccine have been administered).

Lack of information. The presence of the this attribute can be considered a signal of unreliability. It appears with the 5W1H labels to mark the absence of important data in the text (such as the cause/reason of an event, the subject of the action, etc) as well as to indicate the lack of evidence such as scientific studies or official and verified data. Sometimes, the author states that the information is based on scientific studies without specifying which ones, which provides little credibility. As stated by (Mottola, 2020), the lack of data and sources is another typical characteristic of disinformation, turning news into stories that lack informative content. For example, a WHAT label with lack_of_information attribute is: *según algunos científicos* (according to some scientists).

Typos. When the **ORTHOTYPOGRAPHY** label is annotated, it has a negative reliability impact, as spelling mistakes, poor or careless writing style, inadequate punctuation or constant use of capital letters will not be considered a quality news item. Some examples of orthotypography are: whole sentences in capital letters; suspension points in the middle of the text or incomplete sentences; double spaces; many exclamation marks; grammatical errors; spelling mistakes; lack of cohesion; etc. For instance, *aquí en nuestro Pays* (here in our "Countri") is annotated with the ORTHOTYPOGRAPHY label.

3.2.2 Neutrality

In a news item, neutrality is a key component. A news item is more likely to be Reliable when information is provided in an objective manner and does not show the author's stance. Hints about text neutrality (or lack thereof) are considered in the RUN-AS schema as follows:

Personal Remarks and Emotional Messages. When the author speaks in the first person, tells his/her personal experience or that of someone he/she knows, it is a sign of low credibility, as the author is trying to scare, persuade or make the reader feel closer to the story and thus empathise (Rashkin et al., 2017). Futhermore, offensive, hopeful, alarming or exhortative messages are a clear sign of unreliability because the author is try-

ing to manipulate the reader and to play with people's emotions (Zhang et al., 2018).

Through the labeling of **KEY_EXPRESSIONS** we can represent this kind of non-neutral information.

Some examples of KEY_EXPRESSIONS regarding this issue are: *yo lo hago y funciona* (I do it and it works) or *evite que sus amigos y conocidos se enfermen* (keep your friends and acquaintances from getting sick).

Quotes and author stance. The presence of **QUOTE** labels add neutrality to a news item since it indicates that the information comes from an external source (Zhang et al., 2018). However, when the author is clearly in favor or against the quote, an important hint of subjectivity is introduced. Thus, labeling QUOTE with attribute **author_stance=Unknown** would indicate neutrality since the author will only be reproducing the words of a third party to inform and not to influence the reader, while any other value would indicate a lack of it.

Title style and stance. The titles of newspaper articles often provide important clues to the reliability of the content. For example, alarmist, subjective or striking titles are suspected of introducing unreliable information. Also, misleading or opaque titles on a topic may indicate clickbait (Zhang et al., 2018). Even certain morphosyntactic features such as the excessive length of a title, the use of more capitalised words (Horne and Adali, 2017) and punctuation marks (especially exclamation marks) and ellipses can lead to a lack of neutrality (Mottola, 2020). In our annotation proposal, these clues are marked in the **TITLE** by means of the attribute-value **style=Subjective**.

Moreover, the stance of the title regarding the news content indicates misleading information when they disagree (Ferreira and Vlachos, 2016). In this case, the existence of the attribute value **style=Disagree** associated with the TITLE label would clearly indicate that the information is Unreliable.

4 Annotation environment and RUN Dataset

A Reliable and Unreliable News (RUN) dataset in Spanish and focused on health and COVID-19 has been created to test the RUN-AS proposal. The RUN dataset comprises 80 Reliable and Unreliable news items, randomly selected and then sorted (36,659 words in total), of which 51 are Reliable and 29 Unreliable, collected from several digital newspapers. Both the reliability of the internal elements and the global reliability of the news item are annotated. News has been annotated with Brat, an intuitive web-based annotation tool (Stenetorp et al., 2012). An example of the graphical annotation in Brat can be observed in Figure 2[3].

Tables 1 and 2 show the total number of labels in the dataset[2].

Label	% Reliable	% Unreliable	Total
WHAT	74.64	25.09	1100
WHO	84.49	15.37	748
WHEN	78.93	21.07	299
WHERE	94.61	4.79	334
WHY	69.08	30.92	152
HOW	75.74	23.76	202

Table 1: Dataset description (5W1H labels).

Structure and EoI labels	% Appearance
TITLE	100
SUBTITLE	55
LEAD	95
BODY	100
CONCLUSION	62.50
QUOTE	53.75
KEY_EXPRESSION	32.50
FIGURE	63.75
ORTHOTYPOGRAPHY	40

Table 2: Dataset description (Structure and EoI labels).

The methodology for creating the dataset followed five steps. First, the dataset was defined and delimited on the basis of three main criteria: domain (health and COVID-19), language (Spanish) and traditional news content structure. Second, news was collected both manually and by means of a web crawler. Third, RUN-AS annotation scheme was applied, and a reliability rating was assigned for each 5W1H label. Fourth, the global reliability of each news item was assigned by two non-expert annotators with knowledge of NLP, taking into account only the plain text, without the labels of the expert annotator. Finally, two inter-annotator agreements were measured to validate the quality of the annotation.

[3] `https://bit.ly/38AyW7K`

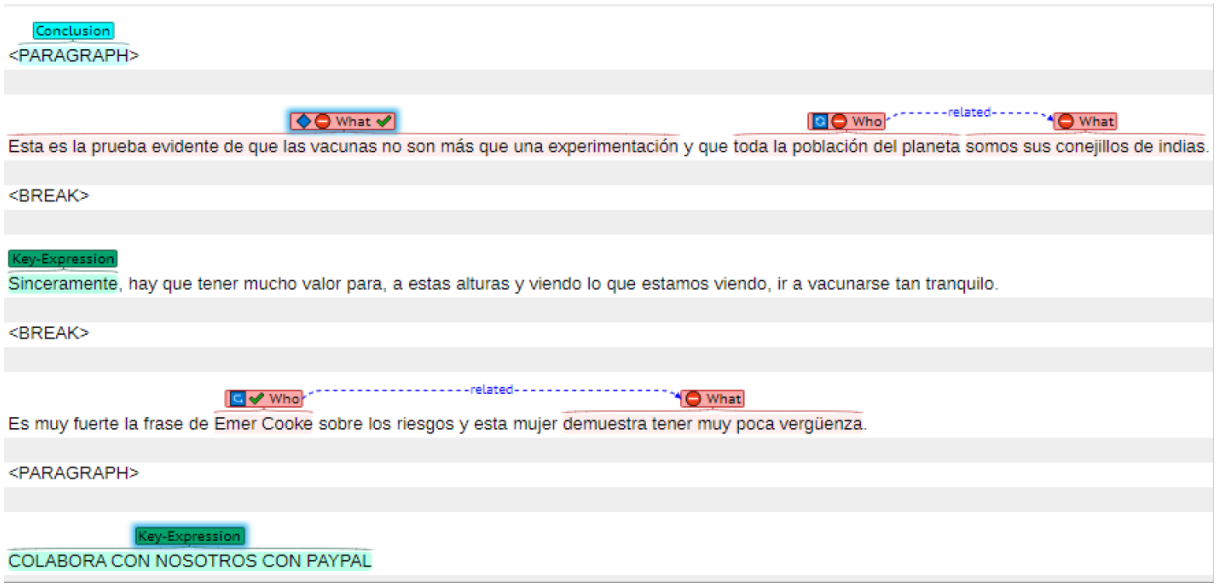


Figure 2: Annotation of 5W1H, Inverted Pyramid and Elements of Interest on Brat.

4.1 Annotation quality

Two inter-annotator agreements were calculated independently using the Cohen's Kappa metric (Vieira, Kaymak, and Sousa, 2010). Firstly, the inter-annotator agreement regarding the three levels of RUN-AS annotation (Structure, 5W1H and EoI) was performed. Secondly, the inter-annotator agreement regarding the annotation of the global reliability of the news item was obtained.

4.1.1 Labels inter-annotator agreement

To measure the agreement in the annotation of the three level labels, the annotation of a set of news items comprising 1,337 words was asked to two non-expert annotators. Without previous training, they had to annotate news according to the annotation scheme proposed. The agreement obtained a score of k=0.80 in the Inverted Pyramid and of k=0.53 in the 5W1H. This inter-annotator agreement allowed us to reach the conclusion that annotating semantic elements has a higher level of difficulty and therefore more intense training needs to be provided to annotators for this purpose.

4.1.2 Global reliability inter-annotator agreement

In order to measure the agreement when annotating global reliability of a news item, two non-expert annotators were used. Their annotations had to be made using plain text only, without labels, and following the reliability criteria defined in the scheme. The agreement obtained in this task was k=0.75 which is considered a fairly high score. When there was no agreement among the annotators, a consensus process was carried out.

5 Validation of RUN-AS scheme: Evaluation framework

Several experiments were conducted to validate our RUN-AS scheme and to support the hypothesis that a fine-grained reliability assessment of the elements in a news story can provide an accurate estimation of its global reliability.

SOTA Machine Learning (ML) and Deep Learning (DL) methods, widely applied in the disinformation classification task, were used to determine whether the information provided by the proposed annotation scheme is feasible to address disinformation detection. From this fine-grained annotation proposal (Structure, Content, and EoI) two types of features were extracted: numerical and categorical. In total, 42 different features were extracted per news item.

From the Structure level, a total of 7 features were extracted as follows: 5 categorical features that indicate the presence of the news structure parts (TITLE, SUBTITLE, LEAD, BODY and CONCLUSION); and, 2 other categorical features extracted from the attributes of the TITLE (stance and style). Concerning the 5W1H content and EoI levels, there is a total of 35 numerical features that refer to the number of labels for each one.

As for the 5W1H content level, 6 features were extracted related to each 5W1H. For each 5W1H label, the number of attributes of type Reliable/Unreliable was counted (12 features), as well as the number of the attributes of type lack_of_information (6 features), the attribute of type role (3 features), and the attribute of type main_event (1 feature). Regarding the level of Elements of Interest, a total of 4 numerical features were extracted (FIGURE, KEY_EXPRESSION, ORTHOTYPOGRAPHY and QUOTE), as well as the number of attributes of type author_stance (3 features). A simplified example of some numerical and categorical features extracted from the TITLE and LEAD of a news piece is presented next.

```
{
    TITLE_style: Objective,
    TITLE_title_stance: Agree,
    TITLE_WHAT_Reliable: 0,
    TITLE_WHAT_Unreliable: 1,
    TITLE_WHO_Reliable: 0,
    TITLE_WHO_Unreliable: 1,
    TITLE_WHEN_Reliable: 0,
    TITLE_WHEN_Unreliable: 1,
    LEAD_WHAT_Reliable: 2,
    LEAD_WHAT_Unreliable: 2,
    LEAD_WHO_Reliable: 0,
    LEAD_WHO_Unreliable: 1,
    LEAD_WHEN_Reliable: 0,
    LEAD_WHEN_Unreliable: 3,
    # ...
}
```

The same type of features will be generated from the other parts of the structure of the document. Each feature indicates the number of 5W1H components with a specific label and reliability attribute that appear in each part of the news. For example, `LEAD_WHAT_Reliable: 2` indicates that the `LEAD` contains two `WHAT` items annotated with a `Reliable` value. The model is trained to predict the overall document reliability label based on these numerical and categorical features.

5.1 Experiments

To confirm the suitability of the RUN-AS proposal, we decided to test classic ML algorithms that obtained good results using numerical and categorical features. In addition, a DL language model, which obtained state-of-the-art results in many tasks within NLP, was used to compare the results. The following experiments were carried out:

ML performance: the following ML classification algorithms are used: Support Vector Machines (SVM); Random Forest (RF); Logistic Regression (LR); Decision Tree (DT); Multi-layer Perceptron (MLP); Adaptive Boosting (AdaBoost); and, Gaussian Naive Bayes (GaussianNB). Two configurations of the aforementioned algorithms are used.

- *Baseline model:* encoding of news texts by using TF-IDF type vectors.
- *Model with RUN-AS features*: concatenation of the TF-IDF vectors with the 42 features obtained from the annotation.

This experiment was implemented using *scikit-learn* library[4]. It can be replicated at the Colab[5] notebook.

DL performance (pre-trained transformer model): the Beto[6] language model based on transformer architecture (Canete et al., 2020) was used to create two classifier models. Both classifier models consist of fine-tuning the model by using the annotated dataset and are composed of two main components: a language model (BETO) and a classification neural network. The architecture of classification presented in Sepúlveda-Torres et al. (2021) is used. The following hyperparameters were used: maximum sequence length of 512, batch size of 2, training rate of 2e-5, and training performed for 3 epochs.

- *Baseline model:* the first is a baseline system that used the news as input to the language model (BETO).
- *Model with RUN-AS features*: the second used the architecture proposed by Sepúlveda-Torres et al. (2021), which modified the BETO baselines to include external features. Both the text and the 42 features were used as input. Features are concatenated with the output of the BETO language model to feed the input to the classification neural network.

To create the classifiers, the *Simple Transformers library*[7] was used, which creates a wrapper around *HuggingFace's Transformers library* for using Transformer models (Wolf et al., 2019). These experiments can be reproduced on the repository[8]. The cross-

[4] `https://scikit-learn.org/stable/`
[5] `https://bit.ly/37KNHnM`
[6] `https://github.com/dccuchile/beto`
[7] `https://simpletransformers.ai/`
[8] `https://bit.ly/3L5LvJg`

validation strategy was performed in all experiments enabling all available data to be used for training and testing (Bergmeir and Benítez, 2012). In these experiments, k-fold cross-validation with k = 5 is used, where 80% of each subset has been used for training and 20% for testing. In order to evaluate the proposal, the commonly used NLP measures (accuracy and macro-averaged F_1 –F_1m–) are used.

6 Validation of RUN-AS scheme: Results and Discussion

This section presents the results obtained in each of the experiments and a discussion of those results. Table 3 presents the performance of experiments explained in Section 5. All the models that used RUN-AS features significantly outperform the proposed baselines. The best results are attained with Decision Tree using RUN-AS annotation, obtaining a 0.948 of macro F_1 (F_1m), and BETO using RUN-AS annotation, obtaining a 0.854 of F_1m. It is noteworthy that when using the whole document annotated with a single reliability value (baselines) the best F_1m value is obtained by AdaBoost with 0.748 F_1m, followed by Random Forest and Decision Tree. However, for the rest of the approaches, the results using the document with a single reliability value are very poor. All approaches are significantly improved by using the information provided by the annotation labels of the RUN-AS scheme. Therefore, these results validate the main hypothesis presented in this research, i.e., that individual 5W1H components reliability are a better predictor of overall news story reliability.

7 Conclusions and future work

The novelty of this work lies in the development of RUN-AS, a fine-grained annotation scheme based on journalistic techniques that classify news and its essential parts into Reliable or Unreliable. This annotation proposal was tested by using ML and DL experiments in a Spanish news dataset called RUN, created ad hoc. Furthermore, inter-annotator agreements were measured, both those related to the three-level RUN-AS label annotation as well as those related to the global reliability of the news item. The results indicate the intrinsic complexity derived from a semantically rich annotation scheme. Experiments conducted have shown that the individual reliability of each of the elements annotated contributes to assessing the overall reliability of a news item with a 0.948 F_1m performance. Therefore, the experiments presented here support the hypothesis that a fine-grained reliability assessment of multiple semantic elements in a news story can provide an accurate estimate of a global reliability score.

This annotation is complementary to other lines of research, such as fact-checking or contradiction detection, as it provides useful information at a first level of a text-only annotation. Our proposal is designed to annotate the style, the structure of the story, the tone, the evidence, the neutrality or the way in which information is provided. These are key characteristics that distinguish Reliable from Unreliable news. As future work, we are developing an assisted annotation methodology that combines both manual and automatic approaches. This semi-automatic system will reduce the time and the effort spent on compilation and annotation tasks, enabling a RUN dataset extension. Furthermore, performance in the veracity detection task of the RUN-AS annotated dataset will be evaluated to determine to what extent reliability detection can support veracity detection.

Acknowledgments

This research work is funded by MCIN/AEI/10.13039/501100011033 and European Union NextGenerationEU/PRTR through the projects “TRIVIAL” (PID2021-122263OB-C22) and “SocialTrust” (PDC2022-133146-C22). It is also supported by Generalitat Valenciana through the project “NL4DISMIS” (CIPROM/2021/21) and Consellería de Innovación, Universidades, Ciencia y Sociedad Digital (ACIF/2020/177).

References

Assaf, R. and M. Saheb. 2021. Dataset for arabic fake news. In *2021 IEEE 15th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–4. IEEE.

Bergmeir, C. and J. M. Benítez. 2012. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, may.

Experiments	**Baseline model (TF-IDF)**		**Model with RUN-AS features**	
	Acc	F_1m	**Acc**	F_1m
SVM	0.662	0.395	0.937	0.925
Random Forest	0.75	0.639	0.912	0.898
Logistic Regression	0.650	0.392	0.912	0.875
Decision Tree	0.737	0.683	**0.950**	**0.948**
MLP	0.712	0.570	0.925	0.912
AdaBoost	**0.787**	**0.748**	**0.950**	0.945
GaussianNB	0.612	0.456	0.687	0.570
	Baseline model		**Model with RUN-AS features**	
BETO	**0.850**	**0.800**	**0.887**	**0.854**

Table 3: Experiments results using ML and DL methods.

Canete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:2020.

Chakma, K. and A. Das. 2018. A 5w1h based annotation scheme for semantic role labeling of english tweets. *Computación y Sistemas*, 22(3):747–755.

Chakma, K., S. D. Swamy, A. Das, and S. Debbarma. 2020. 5w1h-based semantic segmentation of tweets for event detection using bert. In *International Conference on Machine Learning, Image Processing, Network Security and Data Sciences*, pages 57–72. Springer.

DeAngelo, T. I. and N. S. Yegiyan. 2019. Looking for efficiency: How online news structure and emotional tone influence processing time and memory. *Journalism & Mass Communication Quarterly*, 96(2):385–405.

Ferreira, W. and A. Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, California, June. Association for Computational Linguistics.

Figueira, Á. and L. Oliveira. 2017. The current state of fake news: challenges and opportunities. *Procedia Computer Science*, 121:817–825.

Giansiracusa, N. 2021. *How Algorithms Create and Prevent Fake News*. Springer.

Gruppi, M., B. D. Horne, and S. Adali. 2018. An exploration of unreliable news classification in brazil and the us. *arXiv preprint arXiv:1806.02875*.

Hamborg, F., C. Breitinger, M. Schubotz, S. Lachnit, and B. Gipp. 2018. Extraction of main event descriptors from news articles by answering the journalistic five w and one h questions. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, pages 339–340.

Horne, B. and S. Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 759–766.

Khodra, M. L. 2015. Event extraction on indonesian news article using multiclass categorization. In *2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–5. IEEE.

Mottola, S. 2020. Las fake news como fenómeno social. análisis lingüístico y poder persuasivo de bulos en italiano y español. *Discurso & Sociedad*, (3):683–706.

Norambuena, B., M. Horning, and T. Mitra. 2020. Evaluating the inverted pyramid structure through automatic 5w1h extraction and summarization. In *Computational Journalism Symposium*.

Paka, W. S., R. Bansal, A. Kaushik, S. Sengupta, and T. Chakraborty. 2021. Crossstitch semi-supervised neural attention model for covid-19 fake news detection. *Applied Soft Computing*, 107:107393.

Patwa, P., S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty. 2021. Fighting an infodemic: Covid-19 fake news dataset. In *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 21–29. Springer.

Pérez-Rosas, V., B. Kleinberg, A. Lefevre, and R. Mihalcea. 2017. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*.

Posadas-Durán, J.-P., H. Gómez-Adorno, G. Sidorov, and J. J. M. Escobar. 2019. Detection of fake news in a new corpus for the spanish language. *Journal of Intelligent & Fuzzy Systems*, 36(5):4869–4876.

Rashkin, H., E. Choi, J. Y. Jang, S. Volkova, and Y. Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.

Saquete, E., D. Tomás, P. Moreda, P. Martínez-Barco, and M. Palomar. 2020. Fighting post-truth using natural language processing: A review and open challenges. *Expert systems with applications*, 141:112943.

Sepúlveda-Torres, R., E. Saquete Boró, et al. 2021. Gplsi team at checkthat! 2021: Fine-tuning beto and roberta. *CEUR*.

Shahi, G. K., J. M. Struß, and T. Mandl. 2021. Overview of the clef-2021 checkthat! lab task 3 on fake news detection. *Working Notes of CLEF*.

Shao, C., G. L. Ciampaglia, O. Varol, A. Flammini, and F. Menczer. 2017. The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592*, 96:104.

Shu, K., S. Wang, D. Lee, and H. Liu. 2020. Mining disinformation and fake news: Concepts, methods, and recent advancements. In *Disinformation, Misinformation, and Fake News in Social Media*. Springer, pages 1–19.

Silva, R. M., R. L. Santos, T. A. Almeida, and T. A. Pardo. 2020. Towards automatically filtering fake news in portuguese. *Expert Systems with Applications*, 146:113199.

Stenetorp, P., S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Thomson, E. A., P. R. White, and P. Kitley. 2008. "objectivity" and "hard news" reporting across cultures: Comparing the news report in english, french, japanese and indonesian journalism. *Journalism studies*, 9(2):212–228.

Vieira, S. M., U. Kaymak, and J. M. Sousa. 2010. Cohen's kappa coefficient as a performance measure for feature selection. In *International Conference on Fuzzy Systems*, pages 1–8. IEEE.

Vlachos, A. and S. Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22.

Vosoughi, S., D. Roy, and S. Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Wang, W. Y. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zhang, A. X., A. Ranganathan, S. E. Metz, S. Appling, C. M. Sehat, N. Gilmore, N. B. Adams, E. Vincent, J. Lee, M. Robbins, et al. 2018. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of the The Web Conference 2018*, pages 603–612.

Zhou, X. and R. Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 27-38 recibido 29-11-2022 revisado 31-01-2023 aceptado 10-02-2023

Evaluation of transformer-based models for punctuation and capitalization restoration in Catalan and Galician

Evaluación de modelos basados en Transformers para el sistema de recuperación de puntuación y mayúsculas en Catalán y Gallego

Ronghao Pan[1], **José Antonio García-Díaz**[1],
Pedro José Vivancos-Vicente[2], **Rafael Valencia-García**[1]
[1]Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, España
[2]VÓCALI Sistemas Inteligentes S.L., Parque Científico de Murcia,
Carretera de Madrid km 388. Complejo de Espinardo, 30100 Murcia, España
{ronghao.pan, joseantonio.garcia8, valencia}@um.es
pedro.vivancos@vocali.net

Abstract: In recent years, the performance of Automatic Speech Recognition systems (ASR) has increased considerably due to new deep learning methods. However, the raw output of an ASR system consists of a sequence of words without capital letters and punctuation marks. Therefore, a capitalization and punctuation restoration system are one of the most important post-processes of ASR to improve readability and to enable the subsequent use of these results in other NLP models. Most models focus solely on English punctuation resolution, and recently new models of Spanish punctuation restoration have emerged. However, none focus on capitalization and punctuation restoration in Galician and Catalan. In this sense, we propose a system for capitalization and punctuation restoration based on Transformers models for Catalan and Galician. Both models perform very well, with an overall performance of 90.2% for Galician and 90.86% for Catalan, and have the ability to identify proper names, country names, and organizations for uppercase restoration.
Keywords: Automatic Speech Recognition, Transformers, Punctuation Restoration, Capitalization Restoration, Catalan, Galician.

Resumen: En los últimos años, el rendimiento de sistemas de Reconocimiento Automático del habla ha aumentado considerablemente gracias a nuevos métodos de deep learning. Sin embargo, la salida bruta de estos sistemas consiste en secuencias de palabras sin mayúsculas ni signos de puntuación. Recuperar esta información mejora la legibilidad y permite su posterior uso en otros modelos de PLN. La mayoría de las soluciones existentes se centran únicamente en inglés; aunque recientemente han surgido nuevos modelos de restauración de la puntuación en español. Sin embargo, ninguno se centra en gallego y catalán. En este sentido, proponemos un sistema de restauración de mayúsculas y puntuación basado en modelos Transformers para estos idiomas. Ambos modelos tienen un rendimiento muy bueno: 90,2% para el gallego y 90,86% para el catalán. Además, también tienen la capacidad de identificar nombres propios, nombres de países y organizaciones para la restauración de mayúsculas.
Palabras clave: Reconocimiento Automático del Habla, Transformers, Recuperación de puntuación, Recuperación de mayúsculas, Catalán, Gallego.

ISSN 1135-5948 DOI 10.26342/2023-70-2

1 Introduction

In recent years, the performance of Automatic Speech Recognition (ASR) systems has increased significantly due to recent advances in deep learning methods. The improved performance of ASR has enabled the development of a wide range of applications in various fields, such as voice assistants, customer care, and healthcare, making it increasingly important in our daily lives. However, the ASR system often generates a stream of unpunctuated words as output, which noticeably reduces its overall readability and comprehensibility (Jones et al., 2003). Moreover, the most advanced Natural Language Processing (NLP) models are mostly trained with punctuated text, such as Wikipedia texts (Cañete et al., 2020). Thus, unpunctuated texts reduce the possibility of being used in these models (Peitz et al., 2011), because the lack of punctuation would seriously degrade the performance of the language models. For example, in Basili et al. (2015) there is a performance difference of over 10% when the models are trained with newspaper texts and tested with unpunctuated transcripts for the entity recognition system.

Recent developments in transformer-based pre-trained models have proven to be successful in many NLP tasks across different languages, and these models have been explored very litter for the punctuation restoration problem. In this work, we present a model of punctuation and capitalization restoration for Catalan and Galician. Both models are composed of a transformer architecture that uses an adapted pre-trained language model as a starting point for transferring knowledge to a specific task, as in this case, the identification of capital letters and punctuation marks. Currently, for Galician and Catalan there are different monolinguals and multilingual models based on BERT or RoBERTa, with different performances. Thus, this work also analyses the behavior of different pre-trained models for the task of automatic restoration of punctuation and capital letter.

This paper is structured as follows: Section 2 presents an overview of the state of the art of punctuation and capitalization restoration system. In Section 3, materials and methods are presented and described in detail. Section 4 presents the performed experiment and the results obtained by different pre-trained language models. In Section 5, error analysis is conducted with a few representative examples. Finally, in Section 6 the conclusions and future work are discussed.

2 Related work

Nowadays, the task of automatically recovering capitalization and punctuation marks has been extensively studied in many systems. These approaches can be broadly divided into three categories in terms of applied features (Yi et al., 2020): those using prosody features derived from acoustic information, those using lexical features, and the combination of the previous two features-based methods.

In recent years, the problem of punctuation retrieval has been addressed with different approaches, from the use of deep learning algorithms, such as Che et al. (2016), which used pre-trained word embedding to train feedforward deep neural network and Convolutional Neural Network, to architectures based on Recurrent Neural Networks (RNNs) combined with Conditional Random Fields (CRF) and pre-trained vectors (Tilk and Alumäe, 2016). Tilk and Alumäe (2016) used RNNs with an attention mechanism to improve performance over Deep Neural Networks and CNN models. Recent advances in transformer-based pre-trained models have proven to be successful in many NLP tasks, so new transformer-based approaches based on BERT-type architectures have emerged (Courtland, Faulkner, and McElvain, 2020), which have been shown to achieve values of up to 83.9% on the F1-score in the well-known and reference IWSLT 2012 dataset (Federico et al., 2012). Another study (Alam, Khan, and Alam, 2020) explored different transformer-based models for both English and Bangla using different pre-trained models and used bidirectional LSTM (BiLSTM) on top of the pre-trained transformer network. However, most of these models mainly focus on solving the problem of punctuation in the three most common punctuation marks, such as period (.), comma (,), and question (?) in English.

Recently, new models of punctuation restoration in Spanish have emerged, such as punctuation restoration in Spanish customer support transcripts using transfer learning (Zhu et al., 2022), and a BERT-based automatic punctuation and capitalization system for Spanish and Basque (González-Docasal et

al., 2021), but there is no adapted model for Catalan and Galician.

The system presented in this paper addresses 5 different punctuation marks for Catalan and Galician, which are described in Section 3. Both models are composed of a transformer architecture but, unlike prior works that solely studied one architecture (BERT), we experiment with different pre-trained models based on BERT, and RoBERTa (see Section 3.3), thus analyzing the monolingual and multilingual models used. We also propose an augmentation scheme that improves performance. Our augmentation is closed related to the augmentation techniques proposed in (Alam, Khan, and Alam, 2020) where authors consider *unknow* word substitution, random insertion, and random deletion. We propose a different version of it in our approach, which uses the back-translation technique for the substitution task described in Section 3.2.

3 Materials and methods

We frame the restoration of punctuation and capitalization as a sequence token classification problem in which the model predicts the punctuation marks that each word in the input text may have. The main advantage of using this approach is that no dependency information is lost and that, given a word in a sentence or sequence, the model can use which word is on the right or left to predict its punctuation mark.

Instead of covering all possible punctuation marks in Catalan and Galician, we only include 5 types of target punctuation that are commonly used and are important to improve the readability of the transcription: period (.), comma (,), question (?), exclamation (!), and colon (:). More specifically, the model predicts which punctuation mark appears next to a given token. However, our model also has the ability to restore capital letters, so for each type of punctuation two labels are added, e.g. for the comma, we have two labels: '`,u`' indicates that the token is of uppercase type and has the comma, and '`,l`' denotes that the token is of lowercase type. Therefore, there are a total of 12 classes that the model needs to predict: '`l`' (lower case), '`u`' (upper case), '`?u`' (upper case with a question), '`?l`' (lower case with a question), '`!u`' (upper case with an exclamation), '`!l`' (lower case with an exclamation), '`,u`' (upper case with a comma), '`,l`' (lower case with a comma), '`.u`' (upper case with a period), '`.l`' (lower case with a period), '`:u`' (upper case with a colon) and '`:l`' (lower case with a colon).

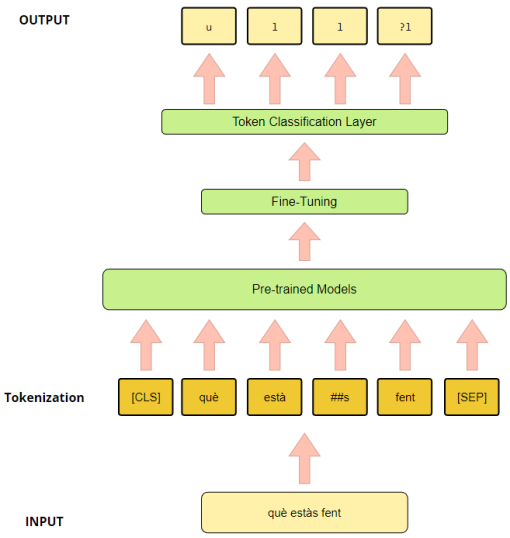


Figure 1: Capitalization and punctuation restoration model structure.

In Figure 1, we report the punctuation restoration model structure. Briefly, it can be described as follows. First, we pre-process the input data through by the tokenization process (see Section 3.4). Second, we use several pre-trained transformer-based models as a starting point, as this reduces computational costs and allows us to use the latest generation models without having to train one from scratch (see Section 3.3). Third, we train the pre-trained models to fit a token classification task, transferring the knowledge from the pre-trained model. This stage is also known as fine-tuning. Finally, the best models from each pre-trained model are evaluated on the test dataset. As can be seen in Figure 1, the input sentence "què estàs fent" does not have any punctuation, and the model predicts that the word "què" is uppercase type, the word "estàs" is lowercase type, and the word "fent" is lowercase and has a question mark after it to produce the output sentence "Què estàs fent?".

We split the dataset into three parts (as shown in Table 2) to evaluate model accuracy: the training set (60%), the validation set (20%), and the test set (20%). The performance of each model is evaluated on the test set after it has been finetuned on vari-

ous combinations of sources and training processes.

3.1 Dataset

We use OpusParaCrawl (Bañón et al., 2020) dataset for Catalan and Galician capitalization and punctuation restoration, which consists of parallel corpora from Web Crawls collected in the ParaCrawl project. These datasets contain 42 languages, and 43 bitexts with a total number of 3.13G sentence fragments by crawling hundreds of thousands of websites, using open-source tools. Usually, parallel corpora are essential for building high-quality machine translation systems and have found uses in many other natural language applications, such as paraphrases learning (Bannard and Callison-Burch, 2005). In this case, the main reason for using this database for capitalization and punctuation restoration is that the texts are already divided into sentences and have all the punctuation marks for each language. After the cleaning, extraction and selection process, a total of 50,000 sentences are selected for Catalan and Galician with 1,136,708 and 1,062,565 words respectively. During the selection process, less common punctuation marks, such as question marks, exclamation marks, and colons, have been preferentially selected to balance the dataset.

3.2 Data augmentation

For this study, we propose an augmentation method inspired by the study of Alam, Khan, and Alam (2020), as discussed above. By training the models with well-trained and correctly punctuated datasets, the trained models lack the knowledge of the typical errors made by an ASR system. Therefore, our augmentation method relies on the type of errors made by the ASR during recognition using the random insertion and deletion technique, and back-translation of sentences to augment the training set.

Currently, most synonym substitution models focus mainly on the English language, so we have chosen the back-translation technique for our synonym substitution task. This technique consists of first translating the text into a given language and then back-translating it into the source language. Thus, when translating the text from one language to another, the translation models usually replace some words with their synonyms or create a new sentence with the same meaning. In this study, the "Helsinki-NLP" models have been used to translate an original text in Catalan or Galician into English and then back-translate it into the source language. In contrast to Alam, Khan, and Alam (2020), we consider all three techniques to have the same prevalence. With this in mind, to process the input text with augmentation, we use three adjustable parameters with the same value (0.33) to control the probability of each of them. Table 1 shows an example of each technique, where Text 1 corresponds to the back-translation technique, Text 2 to the random insertion, and Text 3 to the random detection technique.

3.3 Pre-trained models

Transfer learning and pre-trained transformer-based models have been popular in computer vision and widely adopted for various NLP tasks since the introduction of BERT (Devlin et al., 2019). For Catalan and Galician, available pre-trained resources include multilingual models such as mBERT (Devlin et al., 2019), DistilmBERT (Sanh et al., 2019), and XLM-RoBERTa (Conneau et al., 2019), as well as monolingual such as BERTa for Catalan (Armengol-Estapé et al., 2021), and Bertinho for Galician (David Vilares, 2021). In our experiment, we used such pre-trained language models for capitalization and punctuation restoration tasks. Moreover, we briefly discuss the monolingual language of Catalan and Galician, and the multilingual models used in this study. The following models are used:

- **Bertinho**: It is a robust monolingual model based on the BERT for Galician. It has two versions created with 6 and 12 transformer layers, respectively, and trained with a limited amount of resources (around 45 million words on a single 24GB GPU) (David Vilares, 2021). For our experiment, we have used the 12 transformers layers version.
- **BERTa**: It is a transformer-based masked language model based on the RoBERTa for the Catalan language. It has been trained on a medium size corpus collected from web crawling and public corpora (Armengol-Estapé et al., 2021). The training corpus consists of several corpora gathered from web

	Text	Data augmentation
1	Els nens d' aquesta edat no estan desenvolupats fisicament per carregar gaire pes, i per tant les motxilles son petites i lleugeres.	Els nens d' aquesta edat no es desenvolupen principalment per carregar massa pesats, i per tant les bosses són petites i lleugers.
2	Sessio de formacio per a pares i mares d' adolescents centrada en la promocio d' habits i estils de vida saludables.	Sessio de en per a pares i mares d' adolescents centrada formacio la promocio d' habits i estils de vida saludables.
3	En algun moment despres d' aixo, va atracar l' Illa Shimotsuki per aliments i subministraments, estant a prop d' en Zoro.	d' aixo, atracar l' Illa Shimotsuki aliments subministraments, a prop Zoro.

Table 1: Examples of data augmentation.

Dataset	Total	l	u	?u	?l	!u	!l	,u	,l	.u	.l	:u	:l
Galician													
Train	682,444	522,508	75,880	880	5,501	955	5,027	10,492	32,979	3,732	15,806	2,064	6,620
Train (Augmented)	920,017	704,358	98,404	1,172	7,477	1,250	6,794	15,719	47,814	5,233	14,713	2,728	9,686
Dev	166,551	127,253	18,655	196	1,326	250	1,246	2,428	8,068	943	3,993	583	1,610
Test	213,570	163,697	23,422	260	1,609	286	1,576	3,322	10,449	1,131	4,967	679	2,172
Catalan													
Train	726,486	569,875	73,357	800	5,481	970	5,149	9,709	33,711	3,450	15,710	2,227	6,047
Train (Augmented)	981,345	768,552	96,442	1,080	7,385	1,293	6,966	13,532	48,161	4,827	21,771	2,993	7,385
Dev	181,517	142,034	18,507	231	1,398	271	1,321	2,377	8,284	853	4,178	517	1,546
Test	228,705	179,531	22,794	255	1,717	310	1,621	3,071	10,771	1,102	4,954	650	1,929

Table 2: Distribution of the datasets.

crawling and public corpora: (1) the Catalan part of the DOGC corpus (Tiedemann, 2012), (2) a collection of translated Catalan movie subtitles, (3) the non-shuffled version of the Catalan part of the OSCAR corpus, (4) a web corpus of Catalan called CaWac (Ljubešić and Toral, 2014), and the Catalan Wikipedia articles.

- **RoBERTinha**: It is RoBERTa-like language model trained on Oscar Galician corpus, and based on the approach presented by Ortiz Suárez, Romary, and Sagot (2020).
- **mBERT**: It is a transformer model pretrained on a large multilingual data corpus of about 104 languages with the largest Wikipedia using Masked Language Modeling (MLM) target (Devlin et al., 2019).
- **DistilmBERT**: This model is a distilled version of BERT base multilingual model (mBERT) (Devlin et al., 2019). It has been trained on the concatenation of Wikipedia in 104 languages listed. The model has 6 layers, 768 dimensions and 112 heads, totalizing 134 parameters (Sanh et al., 2019).
- **XLM-RoBERTa**: This model was proposed in Conneau et al. (2019). It is a multilingual version of RoBERTa trained by Facebook AI Research (FAIR). It has been trained on 2.5 TB of filtered CommonCrawl data containing 100 languages and has demonstrated superior performance on task such as text classification and multi-language text generation compared to other existing language models.

3.4 Tokenization

The main feature of transformer networks is their self-attention mechanism, whereby each word in the input can learn what relation it has with the others (Yi and Tao, 2019). As shown in Figure 1, all models are based on different transformers-based models, such as BERT, RoBERTa or XLM-RoBERTa, and all of them need the input data to be preprocessed by the tokenization process, which consists of decomposing a larger entity into smaller components called *tokens*. For tokenization, we use model-specific tokenizers, and Figure 2 shows some examples of each of them. The models used in this study use the tokenization of sub-words with the WordPiece algorithm as BERT or the Byte-Pair Encoding (BPE) algorithm in the RoBERTa and XLM-RoBERTa-based models, so there are words that split into several tokens as in the case of BERT frequent tokens are grouped into one token and less frequent to-

kens are split into frequent tokens (Bostrom and Durrett, 2020). The main differences in the tokenizers used are as follows:

- In BERT, it uses special tokens such as [CLS] and [SEP] to indicate the beginning and end of a sentence.
- In both RoBERTa and XLM-RoBERTa, the first word of the sentence is not prefixed with any special characters and uses $< s >$ and $< /s >$ to indicate the beginning and end of a sentence.
- In RoBERTa, all tokens in the sentence are prefixed with "Ġ", and when a word is split into several sub-words, the first sub-word is prefixed with "Ġ" and the remaining sub-words are not prefixed with any special characters.
- In XLM-RoBERTa all tokens in the phrase are prefixed with "▁", and when a word is split into various sub-words, the first sub-words are prefixed with "▁" and the rest of the sub-words are not prefixed with any special character.

Therefore, it is necessary to adjust the subword labels and treat special tokens so that they are ignored during training. For this purpose, we have applied the following techniques:

- Assign -100 labels to special tokens such as [CLS], [SEP], $< s >$ and $< /s >$ so that they are ignored during training.
- Assign all sub-words the same label as the first sub-word to solve the sub-word tokenization problem.

4 Results and analysis

We evaluated our proposed transfer learning approaches using the dataset described in Section 3.1. As shown in Figure 1, we fine-tune pre-trained models using various data and fine-tuning strategies to demonstrate the performance of each pre-trained model in sequence labeling tasks for capitalization and punctuation restoration. We provide the results obtained using different pre-trained models including both monolingual for Catalan and Galician (Bertinho, and BERTa), and multilingual (mBERT, DistilmBERT and XLM-RoBERTa).

4.1 Galician

In Table 3, we report our experimental results on the Galician models with Macro-f1 and Weighted-f1. All models are evaluated using Macro-f1 over 12 classes to evaluate the individual performance of each punctuation mark and Weighted-f1 to see the overall performance of the models.

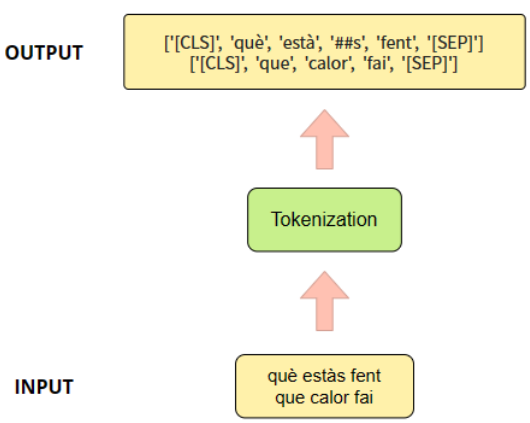
As can be seen in Table 3, all models with augmentation have obtained the best results. Monolingual models, such as Bertinho, perform better than models with a more complex architecture and a larger corpus (DistilmBERT). However, RoBERTinha, which is a monolingual model trained on a reduced corpus, obtained the worst result. XLM-RoBERTa archives a better result than the other models, as it was trained on a large corpus and has a large vocabulary. Our best result is obtained using XLM-RoBERTa with augmentation, and it has a 70,91% accuracy in Macro-f1 and 90.199% overall performance.

In Table 4, the evaluation of each punctuation and capitalization label of the XLM-RoBERTa model with augmentation are displayed. As can be observed, the label that indicates the token is capitalized and has an exclamation mark ('`!u`') or a colon ('`:u`') are the ones that obtain the lowest Macro-f1 because the number of occurrences (see Section 3.1) in this dataset is not sufficient for proper training and evaluation. However, the model predicts capitalized words with question marks ('`?u`') with an accuracy of 67.68%, despite having few occurrences in the training set.

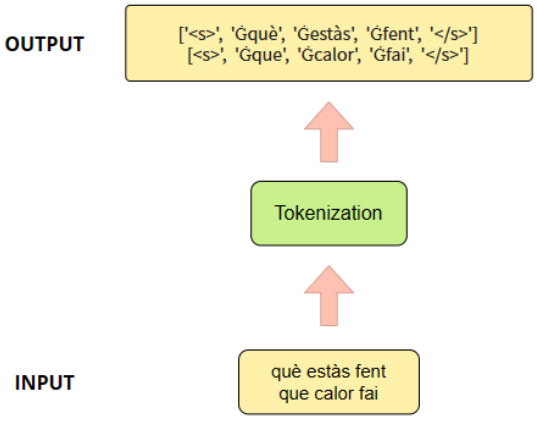
4.2 Catalan

Table 5 shows the macro-averaged *F1* score and weighted-averaged *F1* score for each experiment with the combination of different datasets and the pre-trained model for Catalan. As can be seen, the monolingual models perform better than the multilingual models because the multilingual models (such as mBERT and DistilmBERT) have lower Catalan language content in the training data. BERTa archives a better result than the other models, as it was trained on a large Catalan corpus and has a large vocabulary. Our best result is obtained using BERTa with augmentation, and it has a 69,34% accuracy in Macro-f1 and 90.85% overall performance.

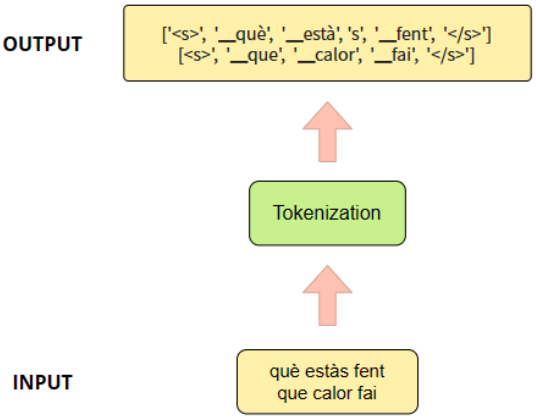
In Table 6, the evaluation of each punctuation and capitalization token of the BERTa



(a) Examples of BERT tokenization



(b) Examples of RoBERTa tokenization



(c) Examples of XLM-RoBERTa tokenization

Figure 2: Examples of tokenization.

	Dataset		Augmented dataset	
Model	Macro-F1 avg	Weighed F1 avg	Macro-F1 avg	Weighted F1 avg
mBERT-cased	66.681	89.297	67.592	89.230
DistilmBERT	62.997	87.843	63.942	87.907
Bertinho	65.687	88.888	66.525	88.906
XLM-RoBERTa	70.448	90.370	**70.941**	**90.199**
RoBERTinha	58.807	87.164	60.252	87.174

Table 3: Results on the dataset and augmented dataset for test sets in Galician.

model with augmentation is shown. As can be seen, the same happens as with the Galician models, that the model is not accurate in classifying the tokens as `!u` and '`:u`' by their number of occurrences in the training set.

Table 4 and 6 illustrate that both models perform well in predicting capitalized words, and with our transfer learning-based sequence labeling approach, the models classify tokens based on the other words. So, they can identify proper names, country names, and organizations well, as shown in Figure 3 and 4.

5 Error analysis

In this section, we analyze the errors of the Catalan and Galician capitalization and punctuation restoration models. For this purpose, we have used the model that has obtained the best result according to Table 5 and 3. To evaluate the performance of these models and to check in which case the models give erroneous predictions, a normalized confusion matrix with truth labels has

	Precision	Recall	F1-score
!l	0.62340	0.62807	0.62573
!u	0.53280	0.50187	0.51688
,l	0.66314	0.69474	0.67857
,u	0.67739	0.68365	0.68050
.l	0.77420	0.73895	0.75616
.u	0.74133	0.75590	0.74854
:l	0.66059	0.62150	0.64045
:u	0.60220	0.56994	0.58563
?l	0.80984	0.78769	0.79861
?u	0.75980	0.61024	0.67686
l	0.95435	0.95795	0.95615
u	0.85912	0.83876	0.84882
Macro avg	0.72151	0.69910	**0.70941**
Weighted avg	0.90205	0.90209	**0.90199**

Table 4: Classification report of the XLM-RoBERTa with data augmentation.

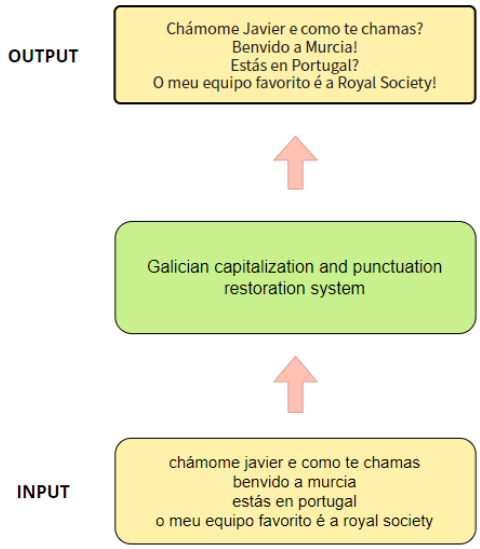


Figure 3: Galician capitalization restoration system examples.

been used, which consists of a table showing the distribution of the predictions of a model compared to the truth label of the data. The confusion matrix of both models is shown in Figure 5.

Concerning the model of capitalization and punctuation retrieval in Catalan, taking into account the confusion matrix (see Figure 5a), it is observed that it does not make many relevant classification errors, such as confusing a comma with a period, and colons with a period marks, which would affect the sentence ending early. Therefore, the focus can be set on the relationship in other punctu-

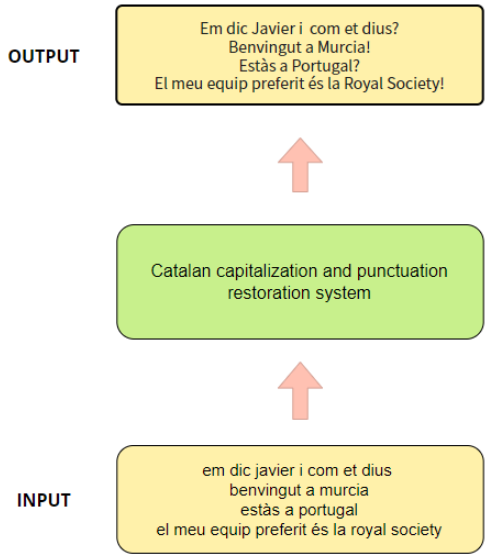


Figure 4: Catalan capitalization restoration system examples.

ations. Through the confusion matrix, it is observed that the model often confuses period marks with the exclamation, and colons with commas. Thus, a set of examples from the test dataset misclassified by the BERTa model trained with the augmented dataset has been analyzed. Table 7 shows the misclassified examples, and it can be seen that it is often very difficult to differentiate between periods and exclamations, as both punctuation marks are placed at the end of the sentence and the only difference is that exclamations are used to show emphasis or an emotional exclamation. Therefore, across texts it is difficult to identify the emotions of a sentence.

Furthermore, in Table 7 we can see that sentences with pronouns such as *what*, *who*, *how*, *where*, *when*, and *which* are ambiguous, as they can be both interrogative and exclamatory pronouns. In our models, when it receives a sentence with only one word and this word is one of the pronouns mentioned above, it always classifies it as an interrogative pronoun (‘`?u`’) instead of an exclamatory pronoun (‘`!u`’). In this case, it is very difficult to solve this problem, as both solutions are valid and the sentence has only one word, so our models cannot use word relation to classify it well.

With respect to Galician capitalization and punctuation restoration model, we have analyzed the confusion matrix (see Figure 5b) and the different misclassified examples of the

	Dataset		Augmented dataset	
Model	Macro-F1 avg	Weighed F1 avg	Macro-F1 avg	Weighted F1 avg
mBERT-cased	65.174	89.341	65.050	89.334
DistilmBERT	60.391	88.042	61.981	88.223
XLM-RoBERTa	69.354	90.753	69.269	90.679
BERTa	68.762	90.735	**69.343**	**90.858**

Table 5: Results on the dataset and augmented dataset for test sets in Catalan.

	Precision	Recall	F1-score
!l	0.63830	0.61224	0.62500
!u	0.52941	0.45652	0.49027
,l	0.68776	0.69414	0.69093
,u	0.65998	0.66643	0.66319
.l	0.77012	0.74903	0.75943
.u	0.69411	0.66018	0.67672
:l	0.59961	0.57685	0.58801
:u	0.57850	0.53147	0.55399
?l	0.79294	0.77802	0.78541
?u	0.69670	0.68764	0.69214
l	0.95894	0.96376	0.96134
u	0.84158	0.82810	0.83479
Macro avg	0.70400	0.68370	**0.69343**
Weighted avg	0.90823	0.90900	**0.90858**

Table 6: Classification report of the BERTa with data augmentation.

XLM-RoBERTa model trained with the augmented dataset. We have seen that the same thing happens as in the Catalan model. The model does not make many relevant classification errors, such as confusing a comma with a period, and colons with period marks, which would affect the sentence ending early. However, it often confuses periods with exclamations and colons with commas, and always classifies pronouns as interrogative.

6 Conclusions and further work

This paper presents two models of capitalization and punctuation restoration, one for Catalan and one for Galician, based on a transfer learning approach through different pre-trained models. The system has been trained for 5 types of punctuation and 2 types of capitalization. In addition, the models are able to identify certain proper names and names of countries and organizations for the capitalization restoration task. Both models have been trained and tested with the OpusParaCrawl dataset. Moreover, we propose an augmentation technique, which improves the performance of the models by up to 1.45% for some models such as RoBERTinha. Our best result is obtained using XLM-RoBERTa with data augmentation for Galician and using BERTa with data augmentation for Catalan. Both have achieved excellent performance with a macro-average *F1* score of 70.94% and overall performance of 90.2% for the Galician, and a macro-average *F1* score of 69.34% and 90.86% of overall performance for the Catalan.

As future work, we would like to use the same approach to create a capitalization and punctuation restoration system for Spanish and compare the performance with other models, such as Zhu et al. (2022), and González-Docasal et al. (2021). And the last proposal is to test a new pre-training model called *Whisper* and see if it works in Catalan and Galician and compare the results with other models, and develop a model that takes into account the relationships of the previous sentence with the following sentence to increase the accuracy of the models and resolve the errors discussed in Section 5.

The models are available on the Huggingface platform[1,2]. In addition, a demo application[3] of these models has also been created for the user to test them in real time. Additional resources concerning to this paper can be accessed.[4]

Acknowledgements

This work is part of the research project (2021/C005/00150076) funded by Spanish Government - Ministerio de Asuntos

[1] https://huggingface.co/UMUTeam/catalan_capitalization_punctuation_restoration

[2] https://huggingface.co/UMUTeam/galician_capitalization_punctuation_restoration

[3] https://huggingface.co/spaces/UMUTeam/punctuation_and_capitalization_restoration

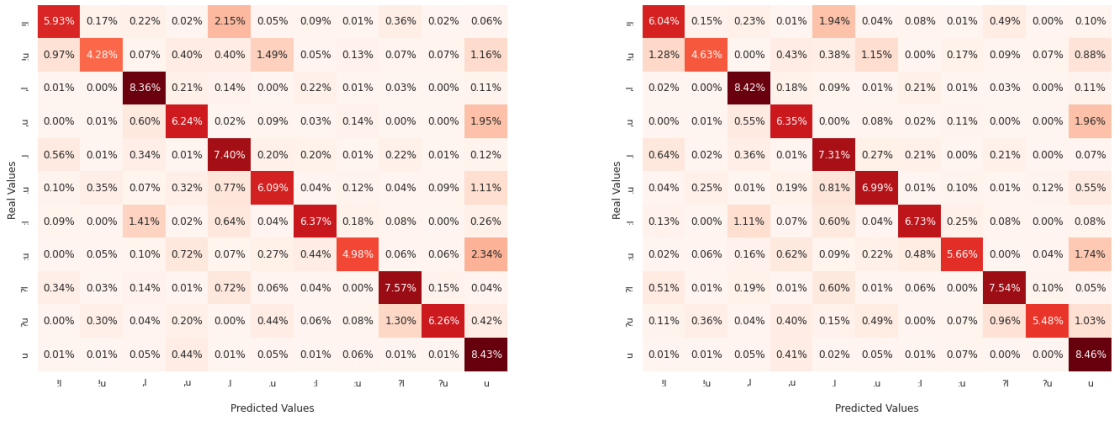
[4] https://github.com/NLP-UMUTeam/capitalization-and-punctuation-restoration

Predicted	Real
Què?	Què!
Com?	Com!
On?	On!
Qui?	Qui!
Quin?	Quin!
Estic bé!	Estic bé.
Et vaig fer el sopar: sopa i truita!	Et vaig fer el sopar: sopa i truita.
Fresca, neta i pura, així és l'aigua de font.	Fresca, neta i pura: així és l'aigua de font.
Aquesta feina no és el meu somni, és una feina amb prou feines.	Aquesta feina no és el meu somni: és una feina amb prou feines.

Table 7: A set of examples misclassified by the BERTa model trained with the augmented dataset for Catalan.

Predicted	Real
que?	Que!
Quen?	Quen!
Cal?	Cal!
Canto?	Canto!
onde?	Onde!
Estou ben!	Estou ben.
Hoxe chegou tarde!	Hoxe chegou tarde.
Querido amigo, hai moito que non sei nada de ti!	Querido amigo: Hai moito que non sei nada de ti.
Naquela libraría había de todo, libros, xornais, cómics.	Naquela libraría había de todo: libros, xornais, cómics.

Table 8: A set of examples misclassified by the XLM-RoBERTa model trained with the augmented dataset for Galician.



(a) Catalan (b) Galician

Figure 5: Confusion matrix of Catalan and Galician capitalization and punctuation restoration system.

Económicos y Transformación and by the European Union NextGenerationEU/PRTR. This work is also part of the research project LaTe4PSP (PID2019-107652RB-I00/AEI/ 10.13039/501100011033) funded by MCIN/AEI/10.13039/501100011033.

References

Alam, T., A. Khan, and F. Alam. 2020. Punctuation restoration using transformer models for high-and low-resource languages. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 132–142, Online, November. Association for Computational Linguistics.

Armengol-Estapé, J., C. P. Carrino, C. Rodriguez-Penagos, O. de Gibert Bonet, C. Armentano-Oller, A. Gonzalez-Agirre, M. Melero, and M. Villegas. 2021. Are multilingual

models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4933–4946, Online, August. Association for Computational Linguistics.

Bannard, C. and C. Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Bañón, M., P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. L. Forcada, A. Kamran, F. Kirefu, P. Koehn, S. Ortiz Rojas, L. Pla Sempere, G. Ramírez-Sánchez, E. Sarrías, M. Strelec, B. Thompson, W. Waites, D. Wiggins, and J. Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online, July. Association for Computational Linguistics.

Basili, R., C. Bosco, R. Delmonte, A. Moschitti, and M. Simi, editors. 2015. *Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project*, volume 589 of *Studies in Computational Intelligence*. Springer.

Bostrom, K. and G. Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. *CoRR*, abs/2004.03720.

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*.

Che, X., C. Wang, H. Yang, and C. Meinel. 2016. Punctuation prediction for unsegmented transcript based on word vector. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 654–658, Portorož, Slovenia, May. European Language Resources Association (ELRA).

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. *CoRR*, abs/1911.02116.

Courtland, M., A. Faulkner, and G. McElvain. 2020. Efficient automatic punctuation restoration using bidirectional transformers with robust inference. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 272–279, Online, July. Association for Computational Linguistics.

David Vilares, Marcos Garcia, C. G.-R. 2021. Bertinho: Galician bert representations. *Procesamiento del Lenguaje Natural*, 66(0):13–26.

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.

Federico, M., M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker. 2012. Overview of the IWSLT 2012 evaluation campaign. In *Proceedings of the 9th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 12–33, Hong Kong, Table of contents, December 6-7.

González-Docasal, A., A. García-Pablos, H. Arzelus, and A. Álvarez. 2021. Autopunct: A bert-based automatic punctuation and capitalisation system for spanish and basque. *Procesamiento del Lenguaje Natural*, 67(0):59–68.

Jones, D., F. Wolf, E. Gibson, E. Williams, E. Fedorenko, D. Reynolds, and M. Zissman. 2003. Measuring the readability of automatic speech-to-text transcripts. In *8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland, September 1-4, 2003*. ISCA, 09.

Ljubešić, N. and A. Toral. 2014. cawac - a web corpus of catalan and its ap-

plication to language modeling and machine translation. In N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

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.

Peitz, S., M. Freitag, A. Mauser, and H. Ney. 2011. Modeling punctuation prediction as machine translation. In *Proceedings of the 8th International Workshop on Spoken Language Translation: Papers*, pages 238–245, 12.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Tiedemann, J. 2012. Parallel data, tools and interfaces in opus. In N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Tilk, O. and T. Alumäe. 2016. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *INTERSPEECH*.

Yi, J. and J. Tao. 2019. Self-attention based model for punctuation prediction using word and speech embeddings. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7270–7274.

Yi, J., J. Tao, Y. Bai, Z. Tian, and C. Fan. 2020. Adversarial transfer learning for punctuation restoration.

Zhu, X., S. Gardiner, D. Rossouw, T. Roldán, and S. Corston-Oliver. 2022. Punctuation restoration in Spanish customer support transcripts using transfer learning. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 80–89, Hybrid, July. Association for Computational Linguistics.

Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 39-51 recibido 12-12-2022 revisado 31-01-2023 aceptado 09-02-2023

RoBERTime: A novel model for the detection of temporal expressions in Spanish

RoBERTime: un nuevo modelo para la detección de expresiones temporales en español

Alejandro Sánchez-de-Castro-Fernández[1], **Lourdes Araujo**[1,2], **Juan Martinez-Romo**[1,2]
[1]Universidad Nacional de Educación a Distancia (UNED), 28040, Madrid
[2]Instituto Mixto UNED-ISCIII IMIENS
{asanchez, lurdes, juaner}@lsi.uned.es

Abstract: Temporal expressions are all those words that refer to temporality. Their detection or extraction is a complex task, since it depends on the domain of the text, the language and the way they are written. Their study in Spanish and more specifically in the clinical domain is scarce, mainly due to the lack of annotated corpora. In this paper we propose the use of large language models to address the task, comparing the performance of five models of different characteristics. After a process of experimentation and fine tuning, a new model called RoBERTime is created for the detection of temporal expressions in Spanish, especially focused in the clinical domain. This model is publicly available. RoBERTime achieves state-of-the-art results in the E3C and Timebank corpora, being the first public model for the detection of temporal expressions in Spanish specialized in the clinical domain.
Keywords: Temporal expressions, TimeML, Language models, Clinical domain.

Resumen: Las expresiones temporales son todas aquellas palabras que refieran temporalidad. Su detección o extracción es una tarea compleja, ya que depende del dominio del texto, del idioma y de la forma de escritura. Su estudio en español y más específicamente en el dominio clínico es escaso, debido principalmente a la falta de corpora anotados. En este trabajo se propone el uso de grandes modelos del lenguaje para abordar la tarea, comparando el rendimiento de cinco modelos de distintas características. Tras un proceso de experimentación y fine tuning, se logra crear un nuevo modelo llamado RoBERTime para la detección de expresiones temporales en español, especialmente centrado en el dominio clínico. Este modelo se encuentra disponible de forma pública. RoBERTime alcanza resultados del estado del arte en los corpus E3C y Timebank, siendo este el primer modelo publico en detección de expresiones temporales en español especializado en el dominio clínico.
Palabras clave: Expresiones temporales, TimeML, Modelos del lenguaje, Dominio clínico.

1 Introduction

The detection of time expressions is a task that can be included in the field of information extraction. The extraction of these terms or expressions is necessary in more complex tasks such as: text summarization, (Ng et al., 2014), *question answering* (Pampari et al., 2018), (Sun, Cheng, and Qu, 2018) or creation of temporal lines (Leeuwenberg and Moens, 2018).

Natural language processing models need to be able to temporally locate certain events that are relevant in the text. For example, a model that works as an assistant answering questions needs to know the order of events in order to be able to answer questions like *'Did a occured before b?'* Or in the case of models that work summarizing texts (Barros et al., 2019), they need to know the temporality of events in order to be able to summarize the information in a consistent manner.

Time expressions are terms that express temporality in some form. Expressions such as: *'yesterday'*, *'at 3:00 p.m.'* o *'every eight hours'* can be considered as time expressions. To detect these expressions, two factors are taken into account, the detection of the expression and the normalization of its value.

ISSN 1135-5948 DOI 10.26342/2023-70-3

Expression detection is the same as scope detection. This can be defined as the detection of at least part of a time expression, which is composed of tokens.

Sometimes temporal expressions are easily detected, because they usually follow syntactic patterns that are easily defined under a system of rules or regular expressions. But these patterns are both language-dependent (Lange et al., 2022), (Lange et al., 2020) as well as the domain of the text in question (Strötgen et al., 2014), (Strötgen and Gertz, 2013). This forces rule-based systems to be adapted, having to adjust existing rules and in many cases adding new rules. (Skukan, Glavaš, and Šnajder, 2014), (Li et al., 2014).

The identification of time expressions can be achieved through several methods, one of them and the most popular for years has been rule-based systems. In more recent years these methods have been displaced by large language models (*LLMs*) and the Transformers architecture (Vaswani et al., 2017). These models are capable of delivering good multi-task performance on small data sets by applying a *fine-tuning* process. This process consists of adjusting the weights of the model, fitting them to a new dataset. And due to the scarcity of annotated corpora these models are a strong candidate to replace the classical systems.

This paper proposes the creation of a new model called RoBERTime based on deep learning and LLMs for the extraction of time expressions, specifically for the detection of their extension or scope, in Spanish in general and in the medical domain in particular. This model is a pioneer in the Hispanic community, since to the authors' knowledge there is no other model based on deep learning/similar characteristics for the solution of this task in Spanish. RoBERTime is the result of a process of experimentation with five LLMs of different nature, on which different fine tuning techniques have been applied in order to understand the adaptability of the LLMs to this task. Finally, the findings of each experiment have been applied to maximize the performance of RoBERTime. For the experiments and for training BioRoBERTa, the Timebank corpus and the E3C corpus have been used.

The model presented has a dual purpose. The first is to serve as part of the task of extracting temporal lines in the medical domain, which is intended to help medical professionals to more easily understand the patient's history. On the other hand, the model is intended to serve as a baseline for the extraction of time expressions in Spanish.

The rest of the article is structured as follows: Section 2 discusses the state-of-the-art and the works related to the proposal. Section 3 presents the corpora that have been used in the process. Section 4 develops in detail our proposal. The methodology and the experiments carried out are explained in Section 5. The results obtained from the experiments are analysed and compared with the current state-of-the-art. Finally, Section 7 presents the main conclusions together with the lines of future work.

2 Related Work

The current scheme regarding the annotation of time expressions is TimeML 1.2.1 (Saurı et al., 2006), also defined as an ISO standard (Pustejovsky et al., 2010), in which time expressions are defined using the TIMEX3 tag. From this point on, when TimeML is mentioned, it will refer to version 1.2.1.

TimeML defines four types of time expressions: *DATE*, *TIME*, *DURATION*, and *SET*. In this order, dates, dates with a granularity of hour or less, durations and repetitions are defined. For example, *'April 12'* would be an expression of type *DATE*, *'3:15'* would be an expression of type *TIME*, *Two months* would be an expression of type *DURATION*, and *'every 8 hours'* would be an expression of type *SET*.

Two of the best-known systems for time expression extraction are HeidelTime (Strötgen and Gertz, 2010) and TIPSem (Llorens, Saquete, and Navarro, 2010). TIPSem was designed to work in both English and Spanish. It is a system based on the use of *conditional random fields* or CRFs (Lafferty, McCallum, and Pereira, 2001) and *semantic role labeling* or SRL (Gildea and Jurafsky, 2002). HeidelTime was designed to work in English but was eventually adapted to multiple languages, including Spanish. It is a rule-based system and is perhaps the most popular, as it is still available for use today, being one of the few systems with this availability.

Other systems with similar characteristics are: ClearTK (Bethard, 2013), a system based on support vector machines (SVMs)

(Vapnik, 1999), (Cortes and Vapnik, 1995) and SUTime (Chang and Manning, 2012), a rule-based system. Both are designed to operate in English only, show similar performance to HeidelTime, and are publicly available[1][2]. Despite showing similar or even superior performance in some aspects to HeidelTime, HeidelTime has maintained its popularity over time by being adapted to multiple languages (Skukan, Glavaš, and Šnajder, 2014), (Li et al., 2014).

In Clinical TempEval (Bethard et al., 2017), a shared task held in 2017, the organizers proposed time expression extraction changing the subdomain for training and test. Specifically, they proposed to train the systems on a dataset dealing with colon cancer and test their performance on a dataset dealing with brain cancer. The results show a drop in performance of more than twenty points compared to systems trained and tested on colon cancer in detecting the scope of time expressions, thus showing the difficulty in adapting to the domain.

A system called Annotador (Navas-Loro and Rodríguez-Doncel, 2020), based on rules for English and Spanish, has recently been released, which performs better in some aspects than HeidelTime. This system is intended for use on general domain documents but is specialized in the legal domain.

Given this context, the most recent systems are based on LLMs and Transformers, as classical systems have probably reached their limit in time expression extraction and everything seems to indicate that the context understanding capability of LLMs can be applied to this task.

Different approaches can be applied to LLMs. Thus in (Almasian, Aumiller, and Gertz, 2021) two different approaches are proposed for detecting the scope of time expressions, that of *token classification* and that of *seq2seq*. In the former, the text is viewed as a sequence of tokens in which each token may or may not be part of a time expression. Time expressions can be composed of several words so it is necessary to identify which is the beginning and which is not. For example in the expression *'April 12'*, *12* will be marked as the beginning of the expression and *'April'* as part of the expression. In this way, the model can be trained to solve the token classification task. This is a relatively similar approach to the one applied with CRFs and SVMs. On the other hand, the *seq2seq* approach is a text generation problem. The model receives the raw text and has to generate the annotated text in an xml annotation format. In the example of *'April 12'* the expression would be annotated as $< TIMEX3 : "DATE" > 12 \quad of \quad April < /TIMEX3 >$. In this paper they leave aside the value normalization of expressions since applying this approach it is understood as a separate problem from that of expression extraction. This paper shows that these models are capable of outperforming rule-based systems.

In (Chen, Wang, and Karlsson, 2019) the performance of BERT is compared with that of a linear model, a *multi layer perceptron* (MLP), a *bidirectional long short term memory* (BiLSTM) and LSTM. On the one hand, BERT is trained on several datasets and its performance is measured on them. On the other hand, BERT and GloVe (Pennington, Socher, and Manning, 2014) are used as a feature extractor. These features are passed to the models mentioned before to measure their subsequent performance. The results show that retraining BERT on the datasets gives better performance in extracting the time expressions, regardless of normalization, than using BERT and GloVe as a feature extractor to train the rest of the models. Also this retrained version of BERT achieves better results in two of the three corpora used than the baseline systems: Syntime (Zhong, Sun, and Cambria, 2017), TOMN (Zhong and Cambria, 2018) and PTime (Ding et al., 2019). These three systems have superior performance to HeidelTime, ClearTK and SUTime and two of them, Syntime[3] and PTime[4] have their code publicly maintained. Despite this, these models are less popular than HeidelTime.

In (Aumiller et al., 2022) it is proposed a web service that works for the extraction of time expressions using some of the models proposed in (Almasian, Aumiller, and Gertz, 2021).

There are works such as (Almasian, Aumiller, and Gertz, 2022) in which the ELECTRA (Clark et al., 2020) architecture is used for time expression extraction in German, re-

[1] nlp.stanford.edu/software/sutime.shtml
[2] cleartk.github.io/cleartk.html
[3] github.com/xszhong/syntime
[4] ws.nju.edu.cn/ptime/

	Timebank	E3C	Total
Date	1589	241	1830
Time	155	30	185
Duration	757	552	1309
Set	49	64	113
Total	**2550**	**887**	3437
None	64541	27928	92469

Table 1: Number of tokens annotated for each type of time expression to train on the Timebank and E3C corpora.

gardless of normalization, outperforming HeidelTime.

3 Corpora

Two corpora have been used for the development of this work: TimeBank (Nieto, Saurí, and Poveda, 2011) and E3C (Magnini et al., 2020). Timebank is a corpus of annotated journalistic documents. E3C is also a multilingual corpus whose resources will only be used in Spanish. In this case, the corpus includes annotated documents from the medical domain, specifically clinical cases. Both corpora have temporal annotations following the TimeML scheme, however, E3C has an extra type of annotations called *'PREPOSTEXP'*. These expressions are of the pre/post-operative type, such as *'postoperative'* or *'post-surgical'*. They will not be taken into account because they only appear in E3C and could hinder the detection of other types of expressions.

The sizes of both corpora can be found in Table 1. As we can see, Timebank is approximately three times larger than E3C. Two of the types of expressions are clearly in the minority, *TIME* and *SET*, so the detection of these types of expressions will be more complicated.

4 Proposal

Five models with different characteristics are explored with the goal of comparing performance based on the main characteristics of the models and creating a new model trained specifically for the task in the clinical domain. All models used are publicly available on HuggingFace. The models considered along with their distinguishing features are described below:

- RoBERTa biomedical clinical (Carrino et al., 2021): A RoBERTa-based model trained on a corpus with biomedical and clinical terminology. It is the only model of those considered that has specialized vocabulary in the clinical domain. For short, this model will be referred to as BioRoBERTa.
- BETO-uncased (Canete et al., 2020): A BERT-based model trained on a Spanish corpus for the purpose of solving a wide range of tasks in Spanish.
- BETO-NER: Model based on BETO and trained on different Spanish CONLL corpora (Sang and Buchholz, 2000), (Tjong Kim Sang, 2002), (Nivre et al., 2007) for the task of *Named Entity Recognition*. With BETO-NER we intend to study the impact of pre-training the models on the task to be worked on, comparing their performance with BETO.
- Tiny BETO-NER: A distilled version of BETO-NER was trained at the same time. A distilled or reduced model is obtained from a distillation process, whereby much of the knowledge is transferred from one model to another by reducing its size. This model has a size of approximately 13 % compared to BETO-NER, maintaining a 78 % of the performance in some tasks. The use of this model will allow studying the adaptability of very small models to other domains.
- DistilBERT-m (Sanh et al., 2019): This model is a distilled version of multilingual BERT, with a relative size of 60 % maintains 97 % of the performance. This model has the largest number of parameters of the five considered. The main feature of this model is its multilingual capability, as it will allow us to study how it adapts to the task in Spanish compared to the other models.

5 Methodology

A series of experiments will be carried out in order to compare the different models considered with each other, in addition to a series of training techniques. The experiments will be performed on a batch size of 16, learning rate of $8e^{-5}$, weight decay 0.1 and 24 epoch. This learning rate has been used and not the standard value of $2e^{-5}$ because based on experience working with these corpora and models it has been found that this value gives

better results. As stated in (Mosbach, Andriushchenko, and Klakow, 2020), to improve training stability when training on small corpora it is preferable to train over a large number of epochs, until the training loss is close to 0. Therefore, the checkpoint of the model with the best f1 over the 24 training epochs will be the one shown in the experiments.

Two metrics will be used to evaluate the models *Seqeval* (Ramshaw and Marcus, 1999), (Nakayama, 2018) and *TempEval-3 toolkit*[5] (UzZaman et al., 2013). Both metrics are designed for the evaluation of *token classification* tasks, but they do not count positive and negative cases in the same way. TempEval-3 toolkit calculates the f1 metric for both fully detected (*strict*) and partially detected (*relaxed*) expressions. A strict match is given when the model predicts the full expression, whereas in a relaxed match the models predicts part of the expression. For example, if the annotated expression is 'El martes 12', and the model predicts 'El martes', it would be counted as a relaxed match. It is the most popular metric for evaluating time expression extraction, so it will be used in the final phase when comparing the performance of RoBERTime with systems from other papers. Seqeval has been used throughout the experimentation and model evaluation phase, as it was much easier to integrate than TempEval-3 toolkit. Seqeval has been used in strict mode and IOB2 scheme.

On the one hand, we are going to test which loss function offers better performance, cross entropy or focal loss (Lin et al., 2017). Both functions are very similar, with the difference that in focal loss a parameter is added to compensate for the most difficult cases to detect. For this purpose, a fine-tuning process is performed on the BioRoBERTa model. For this process we have used the E3C data with a random training/dev split, which has been maintained throughout the experiment. These two functions accept a set of weights representing the importance to be given to each type of expression, since in this case there are many more tokens that are not annotated so this imbalance must be compensated. The weights are calculated with the following formula:

$$W_{n,c} = 1 - \frac{instances_c}{\sum_{c=1}^{n} instances_c}$$

Figure 1: Where n is the total number of classes and c the class for which the weights W are to be calculated.

In order to create a model adapted to the task under consideration, the impact of training the models considered with each of the corpora (or combinations of parts of them) has been studied. In this way it is possible to study the potential of each corpus separately. Cross-folding has been applied with the maximum number of splits allowed for both corpora, two in the case of E3C and three in the case of Timebank with a fixed seed equal to 42, in order to be able to replicate the splits in each experiment. The maximum number of splits is marked by the maximum number of splits that can be made from the data without leaving any of the expression types unrepresented in the split.

To study the impact of merging the two corpora, two experiments have been performed on the BioRoBERTa model. One of them proposes a layer freezing method and the other is based on maximizing the training data of one of the corpora:

- *Join both corpora in a single fine-tuning process:* In this case, one of the two corpora is partially used for evaluation using cross-folding splits, while the other is used in its entirety as training data. In this way we seek to maximize the training set while maintaining a sufficiently representative validation set. Moreover, since the validation sets are the same as those used when training the models on each corpus separately, the results can be directly compared.
- *Train first with Timebank and retrain with E3C freezing different layers of the model:* In this way, the model is trained on the majority corpus, Timebank, performing cross-folding and choosing the best split. Subsequently, this model is trained on the minority corpus, E3C, with cross-folding while freezing different layers of the model. This layer freezing method has been studied in several works such as (Lee, Tang, and Lin, 2019), (Eberhard and Zesch, 2021) for

[5] github.com/naushadzaman/tempeval3_toolkit

language and domain adaptation.

Finally, once the methods described above have been compared, the one that maximizes the performance on E3C is selected and the rest of the models are trained with it.

After completion of the experimentation process, the best performing model is selected for publication and to compare its performance with HeidelTime and Annotador. The results of these systems on Timebank have been obtained from (Navas-Loro and Rodríguez-Doncel, 2020), while the results on E3C have been obtained using the TempEval-3 toolkit by ourselves. To obtain the HeidelTime annotations on E3C, the Philip Hausner repository was used[6] and for Annotador the María Navas repository[7].

All experiments have been performed in blocks of five iterations to minimize the random factor in model training. And to favor reproducibility each block has the same set of seeds: 42, 52, 62, 72, 82. The results shown as f1 metric have been calculated as the arithmetic mean of the five experiments.

6 Results

This section will summarize and analyse the results obtained in the experiments proposed in the Section 5 of this work.

6.1 Focal loss versus Cross entropy

The results of comparing the focal loss function against the cross entropy function can be seen in Table 2. Focal loss is slightly superior in three of the four types of time expressions and in the weighted average. This may be mainly due to the fact that this function gives more importance to the cases that are more difficult for the model to detect, which at the same time are usually the minority cases. This difference is noticeable in the *SET* expressions. As for the *TIME* expressions, it is possible that the difference in favor of the cross entropy function is due to the random factor in the training of the model. This difference will not be taken into account since the performance on the *TIME* expressions is too close to zero. Given these results, focal loss is chosen over cross entropy.

	Cross entropy	Focal loss
Date	0.5972	**0.6174**
Time	**0.0173**	0
Duration	0.6756	**0.678**
Set	0.191	**0.2579**
Mean	0.5966	**0.6099**

Table 2: Comparison of the F1 measure results for each class of the E3C corpus on the two loss functions used to train BioRoBERTa.

	Timebank	Mean	Split
brob	0.8029	0.79716	2
beto	0.766	0.7377	2
btn	**0.8137**	0.**7986**	2
tbtn	0.1938	0.1817	3
mbr	0.748	0.7431	3
	E3C	Mean	Split
brob	**0.5831**	**0.58045**	1
beto	0.4643	0.4482	1
btn	0.5146	0.4978	1
tbtn	0.0617	0.0511	2
mbr	0.5157	0.5057	1

Table 3: F1 measure results of each model for the best Timebank (*tb*) and E3C split, along with the average f1 of all splits. Each abbreviation corresponds, from top to bottom, with BioRoBERTa, BETO, BETO-NER, Tiny BETO-NER, and DistilBERT-m.

6.2 Performance of the models on each corpus

The f1 metric results using the Seqeval approach for each model can be found in Table 3. The results of the best split are presented, which are quite homogeneous.

As for the corpora, it can be seen that the models perform better with Timebank than with E3C. This may be mainly due to the sizes of both corpora. To support this idea, a data augmentation technique based on duplicating the records of the E3C training set while keeping the same test set of Table 3 has been tested. This resulted in improving the performance of the f1 metric by 6.44 %.

The model results show that BETO-NER is the best option for Timebank, while BioRoBERTa is the best for E3C. There are multiple factors that can explain this behaviour. On the one hand E3C is composed of documents from the clinical domain, so a mo-

[6] github.com/PhilipEHausner/python_heideltime
[7] github.com/mnavasloro/Annotador

del that has a specialized vocabulary for it should be able to provide better performance. Similarly, BETO and BETO-NER are two models trained in part with documents and newspaper articles and Timebank is built on news documents. BETO-NER also outperforms BETO, so pre-training the models on the task seems to carry relevant weight.

It can be seen that DistilBERT-m performs better in E3C than BETO and BETO-NER, while the opposite is the case with Timebank. So a model with a larger number of parameters can maintain good results in different tasks and domains. But working on a specific task or domain, a model with fewer and specialized parameters can achieve better performance if the amount of data is sufficient.

Finally, Tiny BETO-NER shows a much lower performance than the other models. Being approximately nine times smaller, Tiny BETO-NER shows approximately four times lower performance in Timebank and nine times lower performance in E3C.

In order to maximize the performance of the model on E3C, two more experiments have been performed. In the next one, the layer freezing technique is tested, to try to make the model fit better to the changes introduced by E3C to a version of the model trained on Timebank. The second one follows the strategy of maximizing the data from one of the two corpora.

6.3 Timebank + E3C with freeze layers

The results of training BioRoBERTa on Timebank and training on Timebank and E3C freezing different layers of the model can be found in Tables 4 and 5. In order to enhance reproducibility, this model is available in a HuggingFace repository.

As we can see in the first row of table 4, when training the model solely on Timebank, the performance on E3C is similar to training with E3C in isolation (see Table 3). Again, it can be seen that there is a consistent difference between the splits. As can be seen in the column *Split 2* of both tables, this split boosts the E3C results the most, while it is the one that most impairs Timebank results and vice versa. The same is true for the number of frozen layers. Timebank results are increased as more layers are forzen, while the opposite is true for E3C. This behaviour can be explained by the performance when freezing layers, since freezing more layers will cause the model to retain more information, whereas freezing few layers will cause the model to update more information. However the results in Timebank were not expected to worsen when retraining on E3C, since both corpora are ultimately composed of time expressions of the same type and in the same language. Therefore, the model seems to present difficulties in generalizing to both domains, giving a trade-off situation between the two corpora. This is also evident when the model is trained with Timebank alone. In Table 5 BioRoBERTa achieves an f1 metric of 0.8159 on the Timebank test set while if trained first with Timebank and subse-

	E3C	
brob pre E3C	0.53361	
brob post E3C	Split 1	Split 2
0 layers	0.7075	0.7234
3 layers	**0.7202**	**0.7365**
6 layers	0.7129	0.7269
9 layers	0.6936	0.6803
Mean	0.7085	**0.7164**

Table 4: F1 measure results for the BioRoBERTa model on E3C test set. The model is first trained on Timebank (brob pre E3C row), selecting the best split. Subsequently, the trained model is retrained on the other E3C splits (split 1 and split 2) and freezing different layers.

	Timebank	
brob pre E3C	0.8159	
brob post E3C	Split 1	Split 2
0 layers	0.7551	0.7508
3 layers	0.7573	0.7456
6 layers	0.7602	0.7525
9 layers	**0.7699**	**0.7587**
Mean	**0.7606**	0.7519

Table 5: F1 measure results for the BioRoBERTa model on Timebank test set. The model is first trained on Timebank (brob pre E3C row), selecting the best split. Subsequently, the trained model is retrained on the other E3C splits (split 1 and split 2) and freezing different layers.

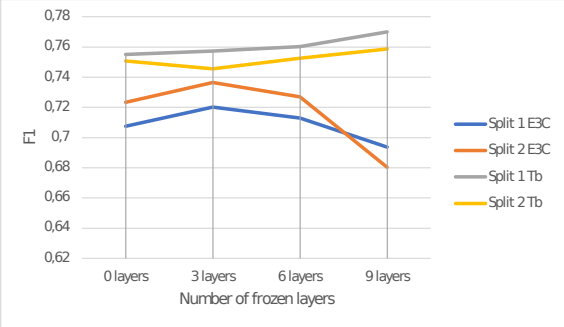


Figure 2: Evolution of the f1 metric according to the number of frozen layers on TimeBank and E3C. Each split corresponds to both E3C evaluation splits. Increasing the number of frozen layers increases TimeBank´s performance, while decreasing the E3C and vice versa.

quently with E3C with freezing, the average performance drops by 0.0553 points for split 1 and 0.064 for split 2. With E3C the opposite happens, the average performance improves by 0.1749 for split 1 and 0.1828 points for split 2.

The difference in performance can be clearly noticed depending on the number of layers that are frozen.

One would expect that the best performance over E3C would be achieved by not freezing any layers, since the model should completely adapt to the new data set. However, this does not occur and the best performance on E3C is given by freezing three layers, even freezing six layers improves performance over not freezing any in both splits. This behaviour may be due to the information shared between the two corpora. It may be the case that when training without any frozen layer the model loses relevant information acquired from Timebank stored in the initial layers. The trade-off is shown in Figure 2. It can be clearly seen how the performance of E3C decreases as more layers are frozen, while for Timebank it increases.

6.4 Timebank + E3C complete versus E3C + Timebank complete

Tables 6 and 7 shows the results of training the BioRoBERTa model on both corpora, using only one of them to perform the training and validation split.

From the mean values it can be seen that maximizing the Timebank set for training

			brob
Timebank split	Timebank	split 1	0.8036
		split 2	0.8028
		split 3	0.7922
		Mean	**0.7995**
	E3C	split 1	0.7406
		split 2	0.7201
		split 3	0.7249
		Mean	**0.7285**

Table 6: F1 measure results for BioRoBERTa model on both test sets, trained on: Timebank (without the validation split), and the whole E3C training sets.

			brob
E3C split	Timebank	split 1	0.8127
		split 2	0.8031
		Mean	**0.8079**
	E3C	split 1	0.7076
		split 2	0.7081
		Mean	**0.7079**

Table 7: F1 measure results for BioRoBERTa model on both test sets, trained on: E3C (without the validation split), and the whole Timebank training sets.

does not bring as much benefit as maximizing the E3C set does. When the entire E3C set is used for training, Table 6, there is a 3 % performance improvement over E3C and a 1 % drop in Timebank as compared to using the entire Timebank for training, Table 7. This may be because since E3C is a corpus with more complex cases for the model, if its representation is maximized in training, the model extrapolates those cases better to Timebank. On the contrary, if Timebank representation is maximized, the model is not able to use those extra cases to extrapolate to E3C.

Comparing these results with those of Tables 4 and 5 it can be seen that maximizing the amount of data for training is better than freezing layers of the model and training separately. The results on both corpora are, on average, better when this strategy is taken. Given these results and since the objective is to maximize the performance of the models for the clinical domain, we will choose the option of training the rest of the models using full E3C for training and Timebank to perform cross-folding splits for the evaluation set.

Tb	split 1	split 2	split 3	Mean
brob	0.8036	0.8028	0.7922	0.7995
beto	0.7384	0.7559	0.7209	0.7383
btn	0.8069	0.8233	0.798	**0.8093**
tbtn	0.2676	0.2612	0.2823	0.2702
mbr	0.7533	0.7604	0.7495	0.7544
E3C	split 1	split 2	split 3	Mean
brob	0.7406	0.7201	0.7249	**0.7285**
bt	0.6349	0.6242	0.6099	0.6229
btn	0.7123	0.6768	0.6987	0.6958
tbtn	0.2024	0.2246	0.2185	0,215
mbr	0.6683	0.6367	0.6629	0.6558

Table 8: Comparison of the different models considered for each Timebank split (*Tb*). The models were trained on Timebank split into validation and training, concatenating to the latter the full E3C training set.

6.5 Final results

The performance of all models on both corpora can be seen in Table 8. BETO-NER manages to have the best performance on Timebank, outperforming BioRoBERTa by 1,2 % on average. On the other hand, BioRoBERTa achieves better results over E3C, being superior by 4,7 %. Significant differences again stand out between BETO and BETO-NER, the latter being better by 11 % over E3C and 8 % over Timebank. DistilBERT-m, on the other hand, performs better than expected. Being a multi-language model and without any specialization in either task or domains, lower performance was expected. Tiny BETO-NER shows similar performance to those seen above. The performance differences based on splits again show the impact of finding a good split for the training data.

Given these results, BioRoBERTa has been chosen as the best model for this task. Because it is the model with the best performance on E3C and with a performance very close to BETO-NER on Timebank. As a result of the whole experimentation process, a model has been trained on the best options found. This new model based on BioRoBERTa, which we have called RoBERTime, is available at[8].

Table 9 shows the comparison of the best version of RoBERTime with HeidelTime and Annotador. It can be seen how the rule-based systems outperform RoBERTime over Timebank. This is mainly due to the fact that these systems were created with the purpose of giving good results on this corpus. This behaviour is shown by measuring the performance of HeidelTime and Annotator over E3C. As can be seen, the performance is considerably reduced with RoBERTime standing out above both.

Timebank	Strict	Relaxed	Type
RoBERTime	0.8152	0.8798	0.8504
Heideltime	**0.8533**	0.8907	0.8363
Annotador	0.8513	**0.9179**	**0.8923**
E3C	Strict	Relaxed	Type
RoBERTime	**0.7606**	**0.9108**	**0.8357**
Heideltime	0.5945	0.7558	0.6083
Annotador	0.6006	0.7347	0.5598

Table 9: Comparison on the f1 metric of TempEval-3 toolkit.

Regarding the performance of RoBERTime over Timebank, it can be seen how this model outperforms HeidelTime in detecting the time expression type. RoBERTime also shows good performance in detecting the type of expression in E3C. RoBERTime excels in this section over HeidelTime and Annotator.

About E3C, the big difference between strict detection and relaxed detection stands out, being the difference between both much bigger than in Timebank. This may be due to the fact that the expressions in E3C are composed of more tokens or are formulated in more varied ways than in Timebank, being more difficult to detect completely.

RoBERTime fails to positively detect some expressions such as *"actualmente"*, *"recientemente"*, four-digit numbers that do not correspond to dates such as *"2006"* or ages such as *"6 años"*. These expressions are particularly difficult to detect since sometimes it is necessary to take into account a large part of the text. There are also other expressions that are not annotated in E3C such as *"Ácido clavulánico 125 mgr, 1 comp. / 8 horas"* or *"isoniazid 300 mgr al día"* that RoBERTime detects.

In Timebank some cases have been detected in which the corpus has annotated age expressions as time expressions, as in the document with identifier *11033_20000817*: *"27 años"*, *"35 años"*, *"53 años"*. It is therefore possible that the model has at least partially acquired this behaviour from the annotations. There are other cases of ambiguities

[8] huggingface.co/asdc/Bio-RoBERTime

in the Timebank annotations that may confuse the model, such as annotating the expression *"hace un año"* in *"Al menos, hace un año, los camiones circulaban en la misma dirección"*, but not doing so in *"Hace un año los camiones te adelantaban a 70 o 80 kilómetros por hora"*.

These ambiguities are hard to treat, because they imply to make changes to the original annotations, and there might be some reason on why those expressions are or are not annotated. So it has to be accepted that this results are limitated by the annotators accuracy in both corpora.

7 Conclusions and future work

In this work we have presented a new model called RoBERTime which achieves state-of-the-art results in the detection of time expressions in Spanish in the journalistic domain with the Timebank corpus and in the clinical domain with E3C. In particular, E3C outperforms some of the most popular systems in all aspects. It has been proved that, unlike other systems, RoBERTime is able to adapt to both domains, showing a balanced behaviour on both corpora. This shows the great potential of LLMs to solve the task of time expression extraction. All this has been achieved through a series of experiments, which have allowed us to make decisions based on empirical results to maximize the performance of RoBERTime.

It has been observed in the performance of BETO-NER and BioRoBERTa that pretraining the models on the task and domain can considerably improve performance. The ability of LLMs to adapt to different tasks and domains has also been shown to be easier to shape than classical rule-based systems.

Although the main objective of the work was to achieve a time expression detection model for the clinical domain, the proposed model has turn out to perform better outside the clinical domain, in spite of the performance in E3C has been prioritized over that of Timebank. This may be due to the quality and quantity of data.

As for future work, the possibility of exploring the multilingualism of the corpora is being considered. On the one hand, this would make it possible to create multilingual models and, on the other hand, to translate annotations from other languages into Spanish, in order to increase the size of the available data. We are also considering exploring data augmentation techniques, since it has been observed that doubling the E3C records improves performance. We also consider the task of obtaining the normalized value of the expressions detected by RoBERTime, exploring rule-based system solutions and LLM-based solutions. For example, seq2seq models in which the expression would be taken as the input and the normalized value as the output. In line with this, the use of this model is proposed for the extraction of time lines, a task that requires the extraction of time expressions.

Acknowledgments

This work has been funded by the following projects DOTT-HEALTH (MCI/AEI/FEDER, UE with identification PID2019-106942RB-C32), OBSERMENH(MCIN/AEI/10.13039/501100011033 and NextGenerationEU"/PRTR with identification TED2021-130398B-C21) and by the project RAICES (IMIENS 2022).

References

Almasian, S., D. Aumiller, and M. Gertz. 2021. Bert got a date: Introducing transformers to temporal tagging. *arXiv preprint arXiv:2109.14927*.

Almasian, S., D. Aumiller, and M. Gertz. 2022. Time for some german? pre-training a transformer-based temporal tagger for german. In *Text2Story@ ECIR*, pages 83–90.

Aumiller, D., S. Almasian, D. Pohl, and M. Gertz. 2022. Online dateing: A web interface for temporal annotations. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3289–3294.

Barros, C., E. Lloret, E. Saquete, and B. Navarro-Colorado. 2019. Natsum: Narrative abstractive summarization through cross-document timeline generation. *Information Processing & Management*, 56(5):1775–1793.

Bethard, S. 2013. Cleartk-timeml: A minimalist approach to tempeval 2013. In *Second joint conference on lexical and computational semantics (* SEM), volume 2: proceedings of the seventh interna-*

tional workshop on semantic evaluation (SemEval 2013), pages 10–14.

Bethard, S., G. Savova, M. Palmer, and J. Pustejovsky. 2017. SemEval-2017 task 12: Clinical TempEval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572, Vancouver, Canada, August. Association for Computational Linguistics.

Canete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:1–10.

Carrino, C. P., J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, and M. Villegas. 2021. Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario. *arXiv preprint arXiv:2109.03570*.

Chang, A. X. and C. D. Manning. 2012. Sutime: A library for recognizing and normalizing time expressions. In *Lrec*, volume 3735, page 3740.

Chen, S., G. Wang, and B. Karlsson. 2019. Exploring word representations on time expression recognition. Technical report, Technical report, Microsoft Research Asia.

Clark, K., M.-T. Luong, Q. V. Le, and C. D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

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

Ding, W., G. Gao, L. Shi, and Y. Qu. 2019. A pattern-based approach to recognizing time expressions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6335–6342, Jul.

Eberhard, O. and T. Zesch. 2021. Effects of layer freezing on transferring a speech recognition system to under-resourced languages. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 208–212.

Gildea, D. and D. Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

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

Lange, L., A. Iurshina, H. Adel, and J. Strötgen. 2020. Adversarial alignment of multilingual models for extracting temporal expressions from text. *arXiv preprint arXiv:2005.09392*.

Lange, L., J. Strötgen, H. Adel, and D. Klakow. 2022. Multilingual normalization of temporal expressions with masked language models. *arXiv preprint arXiv:2205.10399*.

Lee, J., R. Tang, and J. Lin. 2019. What would elsa do? freezing layers during transformer fine-tuning. *arXiv preprint arXiv:1911.03090*.

Leeuwenberg, A. and M.-F. Moens. 2018. Temporal information extraction by predicting relative time-lines. *arXiv preprint arXiv:1808.09401*.

Li, H., J. Strötgen, J. Zell, and M. Gertz. 2014. Chinese temporal tagging with heideltime. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 133–137.

Lin, T.-Y., P. Goyal, R. Girshick, K. He, and P. Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Llorens, H., E. Saquete, and B. Navarro. 2010. Tipsem (english and spanish): Evaluating crfs and semantic roles in tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 284–291.

Magnini, B., B. Altuna, A. Lavelli, M. Speranza, and R. Zanoli. 2020. The e3c project: Collection and annotation of a multilingual corpus of clinical cases. In *CLiC-it*.

Mosbach, M., M. Andriushchenko, and D. Klakow. 2020. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines.

Nakayama, H. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Navas-Loro, M. and V. Rodríguez-Doncel. 2020. Annotador: a temporal tagger for spanish. *Journal of Intelligent & Fuzzy Systems*, 39(2):1979–1991.

Ng, J. P., Y. Chen, M.-Y. Kan, and Z. Li. 2014. Exploiting timelines to enhance multi-document summarization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 923–933.

Nieto, M. G., R. Saurí, and M. A. B. Poveda. 2011. Modes timebank: a modern spanish timebank corpus. *Procesamiento del lenguaje natural*, 47:259–267.

Nivre, J., J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. 2007. The conll 2007 shared task on dependency parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 915–932.

Pampari, A., P. Raghavan, J. Liang, and J. Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. *arXiv preprint arXiv:1809.00732*.

Pennington, J., R. Socher, and C. D. 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.

Pustejovsky, J., K. Lee, H. Bunt, and L. Romary. 2010. Iso-timeml: An international standard for semantic annotation. In *LREC*, volume 10, pages 394–397.

Ramshaw, L. A. and M. P. Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*. Springer, pages 157–176.

Sang, E. F. and S. Buchholz. 2000. Introduction to the conll-2000 shared task: Chunking. *arXiv preprint cs/0009008*.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Saurı, R., J. Littman, B. Knippen, R. Gaizauskas, A. Setzer, and J. Pustejovsky. 2006. Timeml annotation guidelines version 1.2. 1.

Skukan, L., G. Glavaš, and J. Šnajder. 2014. Heideltime. hr: extracting and normalizing temporal expressions in croatian. In *Proceedings of the 9th Slovenian Language Technologies Conferences (IS-LT 2014)*, pages 99–103.

Strötgen, J., T. Bögel, J. Zell, A. Armiti, T. V. Canh, and M. Gertz. 2014. Extending HeidelTime for temporal expressions referring to historic dates. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2390–2397, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Strötgen, J. and M. Gertz. 2010. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 321–324.

Strötgen, J. and M. Gertz. 2013. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298.

Sun, Y., G. Cheng, and Y. Qu. 2018. Reading comprehension with graph-based temporal-casual reasoning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 806–817.

Tjong Kim Sang, E. F. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

UzZaman, N., H. Llorens, L. Derczynski, J. Allen, M. Verhagen, and J. Pustejovsky. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings*

of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 1–9.

Vapnik, V. 1999. *The nature of statistical learning theory*. Springer science & business media.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhong, X. and E. Cambria. 2018. Time expression recognition using a constituent-based tagging scheme. In *Proceedings of the 2018 world wide web conference*, pages 983–992.

Zhong, X., A. Sun, and E. Cambria. 2017. Time expression analysis and recognition using syntactic token types and general heuristic rules. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 420–429.

Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 53-61 recibido 12-12-2022 revisado 02-02-2023 aceptado13-02-2023

Measuring language distance for historical texts in Basque

Cálculo de distancia lingüística para textos históricos en euskera

Ainara Estarrona[1], Izaskun Etxeberria[1], Manuel Padilla-Moyano[2] and Ander Soraluze[1]
[1]HiTZ Center - Ixa, University of the Basque Country UPV/EHU
[2]University of the Basque Country UPV/EHU
{*ainara.estarrona, izaskun.etxeberria, manuel.padilla, ander.soraluze*}*@ehu.eus*

Abstract: Measuring distance between languages, dialects and language varieties, both synchronically and diachronically, is a topic of growing interest in NLP. Based on our Syntactically Annotated Historical COrpus in BAsque (SAHCOBA) and previous work in perplexity-based language distance proposed by Gamallo, Pichel and Alegria (2017, 2020), we have compared historical corpora with current texts in the standard variety and calculated the language distances between them. As the standard Basque is based on the central dialects, the starting hypothesis is that the oldest texts and the dialects on the extremes will be the most distant. The results obtained have largely confirmed the thesis of traditional dialectology: peripheral dialects show a strong idiosyncrasy and are more distant from the rest.
Keywords: Language distance, dialectology, historical texts, perplexity.

Resumen: Medir la distancia entre diferentes lenguas, dialectos o variantes de lengua, tanto sincrónica como diacrónicamente, es un área de interés creciente dentro del PLN. Basándonos en el corpus histórico sintácticamente anotado del euskera (SAHCOBA), y en el trabajo previo realizado por Gamallo, Pichel y Alegría (2017, 2020) en relación con la distancia entre lenguas basada en perplejidad, hemos comparado textos históricos en euskera con textos actuales y hemos calculado la distancia entre ellos. Dado que el euskera estándar se basa en los dialectos centrales, la hipótesis inicial es que los textos más antiguos, así como los textos de los dialectos periféricos serán los más distantes. Los resultados obtenidos confirman de forma contundente las tesis propuestas por la dialectología tradicional: los dialectos periféricos muestran una fuerte idiosincrasia y su distancia respecto al estándar es mayor que la del resto de dialectos.
Palabras clave: Distancia lingüística, dialectología, textos históricos, perplexity.

1 Introduction

Measuring distance between languages, dialects and language varieties, both synchronically and diachronically, is a topic of growing interest in NLP. Under the BIM and SAHCOBA projects[1] we have collected the most relevant historical texts in Basque written in different dialects. As a next step, we hoped to quantify how different these texts are as a means to confirm and modulate theories about the historical and dialectal development of the language. For this purpose, we have compared the historical corpora with current texts in the standard variety and calculated the distances between them. As the standard is based on the central dialects, the starting hypothesis is that the oldest texts and the dialects on the extremes will be the most distant.

For measurements we have used information theory based on *perplexity*. Perplexity-based measures have been employed successfully for language identification (Gamallo et al., 2016), to calculate distance between

[1]Basque in the Making (BIM): A Historical Look at a European Language Isolate project (ANR-17-CE27-0011 - BIM, Agence Nationale de la Recherche, France) and the Syntactically Annotated Historical COrpus in BAsque (SAHCOBA, RTI2018-098082-J-I00) project (Ministry of Science and Innovation (MICINN), Spain).

ISSN 1135-5948 DOI 10.26342/2023-70-4

languages (Gamallo, Pichel, and Alegria, 2017b), and to quantify the diachronic distance in a language (Pichel, Gamallo, and Alegria, 2018). The software is open and, being an unsupervised method, only raw historical corpora are required.

The remainder of this paper is organised into several sections. Specific features of the Basque language and its dialects are introduced in Section 2. Section 3 is devoted to describing the corpus, while Section 4 covers why and how perplexity is applied. In Section 5 we detail the design of the experiments and briefly discuss the results in Section 6. Finally, Section 7 outlines our conclusions and possible future work.

2 *Basque language and dialects*

As a non-Indo-European language, indeed an isolate, Basque grammar differs considerably from that of the neighbouring languages. Basque is agglutinative, head-final, pro-drop, and usually assumed to be a Subject-Object-Verb (SOV) type language (de Rijk, 1969), but it is also described as having 'free word order', meaning that the order of phrases in the sentence can vary (Laka, 1996). Moreover, the Basque language exhibits a high level of dialectal fragmentation over an area of 10,000 km^2. The dialectal split began in the early Middle Ages (Mitxelena, 1981), and over the past few centuries the linguistic distance between dialects has been increasing to the extent that today peripheral varieties are not mutually intelligible in oral speech by non-trained speakers.

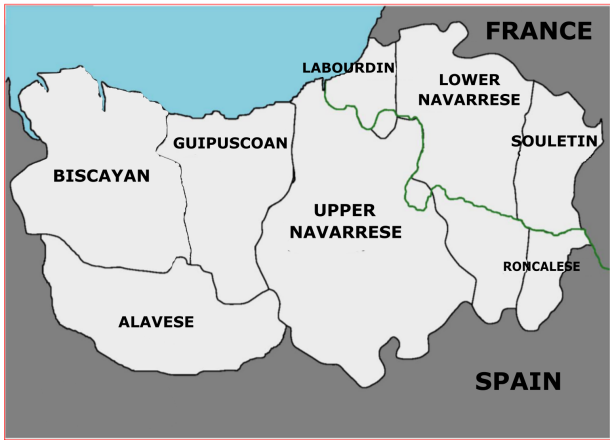


Figure 1: Historical Basque dialects. Alavese and Roncalese are extinct varieties. The green line represents the French-Spanish border.

At present, Zuazo (2014) distinguishes between five main Basque dialects: the Western dialect, traditionally called Biscayan, the Central dialect, traditionally known as Guipuscoan, the Navarrese dialect, the Navarrese-Labourdin dialect, and the Souletin dialect[2].

Figure 2: The five main dialects of the Basque Language (after Zuazo (2014)).

These five dialects are noticeably distinct from each other and, while there were sporadic attempts in the early twentieth century to bring some uniformity to Basque, it was not until 1968 that the Royal Academy of the Basque Language (founded in 1919)[3] decided to standardise it. Standard Basque (*Batua*) is a literary variety constructed upon central dialects of the language and historical dialects differ from standard Basque to varying degrees.

The distance between Basque dialects has often been a matter of discussion among linguists, but a scientific consideration of this problem requires some operational procedure for quantifying linguistic distance (Mitxelena, 1981). To our knowledge, the only attempts to quantify the differences between dialects have been based on dialectometry (Séguy, 1973) and have been carried out by linguists from the Eudia[4] group at the University of the Basque Country, including Aurrekoetxea, Gaminde, and Videgain, among others (Aurrekoetxea, 1992; Aurrekoetxea and Videgain, 2009; Aurrekoetxea et al., 2019). Their research does not deal with historical dialectology, but with dialects spoken

[2]http://euskalkiak.eus/en/ezaugarriak.php
[3]https://www.euskaltzaindia.eus/en/
[4]http://eudia.ehu.es/en/home/

today, from a synchronic perspective. However, language history and dialectology must go hand in hand (Camino, 2008) since every historical text is by definition a dialectal one.

Biscayan and Souletin are the two dialects at the corners. As traditional dialectological studies attest, they display the greatest differences from the others and have the most marked idiosyncrasy. The case of Biscayan is particularly relevant, as in the past certain scholars claimed that there were only two dialects: Biscayan on the one hand, and the central-eastern dialect, which would include the rest of the dialects, on the other (Lacombe, 1924). Although Biscayan has indeed noticeable characteristics that are lacking in the other dialects, it is no less true that many of these idiosyncrasies are innovations due to its peripheral character. Bear in mind that the lateral areas, unlike the central ones, are not only repositories of archaisms but also a breeding ground for innovations fostered by the heat of languages from the surrounding area (Mitxelena, 1981).

As we have already said the standard Basque (*Batua*) is based on the central dialects (mainly Guipuscoan and Labourdin), from which we can deduce that the peripheral dialects are the most distant from the standard. The main goal of this paper is to quantify the distance of the different historical dialects from standard Basque in order to confirm (or reject) existing dialectological theses in a quantitative way, thus contributing to historical dialectology from computational linguistics. To carry out this work we are going to use perplexity-based measures (see Section 4).

3 Corpus

Basque's historical corpus is quite scarce compared to those of neighbouring languages. Moreover, the corpus is asymmetrical geographically and historically: most varieties have significant gaps in their textual history, and at certain periods we do not have written records for all dialect. Along with scarcity and asymmetry, we must also mention homogeneity since, until the nineteenth century, works of a religious nature constituted more than 95% of the corpus (Lakarra, 1997). Most of these are also simple texts (doctrines, catechisms, etc.), which conceal many characteristics of the language, as lexicon, morphology and syntax are constrained by the type of discourse (Ulibarri, 2013).

As mentioned above, two projects have been involved in the creation of the Basque annotated historical corpus: BIM and SAHCOBA. The BIM-SAHCOBA corpus needed be representative of all dialects with a written tradition. Therefore, in these two projects we decided to establish a philologically reliable corpus covering most of the textual production between the fifteenth and mid-eighteenth centuries (Estarrona et al., 2021). This, on the one hand, is the minimal span that includes regular attestations for all Basque dialects and, on the other, is representative of the divide between Archaic and Old Basque from early modern Basque (Gorrochategui, Igartua, and Lakarra, 2018).

For the time being, we are creating a corpus of around one million words. Considering the issues associated with the written past of languages, especially in cases like Basque, this size is considered acceptable for a historical corpus (Claridge, 2009).

We have picked out nine works from the sixteenth and seventeenth centuries for our experiments. The oldest texts have been chosen because, although it is true that over time the dialects have become ever distant from each other (Mitxelena, 1981), it is no less true that the more recent the text, the closer it is to the standard variety. The main criterion for the choice of works was diversity of dialect. Thus, the texts selected are relevant to the history of Basque and that, in addition, reflect the main characteristics of each historical dialect. They are as follows [5]:

- Lazarraga's manuscript (1565)[6]
- *Iesus Krist Gure Iaunaren Testamentu Berria* (New Testament), Leizarraga (1571)
- *Dotrina Christiana. Bigarren impressionean debocionozco othoitz eta Oracino batçuez berreturic*, Materra (1617)
- *Gvero bi partetan partitua eta berecia*, Axular (1643)
- *Iesusen imitacionea*, Pouvreau (1669)

[5] The works are arranged by dialect and chronologically within each dialect.

[6] We will use the following abbreviations in the tables: Lazarraga=Laz; Leizarraga=Lç; Materra=Mat; Axular=Ax; Pouvreau=SP; Beriain=Ber; Kapanaga=Cap; Tartas=Tt: and Belapeire=Bp.

- *Tratado de como se ha de Oyr Missa*, Beriain (1621)
- *Exposición breve de la doctrina christiana*, Kapanaga (1656)
- *Onsa hilceco bidia*, Tartas (1666)
- *Catechima laburra eta Jesus-Christ Goure ginco jaunaren ecagutcia*, Belapeire(1696)

Table 1 shows the description of the selected works:

Author	Century	Dialect	Size
Laz	XVI	Alavese	12,072
Lç	XVI	Labourdin	73,906
Mat	XVII	Labourdin	16,323
Ax	XVII	Labourdin	90,029
SP	XVII	Labourdin	46,363
Ber	XVII	Upper Navarrese	14,995
Cap	XVII	Biscayan	11,408
Tt	XVII	Lower Navarrese	34,505
Bp	XVII	Souletin	23,735

Table 1: Works chosen for the experiments, century and year in which they were written, dialect and size.

Note that we consider Lazarraga's work as written in the Alavese dialect, which is an extinct variety. From today's perspective, Lazarraga's text is commonly considered a western dialect (Pagola, 2006), a classification created by Zuazo (2014).

As can be seen in Table 1, and due to the aforementioned asymmetry of the corpus, we do not have works in all the dialects for each century. As a case in point, there are no sources for the Guipuscoan dialect until the middle of the eighteenth century[7]. An interesting avenue for future work would be to quantify the distance of this central or Guipuscoan dialect with respect to the standard since in principle it should be the closest to it, both because the standard is based on the central dialects and because the texts in Guipuscoan are much more recent than those analysed in this paper.

Finally, we should mention the philological work that we carried out to begin from the best possible transcription of the works that we treated. We compared the transcriptions with their facsimiles (and/or with reliable critical editions) and, depending on the quality of each one, opted for one of the following: i) to correct the transcript, or ii) to create a new one. This task is highly time-consuming, but necessary to ensure our corpus is based on reliable versions of historical texts. The main criterion behind this philological effort is modernising the spelling — not to be confused with the adoption of present-day standard Basque orthography. In our corpus, the updating of spelling preserves the phonological shape of each text. For instance, Eastern Basque dialects have a set of aspirated plosive phonemes *ph, th, kh* that is not represented in the spelling system of standard Basque. However, we decided to maintain this phonological feature in the transcription of the texts (Estarrona et al., 2021).

4 Language distance. Perplexity

The main approaches to measuring language distance for historical or dialectal texts compare phonetic forms (Kondrak, 2005), "but some researchers have argued against the possibility of obtaining meaningful results from crosslingual comparison of phonetic forms" (Singh and Surana, 2007).

In computational linguistics, language models have been utilised for this purpose. The models and the calculation of crosslingual similarity are often based on word co-occurrences (Liu and Cong, 2013; Gao et al., 2014; Asgari and Mofrad, 2016). Recently, Degaetano-Ortlieb and Teich (2018) have used relative entropy for the detection and analysis of periods of diachronic linguistic change.

Perplexity-based measures are related to entropy and have been employed successfully for language identification (Gamallo et al., 2016), to measure the distance between languages (Gamallo, Pichel, and Alegria, 2017b), and to quantify the diachronic distance in a given language (Pichel, Gamallo, and Alegria, 2018). Basque appears in two of the experiments carried out by these authors, one comparing forty-four European languages (Gamallo, Pichel, and Alegria, 2017a) and the other selecting a handful of isolated languages to measure the distance between them (Gamallo, Pichel, and Alegria, 2020).

The method has been quite successful and,

[7]There are, however, small texts of few words collected in Michelena (1964), Sarasola (1983) and Satrustegi (1987).

in addition to language identification and historical linguistics (Scherrer, Samardžić, and Glaser, 2019; Zugarini, Tiezzi, and Maggini, 2020), has been used in other fields, including machine translation (Barrault et al., 2019), sociolinguistics (Chavula and Suleman, 2020) and sociology (Sant'Anna and Weller, 2020).

We use perplexity according to the methodology proposed by Pichel, Gamallo, and Alegria (2020) and the software they offer[8]. A language model's perplexity is defined as the inverse probability of the test text given the model. It is calculated comparing the n-grams (characters) of a text in one language/dialect with the n-gram model trained for another language/dialect (or between two historical periods of the same language). Lower perplexity would indicate lower distance between languages (or language periods). The comparison can be made in both directions because perplexity is a divergence with asymmetric values.

Due to the size of the historical texts, we have calculated the distance only in one direction, building the model for the standard language (larger corpus) and using historical texts as test corpus. In order to have comparable results, we configured the distance and the corpora with the same hyper-parameters as those used by the authors: 7-grams.

5 Design of the experiments

As discussed in the previous section, at least one corpus of standard Basque is required to carry out the experiments that measure distance between today's standard and the various historical dialects of the language. The corpus of historical Basque has been described in Section 3.

Regarding standard Basque, we believed it best to ensure the subject of the standard Basque text (or texts) was 'similar' to that of the historical texts. As most of the latter are religious texts, we elected to use a digital version of the Bible written in standard Basque[9]. However, to determine whether the subject of the text is important when measuring distance between historical and standard Basque, we also relied on a non-religious second corpus written in standard Basque (EPEC, a reference corpus for the processing of Basque (Aduriz et al., 2006)).

In order to obtain the n-gram model of standard Basque (Bible or EPEC) we used the software previously mentioned in section 4. This software carries out a preprocessing step before obtaining the final 7-gram model that consists on: (1) text cleaning: figures and punctuation marks removed, uppercase letters converted to lowercase, extraneous characters eliminated, and so on; (2) tokenization. This preprocess reduced the size of the Bible and EPEC corpora, leaving them at 514,443 and 291,228 words, respectively. After preprocessing step, two n-gram models of standard Basque are trained, one utilising the Bible and the other using EPEC corpora.

The same cleaning process applied to the Bible and EPEC was repeated for each historical text (the resulting sizes of the corpora appear in Table 3). Given that the texts vary significantly in length, from 11,408 to 90,029 words, and because we wished to compare the distances between different dialects, we decided to conduct two experiments for each: one utilised a randomly selected predetermined portion of content similar in size to the shortest text (11,500), while the other used the full text. The results appear in Tables 2 and 3. In addition to cleaning process, n-grams of each historical text were also calculated. These n-grams were then used to compare with the previously obtained n-gram models of standard Basque in order to obtain perplexity-based distance.

6 Results and discussion

In this section we will present and analyse the results. Tables 2 and 3 contain the findings obtained in the two experiments previously described.

6.1 Results

Table 2[10] displays the results obtained for samples of similar size extracted from every source, while in Table 3 we see results for the complete text. As may be appreciated, the findings for the sample experiment differ little from those obtained from that done on the complete work, nor do they vary when we compare the historical texts with the Bible or with a corpus of a different subject matter, such as EPEC.

[8] https://github.com/gamallo/Perplexity

[9] https://www.biblija.net/biblija.cgi?l=eu

[10] In the following tables we will use abbreviations for dialects: Alavese=Al; Labourdin=L; Upper Navarrese=UN; Biscayan=B; Lower Navarrese=LN, and Souletin=S.

Cent.	Auth.	Dial.	Size	Dist. Bible	Dist. EPEC
XVI	Laz	Al	11,501	7.84	7.58
XVI	Lç	L	11,501	6.09	6.32
XVII	Mat	L	11,503	4.93	4.91
XVII	Ax	L	11,507	4.69	4.53
XVII	SP	L	11,510	4.77	4.79
XVII	Ber	UN	11,520	5.52	5.29
XVII	Cap	B	11,408	7.18	6.68
XVII	Tt	LN	11,503	6.07	5.68
XVII	Bp	S	11,500	11.33	9.48

Table 2: The two perplexity values for each historical text based on a portion of similar size. The first value was obtained using a contemporary version of the Bible written in standard Basque. The second was attained using the EPEC corpus.

Cent.	Auth.	Dial.	Size	Dist. Bible	Dist. EPEC
XVI	Laz	Al	12,072	7.83	7.58
XVI	Lç	L	73,906	6.13	6.28
XVII	Mat	L	16,323	4.94	4.90
XVII	Ax	L	90,029	4.72	4.57
XVII	SP	L	46,363	4.74	4.78
XVII	Ber	UN	14,995	5.58	5.31
XVII	Cap	B	11,408	7.18	6.68
XVII	Tt	LN	34,505	6.06	5.66
XVII	Bp	S	23,735	11.43	9.54

Table 3: The two perplexity values for each historical text. In this case, the entirety of each historical text was utilised in the experiment. The two corpora are the same as in the previous experiment.

Unsurprisingly, the works closest to the standard are those belonging to the central dialects: Labourdin, Upper Navarrese and Lower Navarrese. Moreover, the Labourdin texts are somewhat closer than their Navarrese counterparts, which was to be expected since the standard was essentially built on Guipuscoan and Labourdin, as mentioned above. Interestingly, Leizarraga's work, despite being essentialy written in Labourdin, departs somewhat from the standard. Leizarraga was presumably a native speaker of Lower Navarrese and he noted that he translated the Bible so that it would be understood by most readers, i.e. in a sort of northern koiné. We should also mention that Leizarraga's work is one of the oldest and it is therefore logical that it differs most from the standard. These factors may help explain why the results demonstrate a greater distance between this text and the standard. The case of Tartas's contribution may be similar, given that despite being a Souletin writer, he attempted to move away from the Souletin dialect in order to address a wider public.

The next most distant works from the standard are those by Lazarraga and Kapanaga, written in Alavese and Biscayan, respectively (both two western varieties). Once again, this result was expected. Because these varieties move away from the centre, they are more peripheral and, therefore, more distant from that standard variety.

Finally, the gap between the Souletin dialect and the standard should be highlighted. In view of the findings, we can clearly state that the work written in Souletin is the most distant from the standard. Yet again, this is an expected result since Souletin, like Biscayan, is a highly idiosyncratic peripheral variety. One of the most conspicuous features of this Souletin idiosyncrasy is the so-called sixth vowel "ü" (/y/), non-existing in the rest of dialects. We believe that it is this characteristic that makes the distance so quantitatively great. Tartas, however, opted against using a specific spelling for the sixth vowel in his works, or used it in a very defective way and perhaps this also helps explain why the distance is not as great as in other Souletin authors.

In short, we emphasize once more that the results obtained confirm our expectations and that they validate quantitatively what traditional dialectology affirms.

6.2 Comparing with other languages

Although the quantitative values of the distances are not directly comparable to similar experiments with corpora in other languages, we can consider whether the range of values that we have found (4.53 minimum and 11.43 maximum) coheres with those obtained in the diachronic study of other languages or in crosslingual comparisons.

Gamallo, Pichel, and Alegria (2017a) demonstrate that the distance between older English (sixteenth-eighteenth centuries) and today is 5.80, but that for previous centuries

(twelfth-fifteenth) it is up to 15.85 (original spelling in both cases). In the case of Portuguese, the measured values for the same periods in original spelling are 7.40 and 7.73, while for Spanish they are 5.97 and 8.02.

Thus, for the period between the sixteenth and eighteenth centuries, the values for the three languages are 5.80, 7.40 and 5.97. In our case, with the exception of Belapeire, the distances for most of the Basque historical texts are close to these figures.

With respect to distance between languages, Gamallo, Pichel, and Alegria (2017a) compute distances among 44 European languages using perplexity, yielding interesting figures that are close to the values we obtained:

- The smallest distances obtained are the Bosnian-Croatian distance (5) and the Portuguese-Galician distance (6).
- Within the 7-9 range are Bosnian-Slovene, Catalan-Spanish, Czech-Slovak and Portuguese-Spanish.
- The value for the Swedish-Danish distance is 12 and for Swedish-Norwegian 13.

Hence, the smallest diachronic distance between the Basque dialects and standard Basque is similar to the distance between the closest languages, just as the greater distances within Basque are similar to those between languages that are somewhat more differentiated.

7 Conclusions and Future Work

7.1 Conclusions

Utilising our Syntactically Annotated Historical Corpus in Basque and the previous work in language distance by Gamallo, Pichel, and Alegria (2017a), we have compared Basque historical corpora with current texts in the standard variety and calculated the language distances between them. Since the standard is based on the central dialects, the starting hypothesis is that the texts of the dialects of the extremes on the one hand, and the oldest on the other, will be the furthest away.

The results obtained have largely confirmed the thesis of traditional dialectology: peripheral dialects have a strong idiosyncrasy and are more distant from the rest. We have verified this by measuring the distance of all historical dialects from the standard Basque (built upon the central dialects).

We must not forget that these are initial experiments and that the findings, while significant, also raise further questions. For example, we hope to study in depth the effect of the spelling "ü" in Souletin texts, as it is involved in a series of morphophonological changes. In addition, another interesting avenue to pursue is the fact that Lazarraga's work is at the same distance from the standard as Kapanaga's written in the Biscayan dialect. Traditional dialectology tells us that Lazarraga's text is written in what is today known as the western dialect (as is Kapanaga's). But within that dialect, Lazarraga would correspond to a more eastern variety (closer to the central dialects) (Pagola, 2006). Therefore, one would expect that the distance with respect to the standard is not as great as in the case of the westernmost Biscayan.

We believe that this work opens up a new line of research in Basque dialectology and that the next step is to measure the distances of the historical dialects from each other to establish whether the results confirm this study's findings.

7.2 Future work

These first experiments and the results obtained encourage us to continue working along these lines. The next step will be to include all the works in the corpus in the experiments to see what occurs with varieties of the language that are not attested to until later, such as the case of the Guipuscoan dialect.

We would also plan to measure the distance between the different historical dialects, although we foresee that the scarcity of records will be a major roadblock. Once the distances between the historical dialects are calculated, it will be worthwhile to compile a contemporary corpus of the different dialects in order to measure the distances between them and compare the results with those obtained for the historical dialects. In this way, we will be able to test the thesis of traditional dialectology that the distance between Basque dialects increased over time. It nevertheless remains that, in some aspects, the spreading of standard Basque during the last decades favours dynamics of convergence between dialects.

Acknowledgments

We are very grateful to Pablo Gamallo of the University of Santiago de Compostela for his contributions to the development of the experiments. Special acknowledgment are due to José Ramón Pichel and Iñaki Alegría for their expertise in perplexity measure, to Ricardo Etxepare of the IKER UMR 5478-CNRS for his leadership in the BIM project and for always being committed to interdisciplinarity, and finally to Aritz Farwell for his help in revising the text.

This research has been partially supported by the Agence nationale de la recherche of France (ANR-17-CE27-572 0011-BIM); the Ministry of Science, Innovation, and Universities of Spain (RTI2018-573 098082-J-I00); and the Basque Government (IT1570-22).

References

Aduriz, I., M. J. Aranzabe, J. M. Arriola, A. Atutxa, A. D. de Ilarraza, N. Ezeiza, K. Gojenola, M. Oronoz, A. Soroa, and R. Urizar. 2006. Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for automatic processing. In *Corpus linguistics around the world*. Brill, pages 1–15.

Asgari, E. and M. R. K. Mofrad. 2016. Comparing fifty natural languages and twelve genetic languages using word embedding language divergence (WELD) as a quantitative measure of language distance. In *Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP*, pages 65–74, San Diego, California.

Aurrekoetxea, G. 1992. Nafarroako euskara: azterketa dialektometrikoa. *Uztaro*, 5:59–109.

Aurrekoetxea, G., I. Gaminde, J. L. Ormaetxea, and C. Videgain. 2019. *Euskalkien sailkapen berria*. UPV/EHU, Bilbao.

Aurrekoetxea, G. and C. Videgain. 2009. Le projet Bourciez: traitement géolinguistique d'un corpus dialectal de 1895. *Dialectologia*, 2:81–111.

Barrault, L., O. Bojar, M. R. Costa-Jussa, C. Federmann, M. Fishel, and Y. Graham. 2019. Findings of the 2019 conference on machine translation (WMT19). Association for Computational Linguistics (ACL).

Camino, I. 2008. Dialektologiaren alderdi kronologikoaz. *Fontes Linguae Vasconum (FLV)*, 108:209–247.

Chavula, C. and H. Suleman. 2020. Intercomprehension in retrieval: User perspectives on six related scarce resource languages. In *Proceedings of the 2020 conference on human information interaction and retrieval*, pages 263–272.

Claridge, C. 2009. Historical corpora. In *Corpus linguistics. An International Handbook*, pages 242–259, Berlin, Germany.

de Rijk, R. 1969. Is Basque an SOV language? *Fontes Linguae Vasconum (FLV)*, 1:319–351.

Degaetano-Ortlieb, S. and E. Teich. 2018. Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 22–33.

Estarrona, A., I. Etxeberria, R. Etxepare, M. Padilla-Moyano, and A. Soraluze. 2021. The first annotated corpus of historical basque. *Digital Scholarship in the Humanities*, 37(2):391–404.

Gamallo, P., I. Alegria, J. R. Pichel, and M. Agirrezabal. 2016. Comparing two basic methods for discriminating between similar languages and varieties. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 170–177.

Gamallo, P., J. R. Pichel, and I. Alegria. 2017a. From language identification to language distance. *Physica A: Statistical Mechanics and its Applications*, 484:152–162.

Gamallo, P., J. R. Pichel, and I. Alegria. 2017b. A perplexity-based method for similar languages discrimination. *VarDial 2017*, page 109.

Gamallo, P., J. R. Pichel, and I. Alegria. 2020. Measuring language distance of isolated European Languages. *Information*, 11(4):181.

Gao, Y., W. Liang, Y. Shi, and Q. Huang. 2014. Comparison of directed and

weighted co-occurrence networks of six languages. *Physica A: Statistical Mechanics and its Applications*, 393(C):579–589.

Gorrochategui, J., I. Igartua, and J. A. Lakarra. 2018. *Historia de la lengua vasca*.

Kondrak, G. 2005. N-gram similarity and distance. In *International symposium on string processing and information retrieval*, pages 115–126. Springer.

Lacombe, G. 1924. La langue basque. In *Les langues du monde*, pages 255–270, Paris.

Laka, I. 1996. *A brief grammar of Euskara, the Basque language*. UPV/EHU, Bilbao.

Lakarra, J. A. 1997. Euskararen historia eta filologia: arazo zahar, bide berri. *ASJU*, 31(2):447–535.

Liu, H. and J. Cong. 2013. Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin*, 58(10):1139–1144.

Michelena, L. 1964. *Textos Arcaicos Vascos*.

Mitxelena, K. 1981. Lengua común y dialectos vascos. *International Journal of Basque Linguistics and Philology*, 15:289–313.

Pagola, R. M. 2006. Lazarragaren eskuizkribua: grafiak, hotsak eta hitzak. In *Lingüística Vasco-Románica. I Jornadas = Euskal-Erromantze Linguistika. I. Jardunaldiak*, pages 539–561, Donostia.

Pichel, J. R., P. Gamallo, and I. Alegria. 2018. Measuring language distance among historical varieties using perplexity. Application to European Portuguese. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 145–155.

Pichel, J. R., P. Gamallo, and I. Alegria. 2020. Measuring diachronic language distance using perplexity: Application to English, Portuguese, and Spanish. *Natural Language Engineering*, 26(4):433–454.

Sant'Anna, A. A. and L. Weller. 2020. The threat of communism during the cold war: A constraint to income inequality? *Comparative Politics*, 52(3):359–393.

Sarasola, I. 1983. Contribución al estudio y edición de textos antiguos vascos. *ASJU*, pages 69–212.

Satrustegi, J. M. 1987. *Euskal Testu Zaharrak*.

Scherrer, Y., T. Samardžić, and E. Glaser. 2019. Digitising Swiss German: how to process and study a polycentric spoken language. *Language Resources and Evaluation*, 53(4):735–769.

Singh, A. K. and H. Surana. 2007. Can corpus based measures be used for comparative study of languages? In *Proceedings of ninth meeting of the ACL special interest group in computational morphology and phonology*, pages 40–47. Association for Computational Linguistics.

Séguy, J. 1973. La dialectométrie dans l'Atlas linguistique de la Gascogne. *Revue de Linguistique Romane (RLiR)*, 37:1–24.

Ulibarri, K. 2013. Testuak kokatuz dialektologia historikoan: egiteetatik metodologiara. In *Koldo Mitxelena Katedraren III. Biltzarra / III Congreso de la Cátedra Luis Michelena / 3rd Conference of the Luis Michelena Chair*, pages 511–532, Vitoria-Gasteiz.

Zuazo, K. 2014. *Euskalkiak*. Elkar.

Zugarini, A., M. Tiezzi, and M. Maggini. 2020. Vulgaris: Analysis of a corpus for middle-age varieties of italian language. *arXiv preprint arXiv:2010.05993*.

Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 63-71 recibido 15-12-2022 revisado 05-02-2023 aceptado 14-02-2023

Ajuste y evaluación del modelo DialoGPT sobre distintas colecciones de subtítulos de películas y series de televisión

Fine-tuning and evaluation of DialoGPT on several datasets of English movies and TV series subtitles

Raúl Giménez de Dios, Isabel Segura-Bedmar
Departamento de Informática, Universidad Carlos III de Madrid
raulgimenezdd@gmail.com, isegura@inf.uc3m.es

Resumen: Las nuevas plataformas de streaming han generado una proliferación de películas y series, la mayoría de ellas subtituladas. Esta proliferación proporciona una ingente cantidad de textos conversacionales, menos formales, más interactivos, que reflejan mejor la comunicación entre seres humanos. La mayoría de los modelos transformers desarrollados hasta la fecha no han sido entrenados con textos conversacionales. En este artículo, DialoGPT, un modelo GPT-2 entrenado para la tarea de diálogo sobre una colección de mensajes de Reddit, es re-entrenado y evaluado sobre distintas colecciones de subtítulos en inglés de series populares. Los experimentos muestran que DialoGPT es obtiene buenos resultados, y que el uso de los subtítulos y diálogos de películas y series es un excelente recurso para el desarrollo de chatbots.
Palabras clave: GPT-2, DialoGPT, Chatbot, Transformador.

Abstract: The new streaming platforms have generated a proliferation of movies and series, most of them subtitled. This provides a large number of conversational, less formal, more interactive texts that better reflect communication between human beings. Most of the transformative models developed to date have not been trained with conversational texts. In this article, DialoGPT, a GPT-2 model for the dialog task trained on a collection of Reddit posts, is fine-tuned and evaluated on different collections of English subtitles from popular movies and series. Experiments show that DialoGPT performs well and that English subtitles from movies and series can be an outstanding resource for chatbot development.
Keywords: GPT-2, DialoGPT, Chatbot, Transformer.

1 Introducción

Los asistentes conversacionales o chatbots son programas informáticos capaces de simular una conversación hablada o escrita, tal y como lo haría una persona (Adamopoulou y Moussiades, 2020). Durante los últimos años la creación de chatbots ha recibido gran interés tanto en la investigación científica como por parte de muchas empresas tecnológicas, consiguiendo desarrollar tecnologías cada vez más exitosas a la hora de imitar conversaciones entre seres humanos. Estas tecnologías ofrecen a los usuarios multitud de herramientas que facilitan su día a día (Fu et al., 2022), tales como Alexa y Siri, creados por Amazon y Apple, respectivamente. Estos chatbots son capaces de responder gran cantidad de cuestiones realizadas por los usuarios mediante un diálogo fluido e imitando el sistema de conversación humano con un aceptable nivel de calidad.

Los chatbots pueden ser tanto de dominio abierto o cerrado. Alexa o Siri son ejemplos de chatbots de dominio abierto porque son capaces de dar respuestas a preguntas planteadas por el usuario sobre diferentes temas, sin que la conversación esté focalizada en ningún dominio concreto. Por el contrario, los chatbots de dominio cerrado únicamente son capaces de responder cuestiones relacionadas con una temática o determinado campo de conocimiento, tal y como lo haría un técnico de atención al cliente o un apartado de preguntas frecuentes en una página web (Adamopoulou y Moussiades, 2020). Sin embargo, son poco eficaces cuando la conversación gira entorno a cualquier aspecto ajeno a su dominio.

ISSN 1135-5948 DOI 10.26342/2023-70-5

En los últimos años, la aparición de los modelos transformers ha revolucionado el campo de Procesamiento de Lenguaje Natural (PLN), logrando obtener los mejores resultados en muchas de sus aplicaciones (Wolf et al., 2020; Chernyavskiy, Ilvovsky, y Nakov, 2021). La mayoría de estos modelos (Devlin et al., 2019; Zhuang et al., 2021; Radford et al., 2019) han sido entrenados sobre grandes colecciones de texto escrito de fuentes como wikipedia, noticias, etc. Estos textos escritos posiblemente no sean capaces de representar correctamente las interacciones en un conversación humana. Recientemente, en 2022, esta situación ha comenzado a cambiar con la publicación de varios modelos, como DialoGPT (Zhang et al., 2020) o LamDA (Thoppilan et al., 2022), entrenados con textos conversaciones extraídos de redes sociales como Reddit. Más reciente aún ha sido la presentación de ChatGPT[1], por la empresa OpenAI en noviembre de 2022.

El objetivo de este artículo es utilizar el modelo DialoGPT, basado en GPT-2 y entrenado sobre una colección de 147M conversaciones extraídas de Reddit, para desarrollar un chatbot de dominio abierto. Además, el modelo será ajustado utilizando guiones y subtítulos de diferentes películas y series en inglés. Nuestra hipótesis inicial es que estas conversaciones son un excelente recurso para capturar las características del diálogo entre humanos, aún mejor que las conversaciones extraídas de una red social.

El artículo está organizado como sigue: la sección 2 revisa los trabajos más recientes en el desarrollo de asistentes conversacionales utilizando técnicas de PLN. En la sección 3, describimos las colecciones de subtítulos y el modelo DialoGPT que será re-entrenado sobre dichas colecciones. La sección 4 presenta y discute los resultados para cada una de las colecciones. Finalmente, las principales conclusiones y líneas de trabajo futuro serán descritas al detalle en la sección 5.

2 Estado de la cuestión

A continuación, presentamos los últimos avances que se han realizado en el desarrollo de chatbots basados en técnicas de PLN. Dhyani y Kumar (2021) desarrollaron un chatbot de dominio abierto, basado en una arquitectura bidireccional de red recurrente que utiliza mecanismos de atención para procesar textos más largos de forma correcta. Los autores utilizaron una colección de comentarios de usuarios de la plataforma Reddit[2], que fueron recogidos durante enero de 2015. Esta colección está formado por 3.027.254 instancias para el conjunto de entrenamiento y 5.100 para el conjunto de test. Cada instancia se compone de un comentario y la respuesta asociada al mismo. Las métricas empleadas para la evaluación del modelo generado tras el entrenamiento fueron BLEU (Papineni et al., 2002) y perplejidad (Adiwardana et al., 2020), obteniendo un 30,16 y 56,1, respectivamente.

El objetivo del trabajo presentado en (Konapur et al., 2021) fue el desarrollo de un chatbot capaz de mantener conversaciones con personas que se encuentran en una situación estresante, y responder de la misma forma que lo haría un terapeuta profesional. Para ello, los autores aplicaron distintos modelos que van desde redes de neuronas simples, redes convolucionales, varios tipos especiales de redes recurrentes como son las Long Short Term Memory (LSTM) y las Gated Recurrent Units (GRU). Además, los autores también aplicaron el primero de los modelos transformers (Vaswani et al., 2017a), que fue ajustado para la tarea de diálogo utilizando una colección de conversaciones sobre temas de salud mental entre pacientes y terapeutas, recopiladas de redes sociales como Reddit[3], Counsel chat[4] y Quora[5]. Los experimentos mostraron que el transformador obtenía una métrica BLEU de 33,5. El resto de los modelos fueron únicamente evaluados con la métrica accuracy, para medir con que precisión el modelo había generado la respuesta correcta. GRU fue el mejor modelo con una accuracy de 79 %, seguido por el modelo LSTM con una accuracy de 74 %. El modelo CNN y la red neuronal básica, obtienen peores resultados, 49 % y 56 %, respectivamente.

Adiwardana et al. (2020) presentaron Meena, un chatbot basado en un modelo generativo que fue entrenado utilizando conversaciones obtenidas de las redes sociales. En este trabajo, el objetivo es que el chatbot sea capaz de mantener conversaciones más largas y complejas con varios turnos. Para ello,

[1] https://openai.com/blog/chatgpt

[2] https://www.reddit.com

[3] https://www.reddit.com/r/mentalhealth

[4] https://counselchat.com

[5] https://www.quora.com/topic/Mental-Health

el chatbot necesita recordar el contexto durante la conversación y ser capaz de recordar qué información se ha recopilado en turnos anteriores. Así, el modelo fue entrenado considerando como contexto todas los turnos anteriores (hasta un máximo de siete), y como respuesta, el mensaje mostrado en el siguiente turno. En lugar de utilizar un modelo transformer, Meena está basado en "Evolved Transformer"(So, Le, y Liang, 2019), que utiliza técnicas "Neural architecture search (NAS)", para aprender de forma automática nuevas arquitecturas que sean más eficientes que las propuestas por los seres humanos. En concreto, utiliza un algoritmo evolutivo que copia la arquitectura original del modelo transformer, para encontrar una más óptima. El sistema obtuvo una puntuación de 10,2 en perflexity.

Patel et al. (2019) desarrollan un chatbot con la capacidad de detectar las emociones (felicidad, diversión, vergüenza, enfado, disgusto, tristeza, culpa y miedo) de los usuarios a través de los distintos turnos. En concreto, cada turno de un usuario es clasificado con un porcentaje de positividad o negatividad. Los autores evaluaron tres modelos distintos para la clasificación de estas emociones: una red convolucional (CNN), una red recurrente (RNN) y un tercer modelo que combinaba una red recuerrente y un modelo de atención. Los autores utilizaron la colección ISEAR (Satish y Punkit, 2016), formado por textos en inglés y anotados con las emociones anteriormente citadas. La evaluación mostró que CNN obtenía el mejor resultado con una accuracy de 75 %, mientras que los otros dos modelos únicamente obtenían una accuracy de 70 %.

Aunque todos los chatbots de dominio abierto han sido entrenados con conversaciones extraídas de redes sociales, es complicado dar una comparativa final porque estas colecciones son distintas. Además no siempre se han utilizado las mismas métricas de evaluación. Respecto a los enfoques utilizados, estos van desde arquitecturas de deep learning como las redes recurrentes, convolucionales y el primer modelo transformer propuesto por Vaswani et al. (2017a). Aunque dicho modelo fue ajustado con textos conversacionales de Reddit, el modelo base fue entrenado utilizando oraciones del dataset "WMT 14 English-German Sentence pairs"para la tarea de traducción automática.

3 Métodos

3.1 Datasets

Para entrenar nuestro modelo hemos utilizado distintas colecciones de subtítulos de series en inglés, que han sido obtenidos del portal Kaggle[6]. Estos datasets están formados por los diálogos entre los protagonistas en diferentes películas y series televisivas. Las colecciones seleccionados son los siguientes:

- Diálogos de la película "Pulp Fiction". Cada línea en la colección contiene un diálogo de la película. Se proporciona otra información como el número de palabras en la línea, el nombre del protagonista que dice la línea, el tiempo y lugar donde se dice el diálogo, entre otros. En nuestro caso, únicamente utilizaremos el texto del diálogo.
- Diálogos de aproximadamente 600 capítulos de la serie "The Simpson". Cada línea contiene el texto del diálogo y el protagonista que dice la línea, pero únicamente utilizaremos el texto. En total, contiene 131.551 líneas.
- Diálogos de la serie "Rick and Morty". Cada línea también contiene otra información como el número de temporada y capítulo, el nombre del episodio y el nombre del protagonista que está hablando. Sin embargo, esta información no será utilizada en nuestro sistema.
- Diálogos de la serie "Brooklyn-99". En este caso, además del texto, también se proporciona el nombre del protagonista que está hablando, pero este campo también será ignorado en nuestro sistema.
- Diálogos de la serie "The Office (US)". Otros datos como el número de temporada y episodio, título de episodio y protagonista aunque están disponibles serán omitidos por nuestro sistema.

Nuestro entorno de desarrollo para el entrenamiento y evaluación del modelo DialoGPT ha sido Google Colab, un servicio de Google que proporciona el uso gratuito de unidades GPU computacionales. Al ser gratuito, el servicio tiene ciertas limitaciones. Por ejemplo, no siempre existen unidades disponibles, no es posible ejecutar en segundo

[6] `https://www.kaggle.com/datasets`

plano y el tiempo máximo para ejecutar un proceso es de 12 horas. Por ese motivo, ha sido necesario limitar el tamaño de algunos datasets. La selección de las instancias ha sido aleatoria (aunque se ha asegurado que los diálogos fueran consecutivos) y se han eliminado todas las instancias nulas. Además, para cada dataset hemos generado dos subconjuntos con un ratio aproximado de 90:10 para entrenamiento y evaluación (ver tabla 1).

Dataset	Training	Test	Total
Pulp Fiction	1058	118	1.183
Brooklyn-99	981	110	1.098
Rick Morty	1708	190	1.905
The Office	8813	980	9.800
The Simpson	8813	980	9.800

Tabla 1: Conjuntos para el entrenamiento y evaluación del modelo.

Las colecciones han sido procesadas para que todos tengan el mismo formato: un campo denominado "respuesta", que corresponde al texto hablado por un protagonista, y un texto formado por los siete diálogos anteriores, que denominamos "contexto". Para tener un mayor conocimiento de estas colecciones, se han obtenido una serie de histogramas que muestran la distribución de tokens en cada diálogo. En estos histogramas, además se muestran otros valores estadísticos: la media de tokens por diálogo, la mediana, la desviación estándar y el percentil 90. Estos valores son útiles para definir algunos parámetros del modelo, como por ejemplo la longitud máxima de las secuencias de entrada. A continuación, se muestran los histogramas para cada uno de los colecciones que se corresponden con las figuras 1-5.

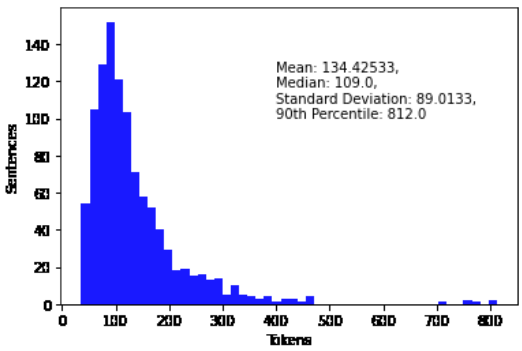


Figura 1: Histograma de la longitud de los textos (número de tokens) en el conjunto "Pulp Fiction".

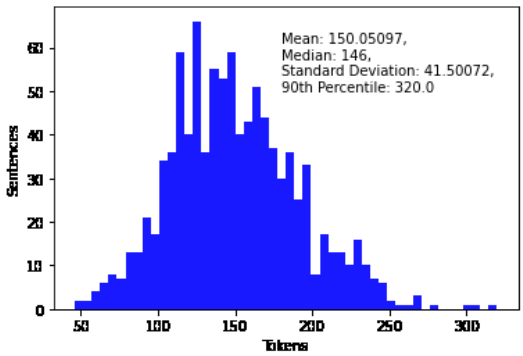


Figura 2: Histograma de la longitud de los textos (número de tokens) en el conjunto "Brooklyn-99".

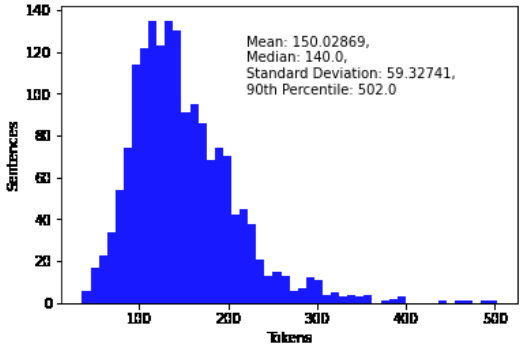


Figura 3: Histograma de la longitud de los textos (número de tokens) en el conjunto "Rick and Morty".

Se puede observar que en el conjunto "Pulp Fiction", la mayoría de las diálogos tienen una longitud máxima de 500 tokens, sin embargo, se observan ciertos casos atípicos que contienen una cantidad de tokens de entre 700 y 850. En el conjunto "Brooklyn-99", los diálogos tienen una longitud media de 150 tokens, siendo 400 tokens el tamaño del diálogo más largo. En el conjunto "Rick and Morty", practicamente el 100 % de los

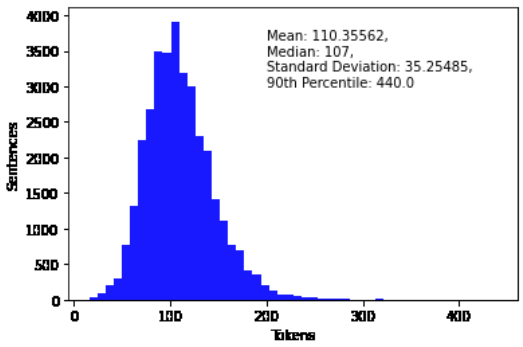


Figura 4: Histograma de la longitud de los textos (número de tokens) en el conjunto "The Simpson".

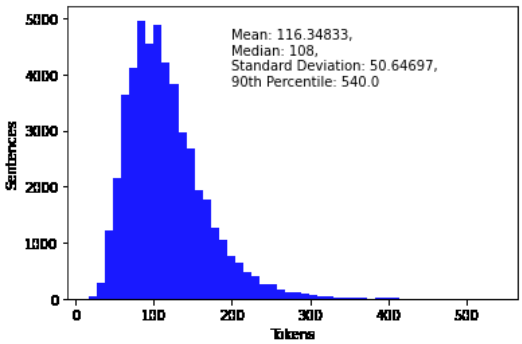


Figura 5: Histograma de la longitud de los textos (número de tokens) en el conjunto "The Office".

diálogos tiene una longitud máxima de 400 tokens, aunque existen algunos diálogos con una longitud mayor. Respecto al conjunto "The Simpson", la mayor parte de los diálogos tienen menos de 200 tokens. Por último, en el conjunto "The Office", la mayoría de los diálogos tienen una longitud máxima de 300 tokens. Para cada conjunto, se ha tomado el tamaño del percentil 90, como el tamaño máximo de la secuencia de entrada, que en ningún caso supera los 512 tokens, tamaño máximo para el modelo GPT-2.

3.2 DialoGPT

La aparición del modelo transformer, descrito en el artículo "Attention all you need"(Vaswani et al., 2017a), ha supuesto una auténtica revolución en el campo del aprendizaje profundo, consiguiendo mejores resultados que los obtenidos por los enfoques desarrollados hasta la fecha en la mayoría de las aplicaciones de PLN (Wolf et al., 2020; Chernyavskiy, Ilvovsky, y Nakov, 2021).

El principal avance del modelo transformer es el uso del mecanismo de atención, que fue diseñado para superar uno de los principales problemas de las redes recurrentes, el procesamiento de las oraciones largas, es decir, con un gran número de palabras. Las redes recurrentes procesan la información de forma secuencial, donde cada palabra está representada con un estado de la red. La información va pasando de un estado a otro, hasta alcanzar el último estado, que será el responsable de generar un vector capaz de codificar toda la información relevante en la oración. Sin embargo, si la oración está compuesta por muchas palabras, la información de sus primeras palabras podría perderse durante el procesamiento de las siguientes palabras. Cuanto mayor sea el número de palabras de la oración, mayor será la probabilidad de que la información relevante descrita en las primeras palabras no esté presente en el último estado (Lavanya y Sasikala, 2021).

El mecanismo de atención tiene en cuenta todos los estados intermedios para generar la salida. Así por ejemplo, en un chatbot, durante la fase de generación de la respuesta para una determinada interacción en el dialogo, el decodificador podrá acceder a todos los estados generados durante la fase de codificación del texto de entrada, y seleccionar aquellos que son más relevantes para generar un elemento específico de la salida. La implementación es bastante sencilla. Para cada elemento de la salida, el decodificador calculará la suma ponderada de los estados del codificador, asignando mayor pesos a los estados que sean más relevantes para el elemento de la salida actual. El mecanismo de atención únicamente está basado en estas sumas, que pueden ser ejecutadas en paralelo, obteniendo un modelo más eficiente. La paralelización es la segunda gran ventaja que ofrecen los modelos transformers respecto a las redes recurrentes (Vaswani et al., 2017a).

Tras la aparición de este primer modelo transformer (Vaswani et al., 2017a), otros modelos transformers han sido propuestos durante los últimos años, tales como BERT (Devlin et al., 2019) o GPT-2 (Radford et al., 2019). Aunque cada uno de ellos presentan características propias, todos se basan en el uso del mecanismo de atención. Además, estos nuevos modelos han sido entrenados como modelos de lenguaje a partir de grandes colecciones de textos. Las principales diferencias entre estos modelos principalmente recaen sobre el tipo de estrategia que utilizan para entrenar sus modelos de lenguaje. Así, mientras BERT utiliza las estrategias de enmascaramiento (donde el objetivo es predecir un token enmascarado en una oración) y predecir si dos oraciones son consecutivas, GPT-2 está basado en un modelo autoregresivo donde el objetivo es dada una secuencia de palabras, predecir la siguiente palabra. Además, BERT se puede considerar un modelo bidireccional, porque cuando predice un token, puede considerar todos los tokens en el contexto del token enmascarado, mientras que GPT-2 únicamente podrá utilizar los tokens anteriores, es decir, los tokens a la izquierda del token a predecir.

Otra importante diferencia entre BERT y GPT-2 se basa en su arquitectura. Mientras BERT únicamente está compuesto por varias capas de codificadores, encargadas de aprender una representación vectorial para cada palabra de una oración, GPT-2 consiste en un bloque de decodificadores, que aplicando un modelo autoregresivo se encargan de predecir el siguiente token que podría continuar a una secuencia de palabras de entrada. Así por ejemplo, dada la entrada "El médico recetó un ", GPT-2 podría generar la palabra "antibiótico", formando así la nueva oración "El médico recetó un antibiótico". Esta oración pasaría a ser la nueva entrada del modelo, y aplicando el mismo procedimiento iterativamente se consigue generar texto nuevo. GPT-2 y BERT, ambos modelos transformers, están basados en el mecanismo de atención, aunque en GPT-2, el mecanismo de atención (denominado "masked self-atention") es ligeramente distinto, ya que únicamente puede considerar los tokens anteriores al token que se están procesando en cada momento (Radford et al., 2019).

El conocimiento codificado en estos modelos de lenguaje puede ser transferido y utilizado para el desarrollo de aplicaciones concretas de PLN. A este proceso, donde un modelo de lenguaje es re-entrenado para una tarea y dataset específico, se denomina en inglés "fine-tuning"(Vrbančič y Podgorelec, 2020). Aunque tanto BERT como GPT-2 pueden ser adaptados para cualquier tarea de PLN, GPT-2 es una elección más apropiada en aplicaciones que implican la generación de nuevo texto como es el caso de los chatbots (Budzianowski y Vulić, 2019).

En 2021, Microsoft presentó un nuevo modelo, DialoGPT (Zhang et al., 2020), que fue específicamente creado para implementar un sistema de diálogo. Su principal ventaja es que es capaz de solventar problemas que presentaban enfoques previos en tareas de diálogo como la falta de consistencia y contextualización (Huang, Zhu, y Gao, 2020). El modelo original GPT-2 fue entrenado con 40GB de páginas webs, consiguiendo un vocabulario de 50.000 tokens. Al igual que BERT, el tamaño máximo de oración fue 512 tokens. A su vez, el modelo DialoGPT fue re-entrenado para la tarea de diálogo utilizando 147M conversaciones extraídas de Reddit. Cada instancia está formada por una secuencia de mensajes consecutivos, y el siguiente mensaje que se considera la respuesta que debería generar el modelo para la secuencia de entrada. Se realizaron algunas tareas de preprocesamiento para eliminar instancias que pueden generar ruido. Por ejemplo, se eliminaron todas las instancias cuya respuesta contenía urls, carácteres especiales como "[.º "]", o lenguaje tóxico (que fue detectado usando palabras claves). También se eliminaron las instancias donde la respuesta no contenía ninguna de las 50 palabras más comunes en inglés o tenía alguna palabra que se repetía más de tres veces. Además, no se consideraron las instancias cuyo texto de entrada y respuesta superaban las 200 palabras.

Como característica especial respecto a otras tareas, en la tarea de diálogo es necesario incorpora un nuevo token especial, "[end_of_turn]", que permita identificar el final de cada interacción en el diálogo. En cada instancia, las interacciones deberán estar separadas por este token. A su vez, el modelo también debe ser capaz de generar dicho token para marcar el final del mensaje de salida. La plataforma HuggingFace proporciona el modelo DialoGPT en tres versiones diferentes: small (117M), medium (345M), y Large (762M). Dichos modelos fueron entrenados con 5, 5, y 3 epochs respectivamente. Además, las interacciones con un número de tokens similar fueron agrupadas en el mismo batch para mejorar el training.

En este trabajo, la versión small del modelo DialoGPT ha vuelto a ser entrenado para colección descrita en el apartado anterior. En cada colección, cada instancia está formada por siete interacciones consecutivas en el diálogo, y una octava interacción, que se considera como respuesta, tal y como se describió en el apartado anterior. Con el objetivo de facilitar la replicabilidad de nuestra experimentación, nuestra implementación está disponible en un repositorio de GitHub[7].

4 Evaluación

Para evaluar el chatbot, hemos utilizado una de las métricas estándar para la evaluación de modelos de lenguaje, la perplejidad (Meister y Cotterell, 2021). Es una medida estadística de la confianza con la que un modelo de lenguaje predice un nuevo de texto. Podríamos definirlo como el grado de incertidumbre o duda que un modelo tienes respecto a si un

[7] `https://github.com/isegura/DialoGPTsepln`

texto es correcto o no. La perplejidad de un texto se calcula con la siguiente ecuación:

$$Perplejidad(W) = P(w_1, ..., w_n)^{-\frac{1}{n}} \quad (1)$$

donde W es una oración, N es el número de palabras, e w_i es la i-ésima palabra de la oración.

Si la perplejidad de un texto es bajo, significa que el modelo es capaz de generar dicho texto. El objetivo deseable es obtener modelos cuya complejidad media sobre una colección de textos sea baja, mientras que una perplejidad alta indicaría que el modelo no es capaz de predecir esos textos. El cálculo de la perplejidad es sencillo y rápido, lo que es especialmente ventajoso cuando estamos evaluando un modelo sobre una gran colección de textos. En nuestro caso, hemos utilizado la implementación proporcionada por HuggingFace para el cálculo de esta métrica, donde únicamente es necesario indica el modelo a utilizar y la colección de textos.

La tabla 2 muestra el grado de perplejidad del modelo DialoGPT (versión small) ajustado para cada una de las colecciones descritas en el apartado 3.1. Además de la perplejidad, también se incluye el error (Loss). Estos valores han sido obtenidos sobre los conjuntos test que fueron creados a partir de las colecciones (ver tabla 1).

Dataset	**Error**	**Perplejidad**
Pulp Fiction	1.09	2.97
Brooklyn-99	1.55	4.28
Rick Morty	1.35	3.84
The Office	1.13	3.11
The Simpson	1.6	4.95

Tabla 2: Resultados de DialoGPT.

A la vista de los resultados obtenidos es posible afirmar que DialoGPT obtiene los mejores resultados cuando es ajustado y evaluado utilizando los diálogos de la película "Pulp Fiction", ya que tanto su error como su perplejidad son los valores más bajos de la tabla. En contraposición, DialoGPT muestra los peores resultados para la colección "The Simpson", seguido muy cerca por "Brooklyn-99". Llama nuestra atención que el tamaño de las colecciones no parece tener un efecto directo sobre el grado perplejidad. Así por ejemplo, aunque "The Simpson.^{es} una de las dos colecciones mayores, su grado perplejidad ha sido el más alto, mientras que la perplejidad obtenida en "Pulp Fiction.^{es} la más baja, aunque dicha colección es una de las más pequeñas. Una posible razón es que los diálogos de esta película contengan menos ruido que los diálogos de "The Simpson".

Aunque no es posible compararnos con los trabajos del estado de la cuestión (ver apartado 2), porque los modelos han sido evaluados sobre colecciones distintas y se han empleado métricas diferentes, sí podemos afirmar que nuestro enfoque parece obtener perplejidades significativamente más bajas que la obtenida por la red bidireccional recurrente utilizada en el trabajo (Dhyani y Kumar, 2021), con perplejidad de 56.1 en una colección de conversaciones de Reddit. Del mismo modo, DialoGPT parece obtener mejores resultados que los obtenidos por el modelo Meena (Adiwardana et al., 2020), cuya perplejidad era 10.2 sobre una colección de textos conversacionales tomados de redes sociales.

Para considerar otra referencia para los valores de perplejidad, podemos considerar el primer modelo transformador (Vaswani et al., 2017b) cuya perplejidad fue 4.33, un valor similar a los obtenidos en en nuestra experimentación. Un reciente estudio (Ngo et al., 2021), mostraba que el modelo GPT-2 tenía una perplejidad de 74.7 evaluado sobre el corpus "One Billion Word Benchmark"(Chelba et al., 2014). Por tanto, podemos considerar que nuestros resultados de perplejidad son significativamente bajos, y que el modelo DialoGPT predice correctamente el texto.

Aunque desgraciadamente no ha sido posible abordar una evaluación basada en usuarios, la ejecución del chatbot nos ha permitido observar que este es capaz de generar texto con total sentido en base a la entrada del usuario, incluso manteniendo y teniendo en cuenta la información de turnos anteriores. El código para el entrenamiento y ejecución del chatbot está disponible en un repositorio de GitHub[8].

5 Conclusiones

La principal contribución de este trabajo ha sido utilizar el modelo DialoGPT para el desarrollo de un chatbot en el dominio abierto. Aunque dicho modelo ya fue entrenado con textos conversacionales obtenidos de re-

[8] `https://github.com/isegura/DialoGPTsepln`

des sociales, en nuestro trabajo, hemos ajustado y evaluado el modelo sobre colecciones de diálogos y subtítulos de distintas películas y series de televisión en inglés.

Nuestros resultados no son directamente comparables con trabajos anteriores porque se utilizan métricas y colecciones distintos, pero basándonos en los valores de perplejidad, podemos afirmar que nuestro modelo es capaz de generar correctamente el texto. La interacción con el chatbot también nos ha permitido comprobar que es capaz de mantener una conversación fluida y coherente, incluso siendo capaz de mantener y tener en cuenta el contexto descrito en los turnos anteriores. Por tanto, DialoGPT ajustado con diálogos y subtítulos de películas y series de televisión son un enfoque adecuado para el desarrollo de chatbots en el dominio abierto.

Entre las líneas de trabajo futuro, planeamos extender nuestro estudio a otros modelos transformers que fueron también preentrenados con textos conversacionales como Lamda y ChatGPT. Aunque hemos visto que los resultados no parecen estar directamente ligados con el tamaño del conjunto entrenamiento, se tratará de trabajar con mayores capacidades de cómputo para poder procesar el mayor conjunto de datos posible. En el actual trabajo se ha demostrado que los subtítulo y diálogos de películas y series de televisión son un buen recurso para la generación de chatbots. Otra de nuestras líneas futuras será aprovechar la existencia de estos subtítulos en diferentes idiomas para entrenar modelos multilingües y aplicarlos al desarrollo de chatbots. Extender nuestra evaluación a otras métricas y abordar una evaluación basada en usuarios son algunos de los retos que nos gustaría abordar en el futuro. Otra línea de trabajo será investigar para mitigar los posibles sesgos y estereotipos de género racismo, implícitos en los datos, y que podrían generar respuestas en los chatbots que sean ofensivas o poco éticas.

Agradecimientos

Esta publicación es parte del proyecto de I+D+i ACCESS2MEET (PID2020-116527RB-I0) financiado por AEI/10.13039/501100011033/.

Bibliografía

Adamopoulou, E. y L. Moussiades. 2020. An overview of chatbot technology. En *IFIP International Conference on Artificial Intelligence Applications and Innovations*, páginas 373–383. Springer.

Adiwardana, D., M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, y others. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Budzianowski, P. y I. Vulić. 2019. Hello, it's GPT-2 - how can I help you? towards the use of pretrained language models for task-oriented dialogue systems. En *Proceedings of the 3rd Workshop on Neural Generation and Translation*, páginas 15–22, Hong Kong, Noviembre. Association for Computational Linguistics.

Chelba, C., T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, y T. Robinson. 2014. One billion word benchmark for measuring progress in statistical language modeling. En *INTERSPEECH 2014*, páginas 2635–2639.

Chernyavskiy, A., D. Ilvovsky, y P. Nakov. 2021. Transformers:"the end of history" for natural language processing? En *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, páginas 677–693. Springer.

Devlin, J., M.-W. Chang, K. Lee, y K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. En *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)*, páginas 4171–4186, Minneapolis, Minnesota, Junio. Association for Computational Linguistics.

Dhyani, M. y R. Kumar. 2021. An intelligent chatbot using deep learning with bidirectional rnn and attention model. *Materials Today: Proceedings*, 34:817–824. 3rd International Conference on Science and Engineering in Materials.

Fu, T., S. Gao, X. Zhao, J. rong Wen, y R. Yan. 2022. Learning towards conversational ai: A survey. *AI Open*, 3:14–28.

Huang, M., X. Zhu, y J. Gao. 2020. Challenges in building intelligent open-domain

dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.

Konapur, S. P., T. Krishna, V. G, U. R, y S. H. 2021. Design of a chatbot for people under distress using transformer model. En *2021 2nd Global Conference for Advancement in Technology (GCAT)*, páginas 1–4.

Lavanya, P. y E. Sasikala. 2021. Deep learning techniques on text classification using natural language processing (nlp) in social healthcare network: A comprehensive survey. En *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, páginas 603–609. IEEE.

Meister, C. y R. Cotterell. 2021. Language model evaluation beyond perplexity. En *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, páginas 5328–5339, Online, Agosto. Association for Computational Linguistics.

Ngo, H., J. G. Araújo, J. Hui, y N. Frosst. 2021. No news is good news: A critique of the one billion word benchmark. En *35th Conference on Neural Information Processing Systems,*.

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.

Patel, F., R. Thakore, I. Nandwani, y S. K. Bharti. 2019. Combating depression in students using an intelligent chatbot: A cognitive behavioral therapy. En *2019 IEEE 16th India Council International Conference (INDICON)*, páginas 1–4.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, y others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Satish, T. y A. Punkit. 2016. Emotion detection in text.

So, D., Q. Le, y C. Liang. 2019. The evolved transformer. En *International Conference on Machine Learning*, páginas 5877–5886. PMLR.

Thoppilan, R., D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, y others. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, y I. Polosukhin. 2017a. Attention is all you need. *Advances in neural information processing systems*, 30.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, y I. Polosukhin. 2017b. Attention is all you need. En I. Guyon U. V. Luxburg S. Bengio H. Wallach R. Fergus S. Vishwanathan, y R. Garnett, editores, *Advances in Neural Information Processing Systems*, volumen 30. Curran Associates, Inc.

Vrbančič, G. y V. Podgorelec. 2020. Transfer learning with adaptive fine-tuning. *IEEE Access*, 8:196197–196211.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, y others. 2020. Transformers: State-of-the-art natural language processing. En *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, páginas 38–45.

Zhang, Y., S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, y B. Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. En *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, páginas 270–278, Online, Julio. Association for Computational Linguistics.

Zhuang, L., L. Wayne, S. Ya, y Z. Jun. 2021. A robustly optimized BERT pre-training approach with post-training. En *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, páginas 1218–1227, Huhhot, China, Agosto. Chinese Information Processing Society of China.

Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 73-83 recibido 14-12-2022 revisado 05-02-2023 aceptado 13-02-2023

On the Poor Robustness of Transformer Models in Cross-Language Humor Recognition

Sobre la Poca Robustez de los Modelos Transformers en el Reconocimiento Translingüístico del Humor

Roberto Labadie Tamayo[1] **Reynier Ortega-Bueno**[1]
Paolo Rosso[1] **Mariano Rodriguez Cisneros**[2]
[1]Universitat Politècnica de València
[2]Harbour.Space University, Barcelona
rlabtam@posgrado.upv.es, rortega@prhlt.upv.es, prosso@dsic.upv.es,
mjasonrc@gmail.com

Abstract: Humor is a pervasive communicative device; nevertheless, its portability from one language to another remains challenging for computer machines and even humans. In this work, we investigate the problem of humor recognition from a cross-language and cross-domain perspective, focusing on English and Spanish languages. To this aim, we rely on two strategies: the first is based on multilingual transformer models for exploiting the cross-language knowledge distilled by them, and the second introduces machine translation to learn and make predictions in a single language. Experiments showed that models struggle in front of the humor complexity when it is translated, effectively tracking a degradation in the humor perception when messages flow from one language to another. However, when multilingual models face a cross-language scenario, exclusive between the fine-tuning and evaluation data languages, humor translation helps to align the knowledge learned in fine-tuning phase. According to this, a mean increase of 11% in F1 score was observed when classifying English-written texts with models fine-tuned with a Spanish dataset. These results are encouraging and constitute the first step towards a computationally cross-language analysis of humor.
Keywords: humor recognition, humor translation, cross-language humor, multilingual models.

Resumen: El humor es un recurso comunicativo muy extendido; sin embargo, su portabilidad de un idioma a otro sigue siendo un reto para las máquinas informáticas e incluso para los humanos. En este trabajo, investigamos el problema del reconocimiento del humor desde una perspectiva translingüística y transdominio. Para ello, recurrimos a dos estrategias: la primera se basa en modelos transformers multilingües para explotar el conocimiento translingüístico que son capaces de destilar, y la segunda introduce la traducción automática para aprender y hacer predicciones en un solo idioma. Los experimentos demostraron que los modelos tienen dificultades ante la complejidad del humor cuando se traduce, lo que supone una degradación de la percepción del humor cuando los mensajes pasan de un idioma a otro. Sin embargo, cuando los modelos multilingües se enfrentan a un escenario translingüístico, exclusivo entre los idiomas de los datos de refinado y de evaluación, la traducción del humor ayuda a alinear los conocimientos aprendidos en la fase de refinado. En consecuencia, se observó un aumento medio del 11% de la puntuación F1 al clasificar textos escritos en inglés con modelos refinados con un conjunto de datos en español. Estos resultados son alentadores y constituyen el primer paso hacia un análisis computacional multilingüe del humor.
Palabras clave: detección de humor, traducción del humor, humor translingüe, modelos multilingües.

ISSN 1135-5948 DOI 10.26342/2023-70-6

1 Introduction

There is a set of evolved emotional functions shared by humans; laughter is part of this universal language of basic emotions that all humans recognize (Savage et al., 2017). Nevertheless, despite its ubiquity, proper comprehension of some humorous expressions goes beyond the semantics involved in messages. It relies on information from the context where jokes are made and the receptor's background knowledge (Tsakona, 2017), which implies a different or even null perception from one person to another. Moreover, when it comes to such creative device as humor, language plays a critical role in perceiving the funny meaning. Particularly, when information flows from one language to another on its way to the receptor, a joke's intended meaning is at risk of vanishing.

Wordplays are examples of language-dependent expressions that can be potentially misunderstood upon literal translation into a different language since they employ the arrangement and phonetics of words to produce humor. For example, in:
Why do male ants float while female ants sink? They're buoy-ant
It is very challenging to translate the phrase to ensure humor understanding by a non-English speaker, regardless of their background knowledge. Whereas, in the case of:

A: *Are you already here?*

B: *No, I'm just a figment of your imagination.*

The literal translation can still provoke laughter.

On the other hand, linguistic diversity on the Internet increases due to its interconnecting nature (Paolillo, 2007). In social media, where people from different cultural backgrounds and ethnicities share information, dealing with this multilingual phenomenon is inherent when identifying and filtering content and behaviors appropriated for specific users.

Many Natural Language Processing (NLP) tasks have been covered from a multilingual perspective in the scenario of social media with machine learning models (Ghanem et al., 2020; Wang et al., 2019; Al-Hassan and Al-Dossari, 2019). Most works tackle the under-representation of some languages by extending the knowledge learned from one language to another. In this sense, multilingual transformer-based architectures have become the state of the art in almost all of them (Wang et al., 2020; Chauhan et al., 2022). Despite the growing interest in humor in many languages such as English (Ermakova et al., 2022a; Meaney et al., 2021; Hossain et al., 2020), Spanish (Chiruzzo et al., 2021), Portuguese (Clemêncio, Alves, and Gonçalo Oliveira, 2019), and Chinese (Wu et al., 2021), to the best of our knowledge few efforts have been made to investigate the task of humor recognition from a computational cross-domain and cross-language perspective.

Machine translation paves the way for facing the challenge of multilingualism in texts[1]. Although these tools have been adequate for translating literal texts, when dealing with figurative language their performance drop considerably. In fact, humorous texts that often appeal to cultural knowledge or play on words become a complex problem in Machine Translation (Attardo, 2002; Zabalbeascoa, 2005; Popa, 2005; Low, 2011). Despite those shortcomings, we consider that some types of *self-contained* funny texts could preserve their meaning from one language to another. That is, the humor purely related to semantics without requiring additional cultural knowledge or information from the context. Moreover, we think that some linguistic features, not necessarily associated with the semantics and pragmatics involved in texts, may help to recognize humor without the need of understanding the text's whole meaning.

In light of the facts above, we consider that more efforts must be paid to investigate humor recognition in cross-domain and cross-languages scenarios. Particularly, we aim at stressing both Machine Translation systems and multilingual transformer-based models in order to identify their feasibility in humor recognition across languages. For that, we address the following research questions:

RQ1. What is the impact of machine translation on the semantics of humorous messages?

RQ2. How robust are multilingual trans-

[1]https://syncedreview.com/2020/05/20/neural-network-ai-is-the-future-of-the-translation-industry/

former models when dealing with translated humorous messages?

RQ3. Is it better to work with the same language of the dataset employed to fine-tune the multilingual transformer model for recognising humor (by automatically translating the target language)?

In RQ1 we aim at investigating how the semantics of humorous messages change upon automatic translation and how transformer models perceive this change. For RQ2, we will study the capability of multilingual transformer models to recognize specifically the presence of humor in translated messages. In RQ3, we are interested in investigating if in the case of multilingual transformers fine-tuned to recognize humor with English-written messages and evaluated on Spanish samples, it is better to automatically translate these samples into English. The latter, also for the opposite direction, when using multilingual transformers fine-tuned with Spanish-written messages and evaluated on English samples.

This paper presents a study on the behavior of transformer-based neural models for addressing humor recognition from a cross-language perspective. We take into account the *self-contained* and *language-dependent* humor phenomena, e.g. puns. Regarding the latter, we explore if the self-contained humor recognition methodology is extensible to its language-dependent kind. The rest of the paper is organized as follows: Section 2 presents some related works on humor recognition also from multilingual and cross-language perspectives. In Section 3, we describe the data and strategies studied as well as the employed methodology. In Section 4, we describe the experimental setup and results. Finally, in Section 5, we discuss the results achieved and provide some directions to explore as further work.

2 Related Works

Computational humor recognition is a widely explored issue. One of the first empirical pieces of evidence of this task's feasibility were given by (Mihalcea and Strapparava, 2005). From there on, several works have been conducted to integrate contextual, visual, and acoustic information in multimodal approaches (Yang, Ai, and Hirschberg, 2019; Vásquez and Aslan, 2021; Song et al., 2021; Chauhan et al., 2022; Tomás et al., 2022). Nevertheless, just a few works examine the phenomenon of humor from a cross-language and multilingual view (Chauhan et al., 2022).

Systems based on large pre-trained language models have outperformed the state of the art in many NLP tasks, including humor recognition (Grover and Goel, 2021; Subies, Sánchez, and Vaca, 2021) and machine translation (Vaswani et al., 2017).
However, humor translation remains a field with a huge room for improvement due to its subjectivity and linguistic complexity. Some of the most recent works (Miller, 2019) provide an interactive method for the computer-assisted translation of puns.

In this line, the task JOKER@CLEF 2022: Automatic Wordplay and Humour Translation Workshop (Ermakova et al., 2022b), where participants were asked to perform translation of humorous texts and identify its nature, was the first attempt to construct a parallel and multilingual humor corpus. Here, most participants' approaches relied again on transformers-based models, this time reinforced with templates featuring (Arroubat, 2022; Anne-Gwenn et al., 2022).

Besides the poor existence of parallel corpora, the above-referred issues in humor translation and recognition have worsened the scarcity of work transferring humor knowledge from one language to another.

3 Methodology

Evaluating the robustness of transformer models from a cross-language perspective for our task requires a parallel corpus with humorous information. Nevertheless, its absence forced us to compile a corpus considering different sources and languages. For this purpose, we focus on the English and Spanish languages, given the extensive amount of work related to humor recognition on them.

Taking into account the growing application of pre-trained transformers models in almost every NLP task, we employ three multilingual variants to evaluate their performance. However, as this work is not intended to outperform the SOTA in humor recognition, we simply stack a ReLU-activated layer between the encoder module and a softmax classification layer for each model. Then, the models are trained and evaluated with

an end-to-end fashion by feeding the ReLU-activated layer with the [CLS] vector from the last encoder block of the transformer (more details in Section 4). We employ multilingual models since we hypothesize they allow to capture and share background knowledge regardless of the language used during the fine-tuning process and the evaluation.

The first model, (*ml-base*) BERT-multilingual-base (Devlin et al., 2018), was pre-trained on the top 104 languages with the largest Wikipedia using a masked language modeling (MLM) and next sentence prediction (NSP) objectives. The second, (*ml-sentiment*), is a fine-tuned version of the latter in a sentiment analysis task on texts from six languages[2], among them English and Spanish, which are the ones addressed in this work. We use this model because although its pre-training knowledge comes from (*ml-base*), the information introduced by the sentiment-tuning could provide us with criteria diversity to characterize the general behavior of humor empirically. Finally, we study another variation of the BERT-base (*ml-distil*) model into a smaller and distilled architecture proposed by (Sanh et al., 2019), trained on the top 104 languages with the largest Wikipedia.

The source code and datasets employed in this study are publicly available in GitHub[3] for reproducibility.

3.1 Datasets

We gathered the monolingual datasets of 4 shared tasks:

(i) SemEval-2020 Task 7: Assessing Humor in Edited News Headlines (Hossain et al., 2020).

(ii) SemEval 2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense (Meaney et al., 2021).

(iii) HAHA at IberLEF 2021: Detecting, Rating and Analyzing Humor in Spanish (Chiruzzo et al., 2021).

(iv) JOKER@CLEF 2022 Task 1: Classify and Explain Instances of Wordplay (Ermakova et al., 2022a).

These datasets are annotated with several aspects related to humor, including whether it is present or not. However, they assess it with texts of different genres and writing styles, including tweets, headlines, or isolated wordplays, and representing different knowledge domains. This enables us to see two perspectives of the aggregation: the language-level, where datasets in the same language are grouped into a single corpus, and the domain-level, where each dataset, regardless of its language, is analyzed separately.

SemEval-2020 Task 7 Dataset

For this task, given a headline in the English language and a micro-edited version (*viz.*, replacement of entity by noun, noun by noun, verb by verb), participants were asked to determine whether this substitution generates a funny message. In the dataset (*Headlines*), for each headline, the replacement was annotated as well as the humor rating given by 6 annotators on a 0 to 3 scale and its mean value. From here, we consider as negative examples the original headline and as positive examples of humor those micro-editions whose mean humor rating was above 2 (*i.e.,* moderately funny and funny).

SemEval 2021 Task 7 Dataset

The dataset from this task (*Hahackathon*) contains English texts from Twitter and the Kaggle Short Jokes dataset, described with the presence of humor as well as humor rating, controversiality, and offensiveness rating of the messages by 20 different annotators. In this work, we only focus on the binary annotation regarding whether a text can be considered as funny.

IberLEF 2021 HAHA Dataset

In the shared task HAHA 2021, it was proposed a dataset (*HAHA*) composed of tweets written in Spanish, annotated regarding the presence of humor, funniness score, the humor mechanism employed (*e.g.,* parody, stereotype, etc.), and the humor target, *i.e.,* for a humorous tweet, the target of the joke from a set of classes such as racist jokes, sexist jokes, etc. We are interested in the binary annotation of humor, even when the remaining annotation is valuable for refining the humor analysis considering the language mechanism, the purpose, and the victims of

[2]https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

[3]https://github.com/labadier/Humor.git

the jokes.

JOKER@CLEF 2022 Task 1 Dataset
For this task, given a wordplay in the English language, participants were asked to classify it, attending different criteria. They also must identify and disambiguate the target words as an explanation of the wordplay. In the dataset (*JOKER*), the criteria annotated for each example include whether the source and the target of the wordplay co-occur in the text (horizontal/vertical), the manipulation type, *viz.* identity, similarity, permutation, abbreviation, if cultural reference is needed in order to understand the instance of wordplay, whether it is offensive or not, and whether the wordplay is in conventional form. Also, the target words and the disambiguation of the wordplay are annotated.

From these data examples, we just take those whose manipulation is by permutation (the textual material is given a new order, as in anagrams or spoonerism. e.g. "Dormitory = dirty room"), similarity (source and target are not perfectly identical, but the resemblance is obvious, e.g. "They're called lessons because they lessen from day to day") or Identity (source and target are formally identical, e.g. "How do you make a cat drink? Easy: put it in a liquidizer").

Table 1 shows the distribution of the examples in each dataset. The balance between positive (humor) and negative (non-humor) classes follows the one proposed by their authors. For training and testing the models from a cross-language perspective, we assume two partitions, one composed of examples originally in English from *SemEval 2021 Task 7 Dataset*, *JOKER@CLEF 2022 Task 1 Dataset* and *SemEval-2020 Task 7 Dataset*, with a representation of the positive class at training of 43%. The second is represented just by *IberLEF 2021 HAHA Dataset* dataset with a 39% of humorous examples for training. The instances in both partitions come from different knowledge domains and humor styles; then, besides the cross-language difficulty, we also have to deal with cross-domain data during evaluation.

To answer the first question, we first study how the humorous perception of these basic models varies for a back-translated instance. Thus, we check if the semantic underlying in the texts (Attardo, 2017) is preserved, even when their humorous incongruity vanishes in the pivot language during back-translation.

We rely on this experiment to disaggregate two sources of errors in the prediction stage when studying RQ2. The first is the one related to the learned parameters and model's architecture for recognizing humor. The second comes when instances are translated to the same language of the samples the multilingual transformer has been fine-tuned with.

Once we determine the impact of the noise introduced directly by machine translation, we can dive into RQ2. For this, we evaluate how translating texts impacts humor recognition in a cross-language scenario under multilingual models.

For evaluating each strategy, we use Micro-F1 over the positive class taking into account that at the domain-level, there are cases of slight unbalance or extreme scenarios where there are no examples for the non-humor class, as in *JOKER*.

4 Experimental Results

In the fine-tuning process of every model, we optimized the parameters with the RMSprop algorithm (Hinton, Srivastava, and Swersky, 2012) by employing an increasing learning rate from the shallower layers to the deeper ones (Howard and Ruder, 2018), starting from 1e-5 and increasing it on each layer with a factor of 0.1 units.

Every translation step involved in this work was accomplished with googletrans library, using Spanish as the complementary language for English, and vice versa. In the same way, when we study approaches based

Language	Dataset	Train ($\mathcal{T}$)		Test ($\mathcal{D}$)	
		Humor	Non-Humor	Humor	Non-Humor
English	*SemEval 2021 Task 7*	3436	5564	385	615
	JOKER@CLEF 2022 Task 1	531	0	4516	0
	SemEval-2020 Task 7 Dataset	890	890	88	88
Spanish	*IberLEF 2021 HAHA*	11595	18405	3000	3000

Table 1: Statistics of the datasets.

on back-translation, this complementarity relation is applied to select the pivot language.

As we mentioned before, the first step is to study how noisy machine translation results for humor regarding the semantics of the message; for this, we applied back-translation over the instances and investigated how humor perception vanishes for every multilingual model. In Table 2 are shown the results in terms of F1 for every model (detailed at *domain-level*), where $\mathcal{D}$ stands for the original version of our test sets described in Table 1, $\mathcal{D}^*$ is the version of $\mathcal{D}$ where each instance is translated into its complementary language and $\mathcal{D}^{**}$ corresponds to the back-translated version of $\mathcal{D}$. We include an estimation of the F1 95%-confidence interval (ci) by Percentile bootstrapping according to (DiCiccio and Efron, 1996).

Model	Dataset	$\mathcal{D}$	ci	$\mathcal{D}^{**}$
ml-base	*Hahack.*	0.921	0.015	0.923
	JOKER	0.941	0.005	0.939
	Headlines	0.778	0.062	0.772
	HAHA	0.870	0.008	0.869
ml-sent	*Hahack.*	0.914	0.017	0.916
	JOKER	0.934	0.005	0.933
	Headlines	0.814	0.050	0.802
	HAHA	0.871	0.008	0.870
ml-distil	*Hahack.*	0.905	0.018	0.903
	JOKER	0.945	0.005	0.944
	Headlines	0.709	0.070	0.716
	HAHA	0.863	0.009	0.861

Table 2: Variation in humor perception by multilingual transformer models after back-translation.

Here we can see that the error in $\mathcal{D}^{**}$ (a perturbed instance of $\mathcal{D}$) is not statistically significant w.r.t. the results on $\mathcal{D}$ if we assume the learned parameters of the models.

Since we only seek an empirical probe of the model's capability to find a similar interpretation of the back-translated data w.r.t. the original, for this experiment we train every multilingual model by employing all the domains and languages at the same time, allowing the knowledge-sharing among all the *domain-level* datasets.

However, we explored how this knowledge-sharing impacts the results for the cross-domain scenario present in the English *language-level* dataset. Table 3 shows different domain combinations for fine-tuning *ml-base* model, where K, H and J, refers to *Hahackathon*, *Headlines* and *JOKER* respectively[4].

Setting	Test Set		
	JOKER	Headlines	*Hahack.*
H	-	0.737	-
K	-	-	0.913
K+H	0.713	-	-
K+J	-	0.667	-
H+J	-	-	0.764
K+H+J	0.906	0.749	0.920

Table 3: Cross-domain settings for English datasets.

As we can see, using a purely cross-domain scenario (rows 3-5) has a negative impact on the model's performance. Nevertheless, when this external knowledge is used as a way of data augmentation (last row), it effectively helped to improve the achieved results. We can notice that in all cases, the results are inferior with respect to those obtained in Table 2, even when the fine-tuning is carried out across all the domains in the English language (last row). The latter suggests that the model employs knowledge from *HAHA* (Spanish corpus) to make inferences in English-written texts. Considering that, we investigated the effectiveness of using a multilingual system in a cross-language scenario by means of a zero-shot approach. We fine-tuned every model with the data from the English *language-level* dataset to evaluate the data from the Spanish *language-level* dataset and vice versa. Table 4 shows the results obtained in each case.

If we compare the results from Table 4 obtained using *ml-bert* with those from Table 3, we can observe that the model performance diminishes in each dataset. This suggests a greater contribution from the cross-domain knowledge for humor recognition with the studied transformer-based models.

Once we have studied the cross-language and cross-domain impact on humor recognition, we can explore how feasible it is to extend the knowledge by means of translation at the evaluation phase. Also, given the

[4]We were not able to evaluate the model trained on the JOKER dataset since it only consists of positive examples of humor.

Fine-tuning Language	Dataset	ml bert	ml sentiment	ml distil
Spanish	*Hahackathon*	0.760	0.753	0.754
	JOKER	0.666	0.661	0.650
	Headlines	0.534	0.528	0.500
English	*HAHA*	0.754	0.713	0.729

Table 4: Cross-language scenario results.

results of Table 2, where we found machine translation did not distort the semantics of funny texts, we are able to explore the issues of the humor recognition systems regardless of any possible *meaning changes* introduced by machine translation.

4.1 Humor Recognition in Translated Instances

As described in the strategy for study RQ3 in Section 1, we introduce a cross-language scenario again, but this time we tried to mitigate it by translating the evaluation instances into the fine-tuning language. Table 5 shows the results of the evaluation in these translated $\mathcal{D}^*$ datasets.

Dataset ($\mathcal{D}^*$)	ml bert	ml sentiment	ml distil
Hahackathon	0.808	0.825	0.787
JOKER	0.736	0.719	0.743
Headlines	0.553	0.554	0.512
HAHA	0.767	0.734	0.731

Table 5: Language inversion to reduce cross-language effect.

Here, we can see an improvement with respect to the previous results in Table 4, which means at least some of the humorous perception is preserved after translation and makes more useful the information learned during the model fine-tuning process.

The latter studies do not allow us to isolate the vanishing of humor recognition introduced when instances are translated. To this end, for evaluating the $\mathcal{D}^*$ dataset, we employed the same model parameters from the experiments referred to in Table 2, where cross-domain knowledge sharing was allowed.

Looking over the results from Table 6 with respect to $\mathcal{D}$ and $\mathcal{D}^{**}$ in Table 2, we can observe a poor robustness of the transformer models associated to humor translation. In

Model	Dataset	$\mathcal{D}^*$
ml-base	*Hahack.*	0.880
	JOKER	0.875
	Headlines	0.659
	HAHA	0.811
ml-sent	*Hahack.*	0.856
	JOKER	0.861
	Headlines	0.641
	HAHA	0.803
ml-distil	*Hahack.*	0.833
	JOKER	0.885
	Headlines	0.616
	HAHA	0.789

Table 6: Results for evaluation in translated instances.

the prediction phase, the models had in common issues associated with polysemous words, phrase ambiguities from the source language as regards the target language, and word rearrangements, particularly in wordplays. Table 7 shows examples from *Hahackathon* and *HAHA* related to this problem.

India is a very peaceful country because nobody has any beef over there. India es un país muy pacífico porque nadie tiene problemas allí.
Two dyslexics walk into a bra Dos disléxicos entran en un sostén
—¿Follamos? —No, que yo recuerde. —"Shall we fuck?" -Not that I remember.

Table 7: Translation ambiguities examples.

In the case of the *Headlines* dataset, which exhibits the greater drop in performance, it can be noticed that besides

the translation degeneration, examples are culturally dependent and related to knowledge and vocabulary distant from the one employed in the pre-training and fine-tuning phase of the evaluated models [5]. That is, HAHA vocabulary represents informal Twitter texts, and *Headlines* involves in some way "journalistic" and more formal vocabulary. Table 8 shows some examples related to the *Headlines* phenomenon.

Gov. Kasich slams President Trump's move on haircut care subsidies
White House spokesman does not rule out Trump-Putin July cuddling in Germany

Table 8: Contextual Dependency of HAHA translated examples.

Experiments developed in this section showed that humor translation helps the model to extend the knowledge learned in one language for inference in examples written in another one, i.e., it helps to mitigate the cross-language effect in some cases. Nevertheless, these models still struggle in front of the humor complexity as a communicative device when it is translated, effectively tracking a degeneration in the humor perception when messages flow from one language to another.

5 Conclusions and Future Work

Humor relies on the incongruences of two semantic planes that, when contrasted by the receptor, produce it in a natural way. Its translation comes with different implications that make pre-trained transformer-based models not robust to recognize it in a cross-language scenario. The main concerns are related to contextual information, background knowledge dependency, and lexical characteristics of the language (RQ2). This vanishing becomes more severe in creative ways of humor, such as wordplays involving phonetics, word polysemy, and phrasal ambiguity. Nevertheless, neural machine translation is capable of individually preserving the humorous semantics, as we examined in our work (RQ1). Also, despite the referred humor recognition vanishing, when we translate and evaluate the samples directly in the language of the models' fine-tuning process, they achieve better performance for recognizing humor in a cross-language scenario (RQ3).

As future work, we plan to extend this analysis towards a broader range of languages and translations provided by ready available machine translation systems to ensure reproducibility. Moreover, since almost every top-ranked system proposed in the shared tasks related to the explored datasets employed transformer-based architectures, we plan to evaluate their proposal on the experiments presented in this study as a way of obtaining more empirical evidence.

Finally, as cultural and contextual knowledge plays an important role in the performance decrease, we plan to explore two strategies. The first is to study how mitigating topic bias in datasets helps the model to address the cross-domain phenomenon. The second strategy consists of partially updating the knowledge of models by determining key examples as domain concepts from the new datasets and incorporating them when fine-tuning the model.

Acknowledgments

This work has been partially developed with the support of valgrAI - Valencian Graduate School and Research Network of Artificial Intelligence and the Generalitat Valenciana, and co-funded by the European Union. The work of Ortega Bueno and Rosso was in the framework of the FairTransNLP research project (PID2021-124361OB-C31) funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe.

References

Al-Hassan, A. and H. Al-Dossari. 2019. Detection of hate speech in social networks: a survey on multilingual corpus. In *6th International Conference on Computer Science and Information Technology*, volume 10, pages 10–5121.

Anne-Gwenn, B., E. Liana, D. de Saint-Cyr Florence, D. L. Pierre, C. Victor, P.-H. Nicolas, A. Benoit, A. Jean-Victor, D. Alexandre, G. Juliette, H. Aymeric, and M.-B. Florian. 2022. Poetic or humorous text generation: Jam event at

[5] In these cases models were fine-tuned with data originally in Spanish (*HAHA*).

pfia2022. In G. Faggioli, N. Ferro, A. Hanbury, and M. Potthast, editors, *Proceedings of the Working Notes of CLEF 2022: Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings.

Arroubat, H. 2022. Wordplay location and interpretation with deep learning methods. In G. Faggioli, N. Ferro, A. Hanbury, and M. Potthast, editors, *Proceedings of the Working Notes of CLEF 2022: Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings.

Attardo, S. 2002. Translation and Humour. *The Translator*, 8(2):173–194.

Attardo, S., 2017. *The General Theory of Verbal Humor*, chapter chapter10. Routledge.

Chauhan, D. S., G. V. Singh, A. Arora, A. Ekbal, and P. Bhattacharyya. 2022. A sentiment and emotion aware multimodal multiparty humor recognition in multilingual conversational setting. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6752–6761, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Chiruzzo, L., S. Castro, S. Góngora, A. Rosa, J. A. Meaney, and R. Mihalcea. 2021. Overview of haha at iberlef 2021: Detecting, rating and analyzing humor in spanish. *Procesamiento del Lenguaje Natural*, 67(0):257–268.

Clemêncio, A., A. Alves, and H. Gonçalo Oliveira. 2019. Recognizing humor in portuguese: First steps. In *Progress in Artificial Intelligence: 19th EPIA Conference on Artificial Intelligence, EPIA 2019, Vila Real, Portugal, September 3–6, 2019, Proceedings, Part II*, page 744–756, Berlin, Heidelberg. Springer-Verlag.

Devlin, J., M. Chang, K. Lee, and K. Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

DiCiccio, T. J. and B. Efron. 1996. Bootstrap confidence intervals. *Statistical science*, 11(3):189–228.

Ermakova, L., T. Miller, F. Regattin, A.-G. Bosser, C. Borg, É. Mathurin, G. Le Corre, S. Araújo, R. Hannachi, J. Boccou, A. Digue, A. Damoy, and B. Jeanjean. 2022a. Overview of joker@clef 2022: Automatic wordplay and humour translation workshop. In A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 447–469, Cham. Springer International Publishing.

Ermakova, L., T. Miller, F. Regattin, A.-G. Bosser, C. Borg, É. Mathurin, G. Le Corre, S. Araújo, R. Hannachi, J. Boccou, et al. 2022b. Overview of joker@ clef 2022: Automatic wordplay and humour translation workshop. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 447–469. Springer.

Ghanem, B., J. Karoui, F. Benamara, P. Rosso, and V. Moriceau. 2020. Irony detection in a multilingual context. In *European Conference on Information Retrieval*, pages 141–149. Springer.

Grover, K. and T. Goel. 2021. Haha@iberlef2021: Humor analysis using ensembles of simple transformers. In M. Montes, P. Rosso, J. Gonzalo, M. E. Aragón, R. Agerri, M. Á. Á. Carmona, E. A. Mellado, J. Carrillo-de-Albornoz, L. Chiruzzo, L. A. de Freitas, H. Gómez-Adorno, Y. Gutiérrez, S. M. J. Zafra, S. Lima, F. M. P. del Arco, and M. Taulé, editors, *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021*, volume 2943 of *CEUR Workshop Proceedings*, pages 883–890. CEUR-WS.org.

Hinton, G., N. Srivastava, and K. Swersky. 2012. Lecture 6a overview of mini–batch gradient descent. *Coursera Lecture slides https://class. coursera. org/neuralnets-2012-001/lecture,[Online.*

Hossain, N., J. Krumm, M. Gamon, and

H. Kautz. 2020. SemEval-2020 task 7: Assessing humor in edited news headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 746–758, Barcelona (online), December. International Committee for Computational Linguistics.

Howard, J. and S. Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July. Association for Computational Linguistics.

Low, P. A. 2011. Translating jokes and puns. *Perspectives*, 19(1):59–70.

Meaney, J. A., S. Wilson, L. Chiruzzo, A. Lopez, and W. Magdy. 2021. SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119, Online, August. Association for Computational Linguistics.

Mihalcea, R. and C. Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Miller, T. 2019. The punster's amanuensis: The proper place of humans and machines in the translation of wordplay. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 57–65, Varna, Bulgaria, September. Incoma Ltd., Shoumen, Bulgaria.

Paolillo, J. C. 2007. How Much Multilingualism?: Language Diversity on the Internet. In *The Multilingual Internet: Language, Culture, and Communication Online*. Oxford University Press, 05.

Popa, D.-E. 2005. Jokes and translation. *Perspectives: Studies in Translatology*, 13(1):48–57.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS 2019*.

Savage, B. M., H. L. Lujan, R. R. Thipparthi, and S. E. DiCarlo. 2017. Humor, laughter, learning, and health! a brief review. *Advances in physiology education*.

Song, K., K. M. Williams, D. L. Schallert, and A. A. Pruitt. 2021. Humor in multimodal language use: Students' response to a dialogic, social-networking online assignment. *Linguistics and Education*, 63:100903.

Subies, G. G., D. B. Sánchez, and A. Vaca. 2021. BERT and SHAP for humor analysis based on human annotation. In M. Montes, P. Rosso, J. Gonzalo, M. E. Aragón, R. Agerri, M. Á. Á. Carmona, E. Á. Mellado, J. Carrillo-de-Albornoz, L. Chiruzzo, L. A. de Freitas, H. Gómez-Adorno, Y. Gutiérrez, S. M. J. Zafra, S. Lima, F. M. P. del Arco, and M. Taulé, editors, *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021*, volume 2943 of *CEUR Workshop Proceedings*, pages 821–828. CEUR-WS.org.

Tomás, D., R. Ortega-Bueno, G. Zhang, P. Rosso, and R. Schifanella. 2022. Transformer-based models for multimodal irony detection. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–12.

Tsakona, V. 2017. Genres of humor. In *The Routledge handbook of language and humor*. Routledge, pages 489–503.

Vásquez, C. and E. Aslan. 2021. "cats be outside, how about meow": Multimodal humor and creativity in an internet meme. *Journal of Pragmatics*, 171:101–117.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, M., H. Yang, Y. Qin, S. Sun, and Y. Deng. 2020. Unified humor detection based on sentence-pair augmentation and transfer learning. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 53–59, Lisboa, Portugal, November. European Association for Machine Translation.

Wang, Z., S. Hale, D. I. Adelani, P. Grabowicz, T. Hartman, F. Flöck, and D. Jurgens. 2019. Demographic inference and representative population estimates from multilingual social media data. In *The World Wide Web Conference*, pages 2056–2067.

Wu, J., H. Lin, L. Yang, and B. Xu. 2021. Mumor: A multimodal dataset for humor detection in conversations. In *Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part I*, page 619–627, Berlin, Heidelberg. Springer-Verlag.

Yang, Z., L. Ai, and J. Hirschberg. 2019. Multimodal indicators of humor in videos. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 538–543.

Zabalbeascoa, P. 2005. Humor and translation - an interdiscipline. *Humor-International Journal of Humor Research*, 18(2):185–207.

Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 85-98 recibido 14-12-2022 revisado 03-02-2023 aceptado 13-02-2023

When humour hurts: linguistic features to foster explainability

Cuando el humor duele: características linguísticas para ganar en explicabilidad

Lucía I. Merlo[1], Berta Chulvi[1,2], Reynier Ortega[1], Paolo Rosso[1]

[1]Universitat Politècnica de València, Spain
[2]Universitat de València, Spain
lumer1@inf.upv.es, berta.chulvi@upv.es, prosso@dsic.upv.es, rortega@prhlt.upv.es

Abstract: The main objective of this research is to use different features for the textual representation of humorous texts and detect which are the characteristics that distinguish non-offensive jokes from the highly offensive ones. For this purpose, we use the data from the HaHackaton task in which jokes are annotated according to their degree of offensiveness. A new classification task is created by using two subsets of the jokes: the non-offensive ones and the highly offensive ones.The features with statistically significant differences in the two groups are used. By applying an ablation test, the most relevant features are used for a second classification task, showing that it is possible to obtain the same results with fewer computational resources.
Keywords: Humour, offensive language, computational lingüistics.

Resumen: El objetivo de esta investigación es utilizar distintas características para representar los textos humorísticos y detectar cuáles son las que mejor distinguen los chistes no ofensivos de los muy ofensivos. Se utiliza los datos de la tarea HaHackaton en la que los chistes están anotados según su grado de ofensa. Se diseña un nuevo problema de clasificación con dos conjuntos de chistes: los nada ofensivos y los muy ofensivos. Los clasificadores se entrenaron con las características que presentan diferencias significativas en las dos clases. Mediante la aplicación de un *ablation test* se identificaron las más relevantes que se han utilizado en una segunda tarea de clasificación mostrando que es posible obtener los mismos resultados con menos recursos computacionales.
Palabras clave: Humor, lenguaje ofensivo, lingüística computacional.

1 Introduction

When a society begins to overcome its prejudices, humour is one of the spaces in which these prejudices remain longer. As the social psychologist Michael Billig stands in his research about humour: if the collective laughter has a shameful, darker side, then, there is a lot that we may wish to hide from ourselves (Billig, 2005). This argument builds upon the insights of Bergson (Bergson, 1900) and Freud (Freud, 1960) who suggest that humour -and mainly the part of humour which serves to ridicule- ensures that members of society routinely comply with the customs and habits of their social milieu to avoid being the objects of jokes.

Certainly, humour has many facets and multiple effects on a social and personal life (Martin and Ford, 2018). It can be rebellious, kicking against the dictates of social norms and defending minority identities (Dobai and Hopkins, 2020). Also, it is well known that humour has an important beneficial function for personal life (Ripoll and Casado, 2010). But sometimes, certain type of humour is more than a simple joke. It has important consequences for some minority groups at personal and at societal level (Ford et al., 2008). For this reason, this research aims to detect how is the language that conveys offense in humour as a first step towards understanding how humour is an effective language device by means of prejudice and stereotypes can be maintained and perpetuated.

The prejudice norm theory (Ford and Ferguson, 2004) stands that *disparagement humour* function as a source of self-regulation for people high in prejudice because it creates

ISSN 1135-5948 DOI 10.26342/2023-70-7

a normative climate of tolerance of discrimination. This could be the reason why offense towards certain groups is well canalised through humour. The effects of these offensive jokes spill over into other spaces with far more serious consequences. For example, research about sexism has demonstrated that for men high in hostile sexism, sexist humour can have important social consequences, for example on rape proclivity (Romero-Sánchez et al., 2017).

Humour as a way to offend is not limited to intergroups relation, but it is also used at a interpersonal level. This kind of humour has been defined as *adversarial humour* (Veale, Feyaerts, and Brône, 2006). This humour occurs when up to a certain point, jokes have the underlying goal of weaken the opponent's position in a given social interaction.

Jokes are also a cultural product very sensitive to the passage of time. Something amusing twenty years ago, nowadays might be considered boring, aggressive or even hateful. Detecting offense in humour is a complex matter (Merlo, 2022). A joke can have abusive language but not being hurtful and the opposite: it can be hurtful without being explicitly abusive (Yin and Zubiaga, 2022). If detecting when a joke is hurtful is complex it is even more difficult to explain the results obtained in a classification task on the basis of the linguistic characteristics of the texts.

1.1 Objectives and research questions

Nowadays, social media platforms are widely extended all over the world and are often used to express hate speech camouflaged into jokes, trying to hide underlying negative attitudes. For hate speech monitoring activities, it is crucial to distinguish between offensive and non-offensive humour. This distinction is also relevant when analysing the communicative climate in a given community. We are conscious that, deep learning models achieve very impressive results in many NLP tasks in terms of effectiveness (Grover and Goel, 2021; Song et al., 2021; Potamias, Siolas, and Stafylopatis, 2020; González, Hurtado, and Pla, 2020), but often they may be quite complex from an explainability point of view. Then, our objective is to identify linguistic patterns present in hurtful humour that could help to automatically recognise these types of communication. This characterization of the language used in offensive humour could serve to gain on explainability when deep learning models are applied in humour recognition tasks. With this aim this work addresses four research questions:

- RQ1. Which are the features that distinguish non-offensive humour from the offensive ones?
- RQ2. How do three standard machine learning classifiers perform using these linguistic features?
- RQ3. Which of these linguistics features contribute more to the classification task?
- RQ4. Is it possible to obtain similar or better results employing only the most relevant linguistic features?

2 Related work

One of the first researches on humour recognition considering linguistic features was presented by Mihalcea and Strappavara (Mihalcea and Strapparava, 2005). The authors carried out a study for distinguishing humorous and non-humorous texts, using a computational approach for humour recognition. Furthermore, humorous examples consisted in one-liners while non-humorous texts were extracted from three resources: Reuters news headlines, proverbs and texts from British National Corpus (BNC). In English context, one-liner is an idiom to refer a short joke or witty remarks. Through classification systems, it was possible for them to detect which linguistic features were relevant. Specifically, classifiers were trained with stylistic features (alliterations, antonyms and slangs), content features and a combination of both. The results showed that stylistic markers help to distinguish a large number of one-liners jokes from Reuters news headlines and from BNC's texts, but not from proverbs. The authors suggest that content features help to differentiate jokes and proverbs although their stylistic similarity, but do not help to distinguish jokes from Reuters news headlines and BNC's texts. They remark on how humorous data mainly include words that refer to human scenarios (man, woman, I, you, person) and negative forms of words (isn't, doesn't, bad).

Sjöbergh and Araki tried to determine whether a text is a joke without considering the meaning of it (Sjöbergh and Araki,

2007). They used a corpus of 6,100 one-liner jokes and phrases from the British National Corpus for non-humorous examples. The features considered were text similarity (word overlap between the training instances and the text to classify, applying a novel weighting scheme), most common words within jokes (e.g. animals are particularly frequent), measure of ambiguity in a phrase, stylistic features (rimes, repeated words, use of you/I/he/she, negations) and idiomatic expressions (e.g. It's a piece of cake). The obtained results yielded that common words in jokes seemed to be the most useful feature for humour distinction, whereas stylistic features did not seem to provide a substantial contribution. Despite this, the article discusses how humorous texts differ from others without recognizing the meaning, although the features extracted from content markers were considered as highly relevant.

A weakness of the research mentioned so far, is that the humorous and the non-humorous texts come from quite different sources, e.g. one-liners vs sentences from British National Corpus. These sources present significant differences between them, regarding topic, vocabulary and target audience. Trying to overcome this issue (Reyes et al., 2010) studied a corpus of online comments retrieved from the Slashdot news website. The authors used a selection of 600,000 comments annotated by users into four categories: funny, informative, insightful and negative. The classification models were trained with linguistic features related to sexual content, semantic ambiguity, polarity, emotions, slang and emojis. By computing a multi-label classification, the authors examined which of the features contributed the most to humour recognition. They observed that the distinction between funny and informative categories was more challenging than the differentiation between funny and insightful and funny and negative ones. Regarding the features, slangs terms and emojis helped to improve humour recognition.

On the other hand, tasks related to humour recognition have attracted many researchers to the field. In English we can find the HashtagWars task in SemEval-2017 (Potash, Romanov, and Rumshisky, 2017) and in SemEval-2020 (Hossain et al., 2020) related with humor in headlines. In SemeEval-2021, the HaHackathon task (Meaney et al., 2021) proposed to distinguish between humorous and non-humorous texts while including several subtasks. The second task mentioned, also set as a subtask the prediction of the rate of offense in texts as we explain in detail in Section 3. In Spanish we find the HAHA task in 2018 (Castro, Chiruzzo, and Rosa, 2018), in 2019 (Chiruzzo et al., 2019) and in 2021 (Chiruzzo et al., 2021). All these tasks proposed a principal task of humour recognition and different subtasks. Furthermore, in the 2021 edition of HAHA, the organisers proposed to predict a funniness score value for each tweet, the mechanism by which the tweet conveys humour belongs to a set of classes (irony, wordplay, hyperbole, or shock) and the content of which the joke is based on, with the main target related to racist jokes, sexist jokes, dark humor, dirty jokes, among others until fifteen categories.

In these evaluation forums, the top-ranking teams made extensive use of pre-trained language models such as BERT, ERNIE (Zhang et al., 2019), ALBERT (Lan et al., 2020) or RoBERTa (Zhuang et al., 2021). These approaches had an excellent performance in accuracy. Still, they cannot distil linguistic knowledge valuable for understanding how language devices (particularly humour) convey offensiveness, stereotypes and prejudice. As a consequence, we do not have an overly recent knowledge about which linguistic features are the most important ones to distinguish offensive humour from non-offensive humour. Moreover, recent works (Ortega-Bueno, Rosso, and Medina Pagola, 2022; Frenda et al., 2022; Cignarella et al., 2020) have shown that reinforcing the deep learning models with linguistic knowledge helps to improve their overall performance. As a result, the aim of the following experiments is to provide some insights in this direction.

3 Data and preprocessing

3.1 Data

For this research we used the dataset of the *HaHackaton* task from *HaHackaton, Detecting and Rating Humor and Offense* organised at SemEval-2021 (Meaney et al., 2021). In the original dataset 80% of texts are originated in Twitter and unsettled 20% is obtained from the Kaggle *Short Jokes* dataset (Moudgil, 2017). Some keywords referring of-

fense to certain groups were used in the data collection strategy. Complete examples of offensive keywords and jokes with them can be found in appendix A in Tables 11 and 12. A total of 10,000 texts compose the original dataset of the *HaHackaton* task. Text annotation was done by US citizenship participants belonging to the following age groups: 18-25, 26-40, 41-55, 56-70. Each text was annotated by 5 members of each group. The task organisers instructed annotators to indicate if the text has the intention to be humorous in a 1 to 5 scale. As a second question they asked if the text was generally offensive in a scale of 1 to 5. They instructed annotators to consider as generally offensive a text which targets a person or group of people, simply for belonging to a certain group or a text that a large number of people were likely to be offended by. The offense rating of each text is the average of all ratings given by the annotators, including 'no offense' as 0.

Our research makes use of "offense rating" annotation to create a new classification task into two new categories: the non-offensive humour vs the most offensive humour. In order to create these two new datasets, we used the offense rating of each joke in the original dataset (0-5) and we created four groups of texts, each one corresponding to a quartile of the offense score variable. For our analysis and for the classification experiments, only the outermost groups are used. Therefore, our dataset is composed by the first quartile (the non-offensive jokes) and the fourth quartile (the highly offensive jokes) of the original dataset of the *HaHackaton* task. Specifically, the non-offensive set has 1,601 instances and the highly offensive set is composed by 1,504 examples. To answer the first three research questions in Section 4 and in Section 5 we use only the training set (1,253 non-offensive and 1,210 highly offensive instances) of the *HaHackaton* dataset. We keep the test set of *HaHackaton* (348 non-offensive and 294 highly offensive) to answer RQ4 in Section 7.

3.2 Checking the manual annotation

We applied several exploratory strategies to evaluate the quality of the manual annotation of the dataset. Firstly, the Spearman correlation between offense rating and humour rating scores over humorous data has been calculated in the two sets of jokes. With a ρ of -0.27 and a p-value $< .001$, we observe that the annotators tend to consider a text with greater amounts of humour if the level of offense in it is low or null.

As a second strategy to evaluate the quality of annotations regarding offense rating variable, we proceed in two steps. The first one consists in computing features from several linguistic resources, for instance: *SenticNet* (Cambria et al., 2016), *Textblob* (Loria and et.al, 2020), *SentiWordNet* (Baccianella, Esuli, and Sebastiani, 2010), *VADER* (Hutto and Gilbert, 2014), *ANEW* (Warriner, Kuperman, and Brysbaert, 2013) and *AFINN* (Nielsen, 2011). The second step consists of calculating either the Mann-Whitney U test or the Wilcoxon Signed-Ranked test, attending whether the observations are paired or not, over the quantitative features, taking as independent variable the offense group (non-offensive jokes vs highly offensive jokes). The complete results can be found in Appendix A in Tables 13 and 14. In summary, it can be seen that we find statistical differences between the two classes of humour in sentiment score, values, polarity, abusive language and subjectivity using the above-mentioned resources.

3.3 Text representation

Linguistic feature extraction conforms the core of this analysis. Hence, vectorized representation of features are achieved it through the *Stanza tool* (Peng Qi and Manning., 2020) and lexicons: *Binary Lexicon of abusive words* (Wiegand et al., 2018), *Hurtlex* (Bassignana, Basile, and Patti, 2018), *EmoSenticNet* (Bandyopadhyay et al., 2013), *SentiSense* (de Albornoz, Plaza, and Gervás, 2012) and *LIWC* (Tausczik and Pennebaker, 2010).

To extract part-of-speech tags, syntactic & morphological information, the *Stanza tagger* for English is used. Each term is assigned to a tag (noun, pronoun, adjective, tenses, 1st/2nd/3rd persons). The information regarding punctuation symbols is also computed by the *Stanza tagger*, by applying it over the original texts.

Variables related to affective and content information are constructed from lexical resources. The feature extraction procedure is equal for both of them. Tokens within tweets, are compared to the list of terms contained

in each one of the lexical resources used. Afterwards, we computed the number of times each word of the terms-list appear within the document. The *LIWC* resource also enables to extract syntactic & morphological markers, besides the affective and the content ones. Finally, the features are obtained by dividing the frequency of terms found in the tweet over the tweet length in terms of number of words. As a result, texts are represented as a frequency weighted term vector. Hence, each i-value of the linguistic feature corresponds to the rate of occurrence of determined category inside the i-tweet.

4 On offensive humour attributes

4.1 Statistical analysis

To select the most suitable features to represent the texts in the experimental phase we decided to identify the ones in which the offense label (non-offensive vs highly offensive) introduces statistically significant differences between the distributions of quantitative data.

Firstly, the Spearman correlation has been computed in order to determine whether or not values of the same feature from each class are independent. If the null hypothesis is true, observations are not paired and the Mann-Whitney U test is used. Rejecting the null hypothesis means that observations are paired and the Wilcoxon Signed-Ranked test is computed. This analysis was carried out by considering a p-value with a significance of 0.05.

The features included in the next section are those with a statistically significant difference with a confidence level of 95%, between the two groups of non-offensive jokes and offensive ones.

4.2 Features for offense

With the selected features, a classification of these into three groups has been carried out, distinguishing affective, content and syntactic & morphological markers.

Syntactic & morphological markers reflect the style of writing and the types of terms used. These are elements which provide of coherence within texts (Weth, 2020) by relating terms within a sentence. In addition, part-of-speech markers such as nouns, adjectives, adverbs, verbs, auxiliary verbs, persons and tenses are considered as part of these markers. Results are shown in Table 1.

Affective markers covers sentiments, emotions and attitude terms within a sentence. In this case, the features derived from sentiment markers quantify negative and positive words/terms, according to the mentioned lexical resources. A similar procedure is followed for features associated with personal states and emotions such as anger, disgust, joy, like, love, sadness, surprise. Results are included in Table 2.

Content markers indicate terms related to the content of a sentence: words from diverse categories used in LIWC dictionary (social, biology or religion) and hateful words, negative stereotypes and moral defects from Hurtlex dictionary, among other categories (see Table 3).

As observed in Table 1, among syntactic & morphological features, first personal pronouns, both singular and plural, and second personal pronouns in singular have a higher ratio of occurrence in non-offensive jokes than in offensive ones. However, the third personal pronoun in plural follows an opposite pattern. Although being highly present in offensive and non-offensive tweets, variables regarding articles (a, an, the), adjectives (cruel, bored, awful) and auxiliary verbs (am, has, might), have a higher frequency in offensive texts. Uniquely considering these variables, articles have the most outstanding difference of occurrence between both types of texts, being mostly used in offensive contexts. As articles define a noun as specific or unspecific, their appliance in line with the explanation about the use of personal pronouns, it might be useful to increment the distance between the sender and the object of the joke. For instance: *"You **the** bomb." "No, you **the** bomb." In America, **a** compliment. In **the** Middle East, **an** argument.* Adjectives also have a wider presence in offensive texts. By taking into account this context, and the fact that these words make reference to an attribute of a thing/person, terms used tend to be hurtful, like in this example: *What **do** you get if you cross **an illiterate african american** with **an illegal hispanic** immigrant looking for **a** green card? **A** United States soldier.*

When inspecting the results for affective features (see Table 2) we observe that negative emotions (anger, disgust, fear and sad-

			Non-offensive		Highly-offensive	
Lexicon	**Feature**	**p-value**	**Mean**	**Variance**	**Mean**	**Variance**
LIWC	I	1.35E-45	0.0706	0.0051	0.0351	0.0036
LIWC	Personal Pronouns	5.50E-11	0.1268	0.0062	0.0964	0.0061
LIWC	Article	9.58E-10	0.0748	0.0038	0.0915	0.0047
PoS	Adjective	1.87E-07	0.0816	0.004	0.0968	0.0049
LIWC	They	2.76E-07	0.0064	0.0004	0.0127	0.001
LIWC	Prepositions	3.87E-07	0.1037	0.0043	0.0893	0.0035
LIWC	Auxiliary Verb	2.67E-06	0.0902	0.003	0.1007	0.0031
PoS	1st Plural Person	3.25E-06	0.0033	0.0002	0.0012	0.0001
PoS	Adverbs	2.91E-05	0.0566	0.0033	0.048	0.0031
PoS	Noun	8.87E-05	0.2511	0.0088	0.2379	0.0092
PoS	2nd Person Singular	4.93E-03	0.0013	0.0001	0.0005	3.0e-05

Table 1: Syntactic & morphological features belonging to non- and highly offensive jokes.

			Non-offensive		Highly-offensive	
Lexicon	**Feature**	**p-value**	**Mean**	**Variance**	**Mean**	**Variance**
EmoSenticNet	Surprise	2.14E-13	0.0409	0.0032	0.0639	0.0057
SentiSense	Fear	2.31E-11	0.0078	0.0005	0.015	0.001
LIWC	Positive Emotions	2.38E-10	0.0322	0.0021	0.0223	0.0015
LIWC	Inhibition	0.00014	0.0056	0.0003	0.0036	0.0002
LIWC	Anxiety	1.65E-04	0.0039	0.0002	0.0027	0.0002
LIWC	Affective Processes	2.70E-04	0.0583	0.004	0.0487	0.0031
SentiSense	Disgust	3.00E-04	0.03	0.002	0.0366	0.0022
LIWC	Anger	1.67E-03	0.0087	0.0005	0.0127	0.0009
SentiSense	Sadness	1.08E-02	0.0033	0.0002	0.0053	0.0004
SentiSense	Like	1.80E-02	0.0276	0.0016	0.0244	0.0015
SentiSense	Joy	1.80E-02	0.0058	0.0003	0.0094	0.0006
SentiSense	Love	2.74E-02	0.0061	0.0003	0.0044	0.0003

Table 2: Affective features belonging to non- and highly offensive jokes.

ness) appear to be highly present through offensive jokes, in contrast to non-offensive ones. Moreover, the offensive set presents a higher amount of terms related to surprise, an emotion that could be either positive or negative. Additionally, affective processes from LIWC and positive emotions in general tend to appear mostly in non-offensive jokes.

A different trend is visible for terms associated to the emotion of joy. Results expose a greater occurrence in offensive texts than in non-offensive ones. When inspecting the linguistic resource the joy variable is extracted from, it is observed that the *gay* term is associated with this emotion −as it was in old English−, although it also is nowadays a term associated to a sexual orientation, as shown in the following examples: *I am **laughing** at these ladies waking up and being like Hey wanna become **gay icons** today?* and *Why do we hate making up **gay jokes**? Because it's always a pain in the as**.

Results regarding the content features are observed in Table 3. Content features are related to the topic of the jokes. It is noticeable that words associated to biology, humans, sexual, social, religion, negative stereotypes, moral and behavioural defects, swear words and ethnic slurs are mostly used in highly offensive jokes than in non-offensive ones. A good example of this kind of use is the following: *Where do most **black people** work? In **jail**.*

The most notorious differences between non-offensive and offensive texts are observed in features with jokes regarding sexuality (gay, lesbian, prostitute), religion (Jewish, christian, Christmas), swear words, negative stereotypes and ethnic slurs (Mexican, Chinese, black people) and moral or behavioural defects (jail, death).

			Non-offensive		Highly-offensive	
Lexicon	**Feature**	**p-value**	**Mean**	**Variance**	**Mean**	**Variance**
LIWC	Social	1.38E-12	0.1161	0.0088	0.1418	0.0091
LIWC	Biology	1.09E-14	0.0363	0.0031	0.0534	0.0038
LIWC	Quantifiers	1.38E-07	0.0206	0.0011	0.0304	0.0019
LIWC	Humans	2.11E-38	0.0103	0.0006	0.0283	0.0017
LIWC	Sexual	2.39E-38	0.0038	0.0002	0.0198	0.0016
LIWC	See	1.23E-09	0.0111	0.0008	0.0197	0.0015
LIWC	Exclusive	1.88E-08	0.0213	0.0012	0.0143	0.0009
LIWC	Leisure	2.63E-07	0.0201	0.0015	0.0136	0.001
LIWC	Religion	5.86E-15	0.0026	0.0002	0.0115	0.0012
Hurtlex	Negative stereotypes and ethnic-slurs	8.64E-40	0.0004	2.2e-05	0.0105	0.0008
Hurtlex	Moral & behavioural defects	2.56E-23	0.0023	0.0001	0.01	0.0006
LIWC	Swear words	6.95E-27	0.0009	5.0e-05	0.0082	0.0006

Table 3: Content features belonging to non- and highly offensive jokes.

5 Classification experiments

This section focuses on the classification of the jokes as non-offensives or highly offensives. The execution of experiments is performed by dividing the training set of the dataset in 80% for training and 20% for testing. As it is a binary classification, the offensive set is considered as the positive class, and the non-offensive set as the negative class. The classifiers applied are: Support Vector Machine (SVM), Random Forests (RF) and Logistic Regression (LR). For evaluating the performance of the classifiers, measures of accuracy, precision (PR) and F_1-score were computed with a five-fold cross validation. As baselines we employed SVM, RF and RL with Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) for text representation. Their results are shown in Tables 4 and 5. In a second set of experiments, the three classifiers were trained with the 35 most relevant linguistic features identified as statistically significant in the exploratory analysis (see Tables 1, 2 and 3). Table 6 provides the results obtained by each classifier trained with the most relevant linguistic features and Table 8 shows the rates achieved.

Compared to the results obtained with the baselines, the three classifiers perform better using the features proposed in this study, especially when classifying offensive samples. As we can see in Table 5, for the baselines the main problem is that highly offensive jokes are classified as non-offensive. In relation to the precision metric, it is observed that the best performance is obtained by RF when using all the linguistic features, similarly when using BoW and TF-IDF.

6 Ablation test

In order to quantify the contribution provided by each group of features in the classifiers performance, an ablation test has been done. Table 7 shows that, in general, all classifiers perform worse when removing any group of features. However, content features are the most important for the classification task. The removal of this set of features decreases in a substantial manner the F_1-score for all classifiers. This can be observed in FPR, FNR, TPR and TNR showed in Table 8. These values worsened when removing this set of features: increasing the FPR and decreasing the TNR. As a consequence, the recall decreases and the precision is altered.

Removing syntactic & morphological features also generates a drop in F_1-score metric, similar for all the classifiers (see Table 7) but not as strong as for the content features case. The performance of the classifiers differ when removing this group of features. All models present a greater percentage of offensive instances misclassified (increase in the FNR) and less capability of classify positive cases properly (decrease in the TPR), as shown in Table 8. However, FPR and TNR improve in SVM and LR, contrary to RF which does not present any variation in these metrics.

Regarding the affective features, a non ex-

	Non-offensive	Highly-offensive		
Model	**F1-score**	**F1-score**	**F1-macro**	**Accuracy**
SVM	0.70	0.61	0.66	0.66
RF	0.66	0.05	0.35	0.49
LR	0.70	0.61	0.66	0.66

Table 4: F1-score & Accuracy of the baselines with BoW+TF*IDF.

	FPR	**FNR**	**TPR**	**TNR**	**PR**
SVM	0.18	0.48	0.52	0.82	0.75
RF	0.004	0.98	0.02	0.99	0.86
LR	0.18	0.48	0.52	0.82	0.75

Table 5: Classification of the baselines with BoW+TF*IDF.

	Non-offensive	Highly-offensive		
Model	**F1-score**	**F1-score**	**F1-macro**	**Accuracy**
SVM	0.76	0.72	0.74	0.74
RF	0.77	0.75	0.76	0.76
LR	0.74	0.72	0.73	0.73

Table 6: F1-score & Accuracy with the 35 linguistic features proposed.

	SVM	**RF**	**LR**
	F1-macro	**F1-macro**	**F1-macro**
All features	0.74	0.76	0.73
Affective	0.73 (↓ 0.01)	0.72 (↓ 0.04)	0.75 (↑ 0.02)
Syntactic & morphological	0.72 (↓ 0.02)	0.73 (↓ 0.03)	0.72 (↓ 0.01)
Content	**0.65 (↓ 0.09)**	**0.66 (↓ 0.1)**	**0.66 (↓ 0.07)**

Table 7: Ablation test for SVM, RF and LR.

		FPR	**FNR**	**TPR**	**TNR**	**PR**
All features	SVM	0.17	0.34	0.66	0.83	0.81
	RF	0.16	0.31	0.69	0.84	0.82
	LR	0.22	0.32	0.68	0.78	0.77
Affective	SVM	0.16	0.36	0.64	0.84	0.81
	RF	0.19	0.36	0.64	0.81	0.78
	LR	0.18	0.31	0.69	0.82	0.80
Syntactic &	SVM	0.15	0.39	0.61	0.85	0.81
morphological	RF	0.16	0.37	0.63	0.84	0.80
	LR	0.20	0.36	0.64	0.80	0.78
Content	**SVM**	**0.34**	**0.37**	**0.63**	**0.66**	**0.66**
	RF	**0.33**	**0.35**	**0.65**	**0.67**	**0.68**
	LR	**0.33**	**0.35**	**0.65**	**0.66**	**0.67**

Table 8: Classification metrics for the classifiers with all features and for the ablation test.

pected result is observed for the LR model (Table 7). This classifier obtains better results when removing this set. Looking at FPR and TNR (Table 8), there is a slight improvement in their values. However, it is noticeable how the decrease of misclassified instances in the positive class (lower FPR) widely contributes to the increase of 0.03 in the precision score in comparison with the LR model trained with all features (Table 8). This res-

ult must be explored in more detail as a future work. A first hypothesis could be related to the confusion effect that the emotion *joy* could introduce in the sense we explained in Section 4.2. One reason could be that, in this situation, SVM and RF models are more robust than the LR model.

7 Classification with less features

The results of the ablation test show that content features were the most relevant for the classification task. Then, a second set of experiments has been carried out only with the content features to answer the RQ4. In these experiments, the classifiers were trained with the complete training set and the test set was used to evaluate their performance. The results of these experiments are provided in Tables 9 and 10.

When the classifiers use the content features RF obtains the same results for accuracy and F_1-score as when using all of the linguistic features, while SVM and LR increase their performance. Taking into account that LR and SVM with linear kernel as hyperparameter are both linear classifiers, the results observed for SVM and LR when trained with all features could be due to the effect that multicollinearity (Bayman and Dexter, 2021) has over both models. That is to say, their vulnerability towards small changes in the data and difficulties on identify feature importance. To explore if the correlation between features could justify the effect of multicollinearity, an exploratory analysis has been done. We could see that a content feature like *social* from LIWC presents a significant correlation ($rho = 0.42$) with a syntactic & morphological feature as it is *personal pronouns*, and content feature *moral & behavioural defects* from Hurtlex correlates with *disgust* ($rho = 0.24$) and with *fear* ($rho = 0.35$) from SentiSense. Therefore, some of these relations could be introducing redundant information, and worsening the classifiers performance.

8 Discussion of results

Regarding RQ1, we identified in our preliminary analysis which are the features that distinguish non-offensive humour from the offensive one. As we see in Tables 1, 2 and 3, a set of content, syntactic & morphological and affective features are useful to differentiate between the two classes of humour. Among the content features *negative stereotypes*, *moral and behavioural defects* and *swear words* are used in a very different way in both classes of humour. A possible reason for this result, could be that offensive humour is mainly reserved to ridicule minority groups or people that present certain behaviours that contradict mainstream values.

Among the syntactic & morphological features, we observe that the first person pronouns, both singular and plural, and second person pronouns in singular have a higher ratio of occurrence in non-offensive jokes than in offensive ones. A possible explanation for this result can rely on the depersonalization of the sender when saying something hurtful. This can be used as a mechanism to take off responsibilities of conveying offensive jokes and removes any possible feeling of guilty. However, third person pronoun in plural follows an opposite pattern, being more frequent in offensive jokes. This result allows to think that offensive jokes share linguistic patterns with other communicative phenomena related to prejudice, as hate speech (Chulvi, Toselli, and Rosso, 2022) and extremism (Torregrosa et al., 2022) where a more frequent use of "they" narratives has been observed. Regarding the features that capture aspects related to emotions, we observe that negative ones (anger, disgust, fear and sadness) appear to be highly present through offensive jokes in comparison to the non-offensive ones. Therefore, at least in this dataset, we can conclude that the offensive jokes are used to convey negative emotions towards particular groups, values and behaviours.

In response to RQ2, we observe that all the classifiers perform better using the proposed linguistic features in comparison with the baselines and all of them perform better distinguishing non-offensive humour from the offensive one. We used standard machine learning classifiers avoiding transformers, given that our main objective focuses on a descriptive analysis of the features that could contribute to the explainability of the results.

In this sense, a result from the ablation tests is that content features are the ones that contribute in a substantial manner for all classifiers (RQ3). This role of content features is in line with some first researches in this area, that showed the importance of cer-

		FPR	FNR	TPR	TNR	PR
BoW+TF-IDF	SVM	0.14	0.49	0.51	0.86	0.76
	RF	0.01	0.90	0.10	0.99	0.86
	LR	0.14	0.49	0.51	0.86	0.76
Content	SVM	0.23	0.26	0.74	0.77	0.73
	RF	0.20	0.28	0.72	0.80	0.75
	LR	0.24	0.25	0.75	0.76	0.72

Table 9: Classification metrics of the baselines and the classifiers with content features.

		Non-offensive	Highly-offensive		
	Model	**F1-score**	**F1-score**	**F1-macro**	**Accuracy**
BoW+TF-IDF	SVM	0.76	0.61	0.68	0.70
	RF	0.72	0.19	0.45	0.58
	LR	0.76	0.61	0.69	0.70
Content	SVM	0.77	0.74	0.75	0.75
	RF	0.78	0.73	0.76	0.76
	LR	0.77	0.74	0.75	0.75

Table 10: F1-score & Accuracy of the classifiers with the baselines and with content features.

tain words in the detection of humour (Mihalcea and Strapparava, 2005) even when the strategy was the opposite (Sjöbergh and Araki, 2007).

Regarding to RQ4, we may conclude that it is possible to adopt a strategy with less computational resources, as long as a previous study is carried out, as shown in Section 4 and in the ablation test (Section 6). It is relevant to consider that the set of the most relevant features, the ones that we called content features in our experiments, come from two different linguistic resources: LIWC and Hurtlex.

9 Conclusions and future work

In this work we have represented two sets of jokes (non-offensive and highly offensive ones) with the use of computational linguistics resources such as LIWC, Hurtlex, SentiSense and EmoSentiNet. The goal was to identify which linguistic features are used differently in non-offensive and offensive humour. We have used these features in a classification task. Subsequently, by applying an ablation test, we were able to detect which groups of features contribute the most. We have used these features in a strategy for using less computational resources, showing that it is possible to obtain the same performance. From a social science point of view, these results allows us to take a step towards a research program that explore how offensive humour is used to construct otherness and underpin prejudice.

As future work, we plan to compare our results with Transformers-based models, although instead of comparing the effectiveness, we plan to focus on identifying similarities and differences between the features highlighted by the attention mechanism and our linguistic features. Moreover, we plan to integrate the most relevant linguistic features in Transformers and deep learning-based models to help explainability during their decision-making process when detecting hurtful humour.

Acknowledgements

This work was done in the framework of the FairTransNLP research project (PID2021-124361OB-C31) funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe. It has been developed with the support of valgrAI - Valencian Graduate School and Research Network of Artificial Intelligence and the Generalitat Valenciana, and co-funded by the European Union.

References

Baccianella, S., A. Esuli, and F. Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference*

on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA).

Bandyopadhyay, S., D. Das, N. Howard, A. Hussain, A. Gelbukh, and S. Poria. 2013. Enhanced SenticNet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems*, 28(02):31–38.

Bassignana, E., V. Basile, and V. Patti. 2018. Hurtlex: A Multilingual Lexicon of Words to Hurt. In CEUR-WS, editor, *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018*, volume 2253, pages 1–6, Turin, Italy.

Bayman, E. O. and F. Dexter. 2021. Multicollinearity in logistic regression models. *Anesthesia & Analgesia*, 133(2):362–365. https://journals.lww.com/anesthesia-analgesia/Fulltext/2021/08000/Multicollinearity_in_Logistic_Regression_Models.12.aspx.

Bergson, H. 1900. *Le rire:essai sur la signification du comique*. Félix Alcan, Paris, France.

Billig, M. 2005. *Laughter and Ridicule: toward a social critique of humour*. Sage, London.

Cambria, E., S. Poria, R. Bajpai, and B. Schuller. 2016. SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2666–2677, Osaka, Japan.

Castro, S., L. Chiruzzo, and A. Rosa. 2018. Overview of the HAHA Task: Humor analysis based on human annotation at IberEval 2018. In *IberEval@ SEPLN*, pages 187–194.

Chiruzzo, L., S. Castro, M. Etcheverry, D. Garat, J. Prada, and A. Rosa. 2019. Overview of HAHA at IberLEF 2019: Humor analysis based on human annotation. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, CEUR Workshop Proceedings. CEUR-WS.

Chiruzzo, L., S. Castro, S. Góngora, A. Rosa, J. A. Meaney, and R. Mihalcea. 2021. Overview of HAHA at IberLEF 2021: Detecting, Rating and Analyzing Humor in Spanish. *Procesamiento del Lenguaje Natural*, 67(0):257–268.

Chulvi, B., A. H. Toselli, and P. Rosso. 2022. Fake news and Hate Speech: Language in Common. Technical report, I International Seminar on Artificial Intelligence and disinformation. https://arxiv.org/pdf/2212.02352.pdf.

Cignarella, A. T., V. Basile, M. Sanguinetti, C. Bosco, P. Rosso, and F. Benamara. 2020. Multilingual irony detection with dependency syntax and neural models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1346–1358, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

de Albornoz, J. C., L. Plaza, and P. Gervás. 2012. SentiSense: An easily scalable concept-based affective lexicon for sentiment analysis. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3562–3567, Istanbul, Turkey. European Language Resources Association (ELRA).

Dobai, A. and N. Hopkins. 2020. Humour is serious: Minority group members' use of humour in their encounters with majority group members. *European Journal of Social Psychology*, 50(2):448–462.

Ford, T. and M. Ferguson. 2004. Social Consequences of Disparagement Humor: A Prejudiced Norm Theory. *Personality and Social Psychology Review*, 8(1):79–94.

Ford, T. E., C. F. Boxer, J. Armstrong, M. Moya, and J. R. Edel. 2008. More than "just a joke": the prejudice-releasing function of sexist humor. *Personality & social psychology bulletin*, 34(2):159–70.

Frenda, S., A. T. Cignarella, V. Basile, C. Bosco, V. Patti, and P. Rosso. 2022. The unbearable hurtfulness of sarcasm. *Expert Systems with Applications*, 193:116398.

Freud, S. 1960. *Jokes and their Relation to the Unconscious*. Norton, Harmondsworth, England.

González, J. Á., L. F. Hurtado, and F. Pla. 2020. Transformer based contextualization of pre-trained word embeddings for irony detection in Twitter. *Information Processing and Management*, 57:1–15.

Grover, K. and T. Goel. 2021. Haha@ iberlef2021: Humor analysis using ensembles of simple transformers. In *IberLEF@ SEPLN*, pages 883–890.

Hossain, N., J. Krumm, M. Gamon, and H. Kautz. 2020. SemEval-2020 task 7: Assessing humor in edited news headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 746–758, Barcelona (online). International Committee for Computational Linguistics.

Hutto, C. and E. Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

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 *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Loria, S. and et.al. 2020. Textblob: Simplified Text Processing. https://textblob.readthedocs.io/en/dev/. Online; accessed 01 March 2022.

Martin, R. A. and T. E. Ford. 2018. Chapter 1 - introduction to the psychology of humor. In R. A. Martin and T. E. Ford, editors, *The Psychology of Humor (Second Edition)*. Academic Press, second edition edition, pages 1–32.

Meaney, J., S. Wilson, L. Chiruzzo, A. Lopez, and W. Magdy. 2021. Semeval 2021 task 7: Hahackathon, detecting and rating humor and offense. In *SemEval 2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense*, pages 105–119.

Merlo, L. I. 2022. When humour hurts: A computational linguistic approach. Final degree project. Technical report. http://hdl.handle.net/10251/188166.

Mihalcea, R. and C. Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Moudgil, A. 2017. Short jokes. https://www.kaggle.com/datasets/abhinavmoudgil95/short-jokes. Online; accessed 01 March 2022.

Nielsen, F. Å. 2011. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In M. Rowe, M. Stankovic, A.-S. Dadzie, and M. Hardey, editors, *Proceedings of the ESWC2011 Workshop on "Making Sense of Microposts": Big things come in small packages*, volume 718 of *CEUR Workshop Proceedings*, pages 93–98.

Ortega-Bueno, R., P. Rosso, and J. E. Medina Pagola. 2022. Multi-view informed attention-based model for Irony and Satire detection in Spanish variants. *Knowledge-Based Systems*, 235:107597.

Peng Qi, Yuhao Zhang, Y. Z. J. B. and C. D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *In Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Potamias, R. A., G. Siolas, and A. G. Stafylopatis. 2020. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, 32(23):17309–17320.

Potash, P., A. Romanov, and A. Rumshisky. 2017. SemEval-2017 task 6: #HashtagWars: Learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 49–57, Vancouver, Canada. Association for Computational Linguistics.

Reyes, A., M. Potthast, P. Rosso, and B. Stein. 2010. Evaluating humour features on web comments. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*

(LREC'10), Valletta, Malta. European Language Resources Association (ELRA).

Ripoll, R. M. and I. Q. Casado. 2010. Laughter and positive therapies: Modern approach and practical use in medicine. *Revista de Psiquiatría y Salud Mental (English Edition)*, 3(1):27–34.

Romero-Sánchez, M., H. Carreterio-Dios, J. L. Megías, M. Moya, and T. Ford. 2017. Sexist Humor and Rape Proclivity: The Moderating Role of Joke Teller Gender and Severity of Sexual Assault. *Violence against women*, 23(8):951–972.

Sjöbergh, J. and K. Araki. 2007. Recognizing humor without recognizing meaning. In F. Masulli, S. Mitra, and G. Pasi, editors, *Applications of Fuzzy Sets Theory*, pages 469–476, Berlin, Heidelberg. Springer Berlin Heidelberg.

Song, B., C. Pan, S. Wang, and Z. Luo. 2021. DeepBlueAI at SemEval-2021 task 7: Detecting and rating humor and offense with stacking diverse language model-based methods. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1130–1134, Online, August. Association for Computational Linguistics.

Tausczik, Y. and J. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29:24–54.

Torregrosa, J., G. Bello-Orgaz, E. Martínez-Cámara, J. D. Ser, and D. Camacho. 2022. A survey on extremism analysis using natural language processing: definitions, literature review, trends and challenges. *Journal of Ambient Intelligence and Humanized Computing*.

Veale, T., K. Feyaerts, and G. Brône. 2006. The cognitive mechanisms of adversarial humor. *Humor-international Journal of Humor Research - HUMOR*, 19:305–339.

Warriner, A. B., V. Kuperman, and M. Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Weth, C. 2020. Distinguishing Syntactic Markers From Morphological Markers. A Cross-Linguistic Comparison. *Frontiers in Psychology*, 11:2082.

Wiegand, M., J. Ruppenhofer, A. Schmidt, and C. Greenberg. 2018. Inducing a lexicon of abusive words – a feature-based approach. 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 1046–1056, New Orleans, Louisiana. Association for Computational Linguistics.

Yin, W. and A. Zubiaga. 2022. Hidden behind the obvious: Misleading keywords and implicitly abusive language on social media. *Online Social Networks and Media*, 30:100210.

Zhang, Z., X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Zhuang, L., L. Wayne, S. Ya, and Z. Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Appendix

Target	Keywords
Sexism	She, woman, mother, girl, b*tch, he, man, blond, p*ssy
Body	Fat, thin, tall, short, bald
Origin	Mexican, Mexico, Irish, Ireland, Chinese, Asian
Sexual orientation	Gay, lesbian, homo, LGBT, trans
Racism	Black, white people, nig**
Ideology	Feminism, lefty
Religion	Muslim, Jewish, Jew, Catholic, Jesus, Christmas
Health	Blind, deaf, r*tard, dyslexic, wheelchair

Table 11: Offensive keywords in the *HaHackaton* dataset (Meaney et al., 2021).

Target	Keyword = Target
A **fat** woman just served me at McDonalds and said "Sorry about the wait". I replied and said, "Don't worry, you'll lose it eventually".	Yes
Don't worry if a **fat** guy comes to kidnap you... I told Santa all I want for Christmas is you.	No

Table 12: Examples of jokes with keywords mentioned in the HaHackaton overview paper (Meaney et al., 2021).

Tool or Lexicon	Feature	p-value	Mean	Variance
SentiWordNet	Sentiment Score	0.0021	0.5188	0.0061
AFINN	Valence Score	2.5e-11	0.6446	0.0041
VADER	Sentiment score	1.282e-11	0.0824	0.1908
TextBlob	Polarity score	1.31e-07	0.0708	0.0758
	Subjectivity score	4.83e-07	0.4114	0.0995
ANEW	Valence score	2.3e-10	5.7582	0.2126
	Dominance score	4.23e-10	5.5608	0.0968
	Arousal score	0.011	4.0808	0.1116
Lexicon of abusive words extended	Score	0.003	0.4715	0.013

Table 13: Tagger features in non-offensive tweets in the humorous subset.

Tool or Lexicon	Feature	p-value	Mean	Variance
SentiWordNet	Sentiment Score	0.0021	0.5084	0.0061
AFINN	Valence Score	2.5e-11	0.6236	0.0049
VADER	Sentiment score	1.282e-11	-0.0476	0.1801
TextBlob	Polarity score	1.31e-07	0.0164	0.0621
	Subjectivity score	4.83e-07	0.35	0.0811
ANEW	Valence score	2.3e-10	5.6276	0.243
	Dominance score	4.23e-10	5.4778	0.1049
	Arousal score	0.011	4.1241	0.1504
Lexicon of abusive words extended	Score	0.003	0.4836	0.0175

Table 14: Tagger features in highly offensive tweets in the humorous subset.

Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 99-110 recibido 15-12-2022 revisado 07-02-2023 aceptado 15-02-2023

Construcción del RomCro, un corpus paralelo multilingüe

Construction of RomCro, a multilingual parallel corpus

Gorana Bikić-Carić[1], Bojana Mikelenić[1], Metka Bezlaj[2]
[1] Facultad de Humanidades y Ciencias Sociales de la Universidad de Zagreb
{gbcaric, bmikelen}@ffzg.hr
[2] Departamento de Lingüística de la Universidad de Zadar
mbezlaj21@unizd.hr

Resumen: En este trabajo se presentan las fases de construcción de un corpus paralelo multilingüe de cinco lenguas romances y croata. El corpus contiene oraciones originales provenientes de textos literarios de los siglos XX y XXI, alineadas con sus traducciones al resto de los idiomas. El orden original de las oraciones ha sido cambiado. El corpus cuenta con 15,9 millones de palabras y está disponible en las plataformas *Sketch Engine* y ELRC.

Palabras clave: corpus multilingüe, corpus paralelo, lenguas romances, croata.

Abstract: In this article we present the phases of construction of a parallel multilingual corpus of five Romance languages and Croatian. The corpus contains original sentences from literary texts from the 20th and 21st centuries, aligned with their translational equivalents in remaining languages. The original order of sentences is scrambled. The corpus counts with 15.9 million words and is available on platforms *Sketch Engine* and ELRC.

Keywords: multilingual corpus, parallel corpus, Romance languages, Croatian.

1 Introducción

A diferencia de los corpus monolingües, los corpus multilingües contienen textos correspondientes a varias lenguas. Estos corpus pueden ser paralelos (formados por textos escritos en una lengua y sus traducciones a otras)[1] o comparables (construidos por textos en varias lenguas que pertenecen al mismo tipo, o se usan para su construcción las mismas técnicas de muestreo) (McEnery, Xiao y Tono, 2006: 47). En los corpus paralelos, es preciso distinguir entre corpus unidireccionales (lengua fuente → lengua meta), bidireccionales (lengua fuente ↔ lengua meta) y multidireccionales (lengua(s) fuente(s) ↔ varias lenguas meta) (McEnery, Xiao y Tono, 2006: 48; Mikelenić y Tadić, 2020: 3932).

En cuanto a la utilidad de los corpus paralelos, es sabido que hoy en día la conveniencia de utilizar esos recursos en las investigaciones lingüísticas es enorme. Se pueden usar para todo tipo de investigaciones lingüísticas (contrastivas, traductológicas, fraseológicas, lexicográficas, glotodidácticas, etc.) (p. ej. Granger, Lerot y Petch-Tyson, 2003; Teubert, 2007), pero también pueden ser muy valiosos tanto en la enseñanza de la traducción (López Rodríguez, 2016) como en el entrenamiento de sistemas de traducción automática (Koehn et al., 2007) y en la extracción terminológica (Lefever, Macken y Hoste, 2009).

[1] Para los diferentes usos del término *paralelo* (ing. *parallel*) en la literatura, v. McEnery, Xiao y Tono (2006: 47).

ISSN 1135-5948 DOI 10.26342/2023-70-8

En este trabajo se presenta el proceso de construcción del RomCro, un corpus paralelo multilingüe que cuenta con seis lenguas. El RomCro está alineado, lematizado, anotado morfosintácticamente y compuesto de textos literarios escritos en cinco lenguas romances (en español, francés, italiano, portugués, rumano) y en croata. Este es el primer corpus paralelo multilingüe y casi completamente multidireccional que alinea textos literarios en diversas lenguas romances y una eslava. En otros corpus similares la lengua eslava era la lengua pivote (Terzić et al., 2020; Grabar et al., 2018; Akimova et al., 2020). Por eso, es importante incidir en que se trata de un corpus multidireccional, ya que incluye traducciones de cada texto a todas las demás lenguas del corpus.

La creación de este corpus empezó en 2019 como parte de un proyecto[2] de la Cátedra de Lingüística Románica del Departamento de Estudios Románicos de la Facultad de Humanidades y Ciencias Sociales de la Universidad de Zagreb bajo la dirección de Gorana Bikić-Carić. Además de la directora, en el proyecto participan los colaboradores Dražen Varga, Bojana Mikelenić y Metka Bezlaj. Las determinadas fases de desarrollo y los resultados de varias investigaciones lingüísticas hechas a base de unos subcorpus del RomCro ya se han presentado en algunas conferencias[3] y publicaciones (Bikić-Carić y Bezlaj, 2022; Bikić-Carić, 2020).

Este artículo se organiza en cinco partes: después de la introducción, en la segunda parte se mencionan algunas bases de textos existentes que contienen las mismas lenguas; en la tercera parte se presenta la composición actual del corpus, mientras que en la cuarta parte describimos todas las fases de desarrollo del corpus con enfoque en varios desafíos y problemas que surgieron a lo largo de su construcción. La quinta parte está dedicada a la conclusión y propone unas pautas para futuras investigaciones.

2 Trabajo previo

Antes de pasar a la descripción de las varias fases de desarrollo del RomCro, cabe destacar las razones por las cuales optamos por construir un nuevo corpus, si bien hay muchos corpus multilingües que ya están disponibles a los investigadores. A pesar de la gran utilidad de los recursos existentes, según nuestro conocimiento, el RomCro es el primer corpus que abarca las cinco lenguas romances ya mencionadas y el croata, tanto en la versión original como en las traducciones.

En cuanto a los corpus existentes, se trata más bien de unas colecciones más extensas en varias lenguas que permiten la extracción de textos paralelos en los idiomas incluidos en el RomCro,

[2] http://www.ffzg.unizg.hr/roman/odsjek/projekti/romcro/

[3] Mikelenić, B. y M. Bezlaj. Desafíos en la construcción de un corpus paralelo multilingüe. *XIII International CORPUS Linguistics Conference – CILC2022*, Università degli studi di Bergamo, mayo 26-28, 2022;

Bikić-Carić, G. y M. Bezlaj. Neke specifičnosti upotrebe određenog člana u romanskim jezicima (s posebnim naglaskom na francuski i španjolski). *70 godina izučavanja romanskih kultura, jezika i književnosti na Filozofskom fakultetu Univerziteta u Sarajevu*, Filozofski fakultet Univerziteta u Sarajevu, diciembre 3-4, 2021;

Bezlaj, M. y G. Bikić-Carić. Le choix entre l'infinitif et une forme conjuguée après les verbes d'opinion dans cinq langues romanes. *Considérations philologiques en contexte français et francophone*, Filološki fakultet Blaže Koneski Sveučilišta Sv. Ćiril i Metod u Skoplju, Skopie, noviembre 19-20, 2021;

Mikelenić, B. y M. Bezlaj. Construcción del RomCro: un corpus paralelo de lenguas romances y croata". *III Encuentro de Jóvenes Hispanistas*, Eötvös Loránd Tudományegyetem, Budapest, marzo 3-5, 2021;

Bikić-Carić, G. y M. Bezlaj. Construcción de un corpus multilingüe y su aplicación en el análisis contrastivo de los artículos. *XLIX Simposio de la Sociedad Española de Lingüística*, Universitat Rovira i Virgili, Tarragona, enero 21-24, 2020.

pero su uso implica numerosos problemas metodológicos. Por ejemplo, existen varios recursos lingüísticos multilingües formados por documentos legales y otros textos de la Unión Europea[4]. El más conocido de estos recursos es el *JRC-Acquis*, un corpus paralelo con más de mil millones de palabras (Steinberger et al., 2006). Estos recursos, aunque traducidos por traductores profesionales, suelen caracterizarse por un estilo jurídico y un vocabulario legal muy específico (Mikelenić y Tadić, 2020: 3932-3933; Mikelenić, 2020: 186).

Por otra parte, recursos como el *OpenSubtitles*[5] (Tiedemann, 2012; Lison y Tiedemann, 2016) están compuestos por traducciones de subtítulos y de esta manera ofrecen una gama más extensa de registros. Sin embargo, hay que tener en cuenta no solo la estructura y el formato determinado de los subtítulos, sino también el hecho de que los traductores de estos textos en la mayoría de los casos se desconocen. Además, la lengua fuente suele ser desconocida también, es decir, podemos suponer que es una sola, normalmente el inglés (Mikelenić, 2020: 186-187). Ambos tipos de recursos, es decir, tanto los subtítulos como los corpus de la UE, están recopilados y procesados automáticamente, lo que aumenta la posibilidad de errores ortográficos, de alineación, etc. (Mikelenić y Tadić, 2020: 3932).

Teniendo en cuenta todas las desventajas de los recursos existentes, decidimos crear un nuevo corpus multilingüe con las características que se irán presentando a lo largo de este trabajo. En breve, quisiéramos destacar que el RomCro es un corpus diferente de los ya disponibles, dado que está compuesto de textos literarios traducidos por traductores profesionales, lo que implica una calidad de traducción del más alto nivel. Como se verá en el capítulo siguiente, consideramos que el lenguaje de las novelas se acerca más al lenguaje general (si lo comparamos con el lenguaje legal, por ejemplo).

Asimismo, los textos que forman parte del RomCro se seleccionaron manualmente. Tanto el procesamiento inicial de los textos como la alineación se hicieron automáticamente, pero los resultados luego fueron revisados y corregidos a mano (v. subcapítulos 3.3 y 3.4), disminuyendo así la contaminación del corpus por errores de diversa índole y creando un recurso más fiable.

Finalmente, se debe subrayar que el RomCro es un corpus multidireccional, lo que significa que está compuesto de textos originales en cada lengua y de sus traducciones a todas las demás lenguas del corpus. En otras palabras, la lengua fuente es siempre conocida y sabemos exactamente, entre otras cosas, qué texto se tradujo a partir de qué original. Lo dicho supone que el RomCro es especialmente valioso a los traductores y traductólogos. Los beneficios del uso de corpus lingüísticos en los estudios de traducción están ya muy bien destacados y argumentados (v. Baker, 1993; Laviosa, 1998; Johansson y Oksefjell, 1998). Un corpus multilingüe como este ofrece varias posibilidades en este sentido también, p. ej. en investigaciones sobre las características de la lengua de traducción.

Conviene decir también que, al incluir el croata, una lengua eslava de bajos recursos digitales, el RomCro abre un nuevo camino para investigaciones contrastivas entre la lengua croata y las lenguas romances. Hasta ahora, el croata se ha incluido en los corpus paralelos que incluyen el inglés u otras lenguas eslavas (como el checo, el búlgaro o el macedonio) (Mikelenić y Tadić, 2020: 3933). Según nuestro conocimiento, el único corpus paralelo que incluye el croata junto con una lengua romance es el corpus bilingüe español-croata compuesto por Mikelenić (Mikelenić y Tadić, 2020; Mikelenić, 2020).

[4] https://joint-research-centre.ec.europa.eu/language-technology-resources_en

[5] https://www.opensubtitles.org/es

Asimismo, el croata está presente en algunos corpus literarios multilingües, como, por ejemplo, *TransLiTex* (Fraisse et al., 2018) o *InterCorp* (Čermák, 2019), del que forman parte también los textos literarios. Sin embargo, *TransLiTex* contiene traducciones de un solo libro a 23 lenguas e *InterCorp* se compone de 40 lenguas, entre ellas todas las que están incluidas en el RomCro, pero como lengua pivote tiene el checo, lo que significa que todos los textos están traducidos del o al checo, pero no necesariamente entre sí.

3 Composición actual del corpus

En la Tabla 1 se muestran todos los títulos originales que forman parte del RomCro, 27 en total. Aunque en la primera fase del proyecto elegimos dos textos originales en cada lengua, al final optamos por incluir otros títulos disponibles. Actualmente en portugués y en croata hay 3 títulos, en italiano y en rumano 4, en francés 6 y en español 7. De esta manera, el RomCro se convirtió en una especie de corpus oportunista (McEnery y Hardie, 2012: 11). Todos estos títulos están traducidos a las demás lenguas del corpus, con 5 excepciones[6]. Construido así, el corpus actualmente cuenta con 157 textos (lo que incluye 27 textos originales y 130 traducciones), en lugar de 162 textos. En total, el corpus compuesto de esta forma contiene 15,9 millones de palabras y 18,5 millones de *tokens*.

		Títulos:
1	**ES**	La sombra del viento (C.R. Zafón, 2001)
2		La catedral del mar (I. Falcones, 2006)
3		El juego del ángel (C.R. Zafón, 2008)
4		El asombroso viaje de Pomponio Flato (E. Mendoza, 2008)
5		Soldados de Salamina (J. Cercas, 2001)
6		El mapa del tiempo (F. J. Palma, 2008)
7		El tiempo entre costuras (M. Dueñas, 2009)
8	**FR**	Seras-tu là ? (G. Musso, 2006)
9		HHhH (L. Binet, 2010)
10		Un barrage contre le Pacifique (M. Duras, 1950)
11		La Fée Carabine (D. Pennac, 1987)
12		L'amant (M. Duras, 1984)
13		A l'ombre des jeunes filles en fleur (M. Proust, 1919)
14	**IT**	Imprimatur (Monaldi & Sorti, 2002)
15		Le otto montagne (P. Cognetti, 2017)
16		La forma dell'acqua (A. Camilleri, 1994)
17		L'amica geniale (E. Ferrante, 2011)
18	**RO**	Maitreyi (M. Eliade, 1933)
19		Întâmplări în irealitatea imediată (M. Blecher, 1936)
20		Nostalgia (M. Cărtărescu, 1993)
21		Cartea şoaptelor (V. Vosganian, 2009)
22	**PT**	A viagem do elefante (J. Saramago, 2008)
23		Nenhum olhar (J. L. Peixoto, 2000)
24		As intermitências da morte (J. Saramago, 2005)
25	**CR**	Muzej bezuvjetne predaje (D. Ugrešić, 1998)
26		Mediteranski brevijar (P. Matvejević, 1987)

[6] No se pudo conseguir la traducción al portugués de la novela *Maitreyi* y al italiano de *Muzej bezuvjetne predaje*, aunque estas existen. Por otra parte, *El asombroso viaje de Pomponio Flato* no está traducido al rumano, mientras que *Dora i Minotaur: Moj život s Picassom* no tiene la versión española ni portuguesa.

27		Dora i Minotaur: Moj život s Picassom (S. Drakulić, 2015)

Tabla 1: Títulos originales incluidos en el RomCro (según las lenguas).

La Tabla 2 presenta la distribución de todos los textos del corpus por lenguas. Cada lengua del corpus cuenta con alrededor de tres millones de *tokens*, lo que incluye tanto los textos originales como las traducciones. El subcorpus francés es el más grande con casi 3,5 millones de *tokens*, mientras que el croata es el más pequeño con 2,8 millones.

En cuanto a los textos originales, la mayoría de los subcorpus tiene menos de 1 millón de palabras: el subcorpus español es el único que cuenta con un número mayor de *tokens* y el croata es otra vez el más pequeño. Naturalmente, la distribución de *tokens* en las traducciones refleja una situación contraria: el español tiene el menor número de *tokens*, mientras que las demás lenguas cuentan con más de 2,5 millones.

	Número total de *tokens*	**Número de *tokens* (textos originales)**	**Número de *tokens* (traducciones)**
Español	3,146,711	1,170,843	1,975,868
Francés	3,419,815	676,806	2,743,009
Italiano	3,024,176	483,537	2,540,639
Portugués	2,997,771	212,016	2,785,755
Rumano	3,158,078	437,474	2,720,604
Croata	2,838,744	174,781	2,663,963

Tabla 2: Distribución de *tokens* por lenguas y en textos originales y traducidos.

La distribución de *tokens* en textos originales y traducidos se ha proyectado igualmente en el Gráfico 1, que muestra los mismos datos en forma porcentual. El porcentaje de los textos originales desciende desde el español (37,2 % de textos originales) hacia el portugués (7 % de textos originales) y el croata (6,2 % de textos originales), mientras que los subcorpus francés, italiano y rumano son bastante similares en cuanto a su composición: el francés tiene un 19,8 % de textos originales; el italiano, 16 %, y el rumano, 13,9 %.

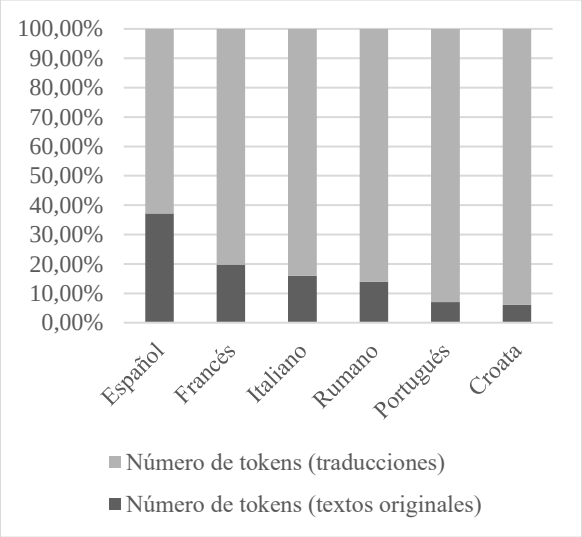


Gráfico 1: Distribución porcentual de *tokens* en textos originales y traducidos.

4 Fases de desarrollo

El desarrollo del RomCro se llevó a cabo de 2019 a 2022 y consistió en varias fases, comenzando por la selección de textos y terminando por hacer el recurso disponible en las plataformas *SketchEngine*[7] y ELRC[8].

4.1 Selección y recogida de textos

Teniendo en cuenta que la elaboración de cada corpus requiere la adopción de varios criterios en cuanto a su estructura y extensión, en un primer momento nos dedicamos al diseño del RomCro. Los criterios adoptados fueron los siguientes: 1) disponibilidad de textos traducidos a todas las lenguas del corpus, 2) textos literarios y 3) contemporáneos, 4) homogeneidad

[7] https://www.sketchengine.eu/

[8] https://www.lr-coordination.eu/

dialectal/variedades europeas, 5) formato digital. Como se verá a continuación, al compilar el corpus, algunos de estos criterios fueron modificados ligeramente por razones de naturaleza práctica.

El primer criterio que adoptamos fue la disponibilidad de los textos traducidos. A pesar de que todas las lenguas que forman parte del corpus sean idiomas oficiales de por lo menos un país, encontrar textos literarios originalmente escritos en cada una de las lenguas del corpus y traducidos a las demás lenguas resultó una tarea bastante difícil (sobre todo en cuanto a los textos en portugués, rumano y croata). Según nuestro conocimiento, la información sobre las traducciones a varias lenguas no está centralizada en ningún banco de datos, así que tuvimos que buscar las traducciones de cada título individualmente.

Se optó por las novelas para excluir del corpus los textos terminológicamente muy específicos, como p. ej. los de la legislación europea, es decir, para que el lenguaje del corpus fuera lo más neutral y general posible. Aunque somos conscientes de que la lengua literaria puede ser altamente idiosincrásica (Lawson, 2001: 294), argumentamos que estos rasgos distintivos no están tan presentes en las novelas como en otros géneros literarios (p. ej. cuentos u otras formas cortas, poesía o ensayos) (Mikelenić y Tadić, 2020: 3933).

Por otra parte, hay que tener en cuenta que la construcción de un corpus a partir de un solo tipo de texto puede ejercer una influencia en los resultados obtenidos. A pesar de esta limitación, vale destacar otra vez que el RomCro es el primer corpus paralelo que abarca esta combinación de lenguas y, por lo tanto, creemos que es una herramienta excelente para muchos tipos de investigaciones.

En cuanto al tercer criterio, es decir, a los límites temporales del corpus, queríamos confeccionar un corpus lo más sincrónico posible y, por consiguiente, intentamos encontrar textos del siglo XXI. No obstante, la versión final del corpus incluye algunos textos publicados en el siglo XX (una obra de la década de 1910, dos de la década de 1930, una de la década de 1950, tres de la década de 1980 y tres de la década de 1990) que decidimos incluir por el hecho de estar disponibles (v. Tabla 1).

Además, para evitar la diversidad dialectal y mantener el corpus lo más homogéneo posible, por el momento decidimos excluir los textos escritos en variedades no europeas del corpus. Esto es especialmente importante para el español y el portugués que tienen una producción literaria y traductora muy prolífica en América Latina. Las variedades americanas presentan unas características que las distinguen de las europeas y que muchas veces difieren según el país o la región. Dichos rasgos incluyen fenómenos a todos los niveles lingüísticos: desde el fonológico, morfológico y sintáctico hasta el léxico y pragmático.

Sin embargo, aunque todos los títulos originales incluidos en el corpus pertenecen a las variedades europeas, algunas de las traducciones se publicaron exclusivamente fuera de Europa, sobre todo en el caso del portugués. Por esta razón el RomCro en su versión actual cuenta con cuatro textos traducidos al portugués brasileño[9], pero el usuario puede elegir si quiere excluir estas traducciones de sus búsquedas a la hora de consultar el corpus, como se verá más adelante.

Finalmente, a fin de facilitar el procesamiento informático del corpus, nuestro objetivo fue encontrar el mayor número posible de textos seleccionados en formato digital. En varias ocasiones no fue posible cumplir con ese criterio,

[9] Se trata de las siguientes traducciones al portugués: *A fada carabina* de Daniel Pennac, *A forma da água* de Andrea Camilleri, *Acontecimentos na Irrealidade Imediata* de Max Blecher y *Nostalgia* de Mircea Cărtărescu.

así que muchos textos fueron escaneados y digitalizados automática y manualmente.

4.1.1 Aspectos legales en la construcción del corpus

Las cuestiones legales que afectan cada tipo de corpus son muchas. Dado que el RomCro está diseñado para que pueda usarse y citarse libremente con fines académicos y además está compuesto de textos contemporáneos, tuvimos que asegurarnos de que su composición no pudiera incurrir en una violación de los derechos de autor. Después de consultar los servicios jurídicos disponibles, resultó evidente que ni la legislación croata ni la legislación europea ofrecen pistas claras en cuanto a la construcción de los corpus digitales a partir de los textos literarios y sus traducciones.

No obstante, las fuentes consultadas nos corroboraron que los datos del corpus pueden usarse libremente para fines no-comerciales, siempre y cuando se citen adecuadamente, es decir, si figuran junto al título y nombre del autor. Además, la probabilidad de incumplir la normativa de los derechos de autor disminuye si las búsquedas del corpus se pueden hacer únicamente a nivel de la frase y si el corpus se diseñó con el único objetivo de ser utilizado en investigaciones científicas, es decir, con fines académicos y no-comerciales (p. ej. Wilkinson, 2006).

Por esta razón decidimos cambiar el orden de los segmentos del corpus para que nadie pudiera recuperar ninguno de los textos en su totalidad. Es más, la alineación se hizo a nivel de la oración (o dos oraciones, dependiendo de la traducción), así que nunca es posible recuperar más que esto. Por un lado, este procedimiento impide el análisis en los niveles mayores (p. ej. el análisis del discurso), pero, por otro lado, se protegen los derechos del autor.

4.2 Digitalización de los textos y preparación para la segmentación

Después de seleccionar y obtener los textos, tuvimos que escanear los que no pudimos encontrar en formato digital y hacer un reconocimiento óptico de caracteres automático (ing. OCR) con el programa *Abbyy FineReader*. Varios estudiantes de grado y de máster colaboraron en el proyecto, revisando y corrigiendo los resultados de esta digitalización, es decir, preparando los textos para la alineación automática. En esta fase se borraba todo lo que no era el cuerpo del texto (p. ej. imágenes, números de páginas) y se excluía todo lo que no se podía alinear con su traducción a otros idiomas (p. ej. dedicatorias, información sobre el autor). Este paso prolongó sustancialmente el trabajo, pero fue necesario para que el proceso de alineación fuera lo más fácil y rápido posible y el resultado final, el más correcto.

4.3 Segmentación, alineación y corrección manual

A continuación hemos segmentado y alineado los textos automáticamente con el programa *LF-Aligner*[10], un programa de acceso libre mediante el cual se puede utilizar la herramienta de alineación *Hunalign*[11] (Varga et al., 2005). La revisión del resultado de este proceso se hizo en el programa *Microsoft Excel*. Aunque trabajamos con 15 estudiantes que manejan varias lenguas romances, la mayoría de ellos no ha podido cumplir la tarea de revisar los segmentos alineados en todas las lenguas a la vez, lo que significaba que los textos se dividían entre dos o tres personas y se juntaban de nuevo. Al final, se hizo otra revisión por parte de uno de los tres profesores que participaron en el proyecto. En esta tarea se ha trabajado continuamente durante tres años académicos, combinándola con la revisión de resultados de la digitalización de nuevos títulos que se añadían al corpus.

Lo que también complicaba esta tarea fueron las diferencias en las traducciones de algunos textos, p. ej. libros con varias ediciones, aunque por la naturaleza del tipo de texto (novela), esto lo encontramos solo con títulos y autores ya conocidos en este sentido. Un ejemplo extremo es el del autor croata Matvejević, que conocía la mayoría de estas lenguas romances y traducía sus obras solo o en colaboración con el traductor, a menudo añadiendo o reformulando partes del texto.

[10] https://sourceforge.net/projects/aligner/

[11] https://github.com/danielvarga/hunalign

Hay que destacar que en ningún momento se añadió nada, tampoco en el caso de errores y omisiones no deliberadas (p. ej. la traducción de una oración no aparece), donde se optó por dejar el segmento vacío en el idioma en cuestión. Finalmente, como ya se destacó, en la última versión del corpus se cambió el orden de los segmentos para proteger los derechos del autor.

4.4 Lematización y anotación morfosintáctica

El paso siguiente fue la lematización y anotación morfosintáctica de los textos. Nos enfrentamos con la duda de seleccionar un etiquetador diferente para cada idioma u optar por uno que otorga resultados uniformes para todos, arriesgando tal vez la pérdida de algunos rasgos distintivos. Se presentaron dos opciones: los anotadores del proyecto *Universal Dependencies* – UD[12] (Nivre et al., 2016), que ofrecen etiquetas comparables para todos los idiomas o los anotadores del *Sketch Engine*[13] (Kilgarriff et al., 2004), que son *FreeLing*[14] (Padró, 2011) para español, francés, italiano y portugués y *MULTEXT-East*[15] (Erjavec et al., 2003; Erjavec 2017) para rumano y croata. Para poder tomar una decisión informada, hicimos pruebas con los dos grupos de anotadores, cuyos resultados se presentan en el apartado siguiente.

4.4.1 Selección de los etiquetadores

Se etiquetaron 50 oraciones en cada idioma, tanto de textos originales como de los traducidos, con ambos grupos de anotadores[16], marcando errores. La selección de los etiquetadores se basó en el análisis cualitativo de estos errores y a continuación comentamos algunos que se repetían.

La dificultad de la diferenciación entre los adjetivos y los participios es bien conocida en la literatura, dado que comparten la forma. Por eso no es de sorprender que ambos grupos de anotadores etiquetaban algunos adjetivos como participios (verbos), aunque del contexto era evidente de qué se trataba. Por ejemplo, en *un alma bendita* el adjetivo fue etiquetado como participio por el anotador del *Sketch Engine*, al igual que algunos adjetivos en portugués (*uma noite lançada*), francés (*les guerres perdues*) e italiano (*disgustata com'è dall'idea che Rino…*) por los anotadores de UD.

En cuanto a la elisión, es decir, la supresión de la vocal en francés y en español, para el francés los anotadores del *Sketch Engine* cometen errores de clase de palabra (*qu'* etiquetado como verbo, *l', n'* como sustantivos, *d'* como adjetivo), que por otro lado se marca bien en los anotadores de UD, aunque están presentes los errores de lematización. Las contracciones *del* y *al* en español se etiquetan en el *Sketch Engine* solo como preposiciones, mientras que el anotador de UD les asigna dos etiquetas separadas (preposición + artículo).

Por último, el problema de la diferenciación entre algunas conjunciones y pronombres relativos que comparten la forma lo ilustramos en español e italiano. Aquí el anotador del *Sketch Engine* para español distingue bien entre las dos clases (p. ej. *que*), mientras que el anotador para italiano le designa su propia etiqueta a *che*, de esta manera evitando el problema. Los anotadores de UD para ambos idiomas en algunos casos funcionan bien, pero se encontraron varios errores (pronombres etiquetados como conjunciones).

En estos ejemplos se podía ver que ambos grupos de etiquetadores tenían dificultades, pero hay que destacar que en general funcionaron bien. Es más, conocidos los puntos débiles de cada etiquetador, en el futuro se podrá ajustar las búsquedas para compensar los errores. Al final, la elección se hizo en conexión con el siguiente asunto: el modo de acceder al corpus.

[12] https://universaldependencies.org/

[13] https://www.sketchengine.eu/

[14] http://nlp.lsi.upc.edu/freeling

[15] http://nl.ijs.si/ME/

[16] Como parte de UD, para algunos idiomas existen varios etiquetadores desarrollados en proyectos diferentes. En estos casos, primero averiguamos qué etiquetador para el mismo idioma funcionaba mejor para nuestros textos. Los etiquetadores elegidos fueron los siguientes: portuguese-bosque-ud-2.6-200830 (portugués), french-gsd-ud-2.6-200830 (francés), italian-isdt-ud-2.6-200830 (italiano), romanian-nonstandard-ud-2.6-200830 (rumano), croatian-set-ud-2.6-200830 (croata), spanish-gsd-ud-2.6-200830 (español).

4.5 Acceso al corpus

En cuanto al acceso al corpus, se presentaron dos posibilidades: crear una interfaz propia o usar una ya existente. La primera opción significa más libertad en diseño y funcionalidad, pero también más gastos de creación y mantenimiento. Por otro lado, utilizar una interfaz ya existente da acceso a asistencia técnica y más visibilidad. Se decidió elegir el *Sketch Engine* y también sus anotadores, por la funcionalidad y varias posibilidades de búsqueda y análisis que otorga este programa.

Así, los corpus paralelos se pueden consultar individualmente o en conjunto, con la opción de las concordancias paralelas (ing. *Parallel Concordance*), como se aprecia en la Imagen 1. El orden de idiomas es arbitrario y cambiable. Es más, según las necesidades del usuario, los subcorpus en cada idioma se pueden añadir u omitir (con el signo "X" en el rincón derecho de cada casilla), así que uno tiene la opción de, por ejemplo, usar solamente dos o tres subcorpus. Asimismo, es posible hacer la búsqueda solo en un idioma o proponer traducciones en uno o varios idiomas (en la Imagen 1: *Translated as (optional)*).

El programa ofrece varias posibilidades bajo *búsqueda avanzada* (en la Imagen 1: *Advanced*), donde los metadatos que contiene el RomCro se pueden usar como filtros para crear los subcorpus deseados. Como se observa en la Imagen 2, estos son el título del libro (*segment.title*), el autor del libro (*segment.author*) y el idioma original (*segment.original_lang*). De esta manera es posible hacer un subcorpus del portugués brasileño eligiendo los títulos mencionados antes (v. la nota a pie de página no. 9) o un subcorpus de textos originales en croata y sus traducciones a otros idiomas. Por supuesto, los filtros se pueden combinar. Cabe mencionar que el uso de los filtros es opcional y depende de las necesidades del usuario. Así, si uno desea consultar el corpus de portugués sin diferenciar entre las variantes, solo marcará *pt* como el idioma original.

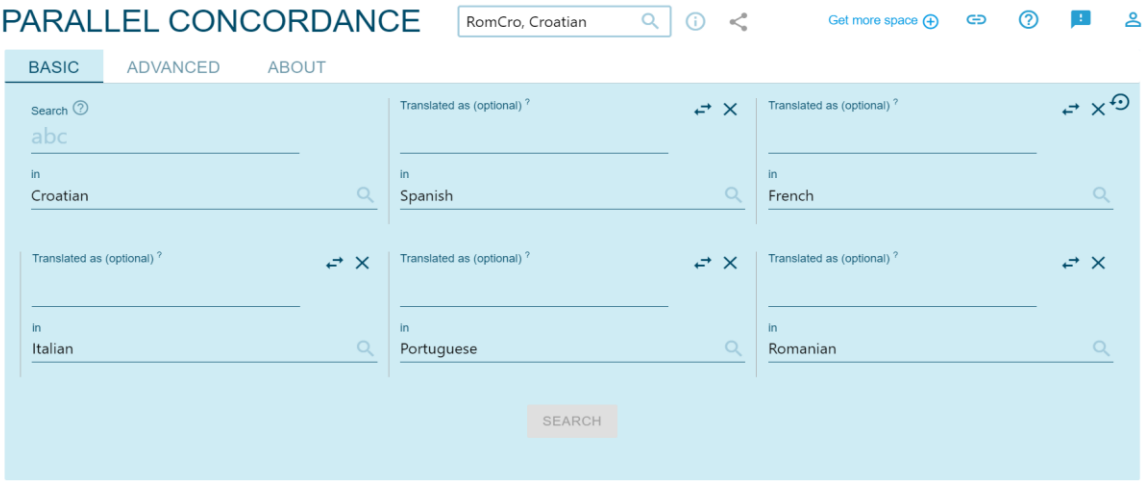


Imagen 1: La interfaz del *Sketch Engine* para obtener concordancias paralelas.

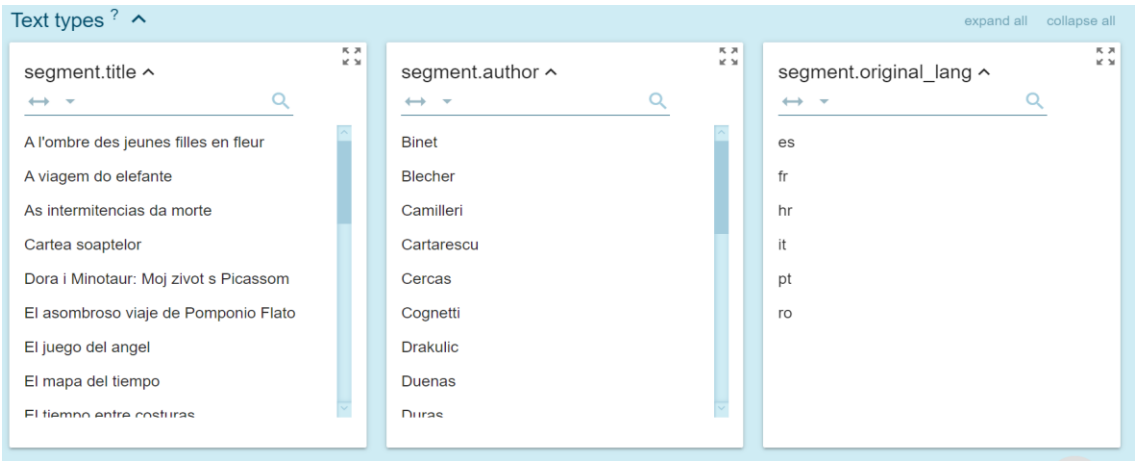


Imagen 2: Los metadatos del RomCro en el *Sketch Engine*.

El acceso al corpus es por ahora libre para los colegas de la Facultad de Humanidades y Ciencias Sociales de la Universidad de Zagreb. A otros usuarios del *Sketch Engine* les rogamos que nos contacten para otorgarles el acceso. Estamos haciendo lo posible para que el RomCro pronto sea accesible a todos los investigadores que deseen utilizarlo en esta forma.

Para los usuarios que prefieren trabajar directamente con el corpus, una versión en el formato TMX y otra en el formato TSV, también están disponibles en la plataforma ELRC (*European Language Resource Coordination*)[17], bajo la licencia CC-BY-NC-4.0. En ambos formatos, el orden de las lenguas es: español (es), francés (fr), italiano (it), portugués (pt), rumano (ro), croata (hr). Estas versiones no están lematizadas ni anotadas, pero ambos documentos contienen información sobre la lengua original, el escritor y el título de texto del que proviene cada segmento.

5 Conclusión y planes para el futuro

En este trabajo se ha presentado la construcción del RomCro, un corpus paralelo multilingüe de lenguas romances y croata que está disponible a través de las plataformas *Sketch Engine* y ELRC. Este corpus está alineado, lematizado y anotado morfosintácticamente. Es importante insistir aquí en que el RomCro es el único corpus multidireccional que cuenta con esta combinación de lenguas, lo que le hace muy valioso para lingüistas, traductores, profesores y otros profesionales.

En el futuro, esperamos ampliar el corpus con más textos literarios y/o de otros géneros e incluir más textos en variedades no europeas. Además, nos parece muy importante hacer el recurso disponible en su totalidad según los principios de la ciencia abierta, no solo a los investigadores croatas de nuestra facultad, sino también a la comunidad académica internacional.

También creemos que es necesario seguir con la realización de investigaciones de diversa índole. De hecho, este año hemos iniciado una continuación de nuestro proyecto que se dedica al análisis contrastivo de la oposición virtualidad-realidad en las lenguas romances y croata. Finalmente, todos los colaboradores en el proyecto están trabajando en la popularización de este recurso entre los colegas y estudiantes. Creemos que es tan importante fomentar el uso del RomCro en la enseñanza de las lenguas y de la traducción como animar a los estudiantes a que lo usen en la redacción de sus trabajos académicos.

Agradecimientos

Las autoras agradecen la financiación recibida por la Facultad de Humanidades y Ciencias Sociales de la Universidad de Zagreb. Igualmente

[17] https://elrc-share.eu/repository/search/?q=romcro

agradecemos a los estudiantes que han participado en el proyecto por su inestimable colaboración.

Bibliografía

Akimova, M., A. Belousova, I. Pilshchikov, y V. Polilova. 2020. En *Actas de Parallel Corpora as Digital Resources and Their Applications*: https://parallelcorporadhn2020.github.io/talks/Akimova_Belousova_Pilshchikov_Polilova.html.

Baker, M., 1993. Corpus Linguistic and Translation Studies: Implications and Applications. En *Text and Technology: In Honour of John Sinclair*, páginas 233-250, John Benjamins Publishing Company, Philadelphia / Amsterdam.

Bikić-Carić, G. 2020. Quelques particularités dans l'expression de la détermination du nom. Comparaison entre cinq langues romanes. *Studia Universitatis Babes-Bolyai-Philologia*, 65(4):39-54.

Bikić-Carić, G. y M. Bezlaj. 2022. Neke specifičnosti upotrebe određenog člana u romanskim jezicima (s posebnim naglaskom na francuski i španjolski). Filozofski fakultet Univerziteta u Sarajevu. To appear.

Čermák, P. 2019. InterCorp. A parallel corpus of 40 languages. En *Parallel Corpora for Contrastive and Translation Studies: New resources and applications*. páginas: 93-101, John Benjamins Publishing Company, Philadelphia / Amsterdam.

Erjavec, T. 2017. MULTEXT-East. En *Handbook of Linguistic Annotation*. páginas 441-462, Springer, Dordrecht.

Erjavec, T., C. Krstev, V. Petkevič, K. Simov, M. Tadić, y D. Vitas. 2003. The MULTEXT-East Morphosyntactic Specifications for Slavic Languages. En *Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages*, ACL (Budapest).

Fraisse, A., Q.-T. Tran, R. Jenn, P. Paroubek, y S. Fisher Fishkin. 2018. TransLiTex: A Parallel Corpus of Translated Literary Texts. En *Proceedings of the 11th Language Resources and Evaluation Conference*, European Language Resource Association.

Grabar N., O. Kanishcheva, y T. Hamon. 2018. Multilingual aligned corpus with Ukrainian as the target language. *SLAVICORP* (Prague).

Granger, S., J. Lerot, y S. Petch-Tyson (eds.). 2003. *Corpus-based Approaches to Contrastive Linguistics and Translation Studies.* Rodopi, New York.

Johansson, S. y S. Oksefjell (eds.). 1998. *Corpora and Cross-linguistic research: Theory, Method, and Case Stuidies.* Rodopi, Amsterdam / Atlanta.

Kilgarriff, A., P. Rychlý, P. Smrž, y D. Tugwell. 2004. The Sketch Engine. En *Proc Eleventh EURALEX International Congress*, páginas 105-116, Lorient, France.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, et al. 2007. *Moses: Open Source Toolkit for Statistical Machine Translation.* En *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, páginas 177-180, Association for Computational Linguistics.

Laviosa, S. 1998. The Corpus-based Approach: A New Paradigm in Translation Studies. *Meta : journal des traducteurs / Meta: Translator's Journal*, 43(4):474-479.

Lawson, A. 2001. Collecting, aligning and analysing parallel corpora. En *Small Corpus Studies and ELT: theory and practice*, páginas 279-309, John Benjamins Publishing Company, Philadelphia / Amsterdam.

Lefever, E., L. Macken, y V. Hoste. 2009. Language-independent bilingual terminology extraction from a multilingual parallel corpus. En *Proceedings of the 12th Conference of the European Chapter of the ACL*, páginas 496-504 (Athens).

Lison, P. y J. Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. En *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, páginas 923–929,

European Language Resource Association (Portorož, Slovenia).

López Rodríguez, C. I. 2016. Using Corpora in Scientific and Technical Translation Training: Resources to Identify Conventionality and Promote Creativity. *Cadernos de traduçao*, 1:88-120.

Mikelenić, B. y M. Tadić. 2020. Building the Spanish-Croatian Parallel Corpus. En *Proceedings of the 12th Language Resources and Evaluation Conference*, páginas 3932-3936, European Language Resource Association.

Mikelenić, B. 2020. *Kontrastivna korpusna analiza prijedložne dopune u španjolskome i njezinih ekvivalenata u hrvatskome*, tesis doctoral. Filozofski fakultet, Zagreb.

McEnery, T. y A. Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.

McEnery, T., R. Xiao, y Y. Tono. 2006. *Corpus-based language studies: An advanced resource book*. Routledge, London/New York:.

Nivre, J., M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. D. Manning, . . . y D. Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. En *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, páginas 1659-1666. European Language Resource Association (Portorož, Slovenia).

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

Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, y D. Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. En *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, páginas 2142-2147. European Language Resource Association (Genova, Italy).

Terzić, D., S. Marjanović, D. Stosic, y A. Miletic. 2020. Diversification of Serbian-French-English-Spanish Parallel Corpus ParCoLab with Spoken Language Data. En *Text, Speech, and Dialogue. TSD 2020*, páginas 61-70, Springer.

Teubert, W. (ed.). 2007. *Text Corpora and Multilingual Lexicography*. John Benjamins Publishing Company, Amsterdam / Philadelphia.

Tiedemann, J. 2012. *Parallel Data, Tools and Interfaces in OPUS.* En *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, páginas 2214-2218, European Language Resource Association (Istanbul, Turkey).

Varga, D., P. Halácsy, A. Kornai, V. Nagy, L. Nemeth, y V. Tron. 2005. Parallel corpora for medium density languages. En *Proceedings of the RANLP 2005*, páginas 590-596 (Borovets, Bulgaria).

Wilkinson, M. 2006. Legal aspects of compiling corpora to be used as translation resources: questions of copyright. *Translation Journal* 10(2): https://translationjournal.net/journal/36corpus.htm.

Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 111-122 recibido 15-12-2022 revisado 09-02-2023 aceptado 15-02-2023

Tuning BART models to simplify Spanish health-related content

Ajuste de modelos BART para simplificación de textos sobre salud en español

Rodrigo Alarcón, Paloma Martínez, Lourdes Moreno
Human Language and Accessibility Technologies group (HULAT)
Universidad Carlos III de Madrid
Leganés, Madrid, Spain
{ralarcon, pmf, lmoreno}@inf.uc3m.es

Abstract: Health literacy has become an increasingly important skill for citizens to make health-relevant decisions in modern societies. Technology to support text accessibility is needed to help people understand information about their health conditions. This paper presents a transfer learning approach implemented with BART (Bidirectional AutoRegressive Transformers), a sequence-to-sequence technique that is trained as a denoising autoencoder. To accomplish this task, pre-trained models have been fine-tuned to simplify Spanish texts. Since fine tuning of language models requires sample data to adapt it to a new task, the process of creating of a synthetic parallel dataset of Spanish health-related texts is also introduced in this paper. The results on the test set of the fine-tuned models reached SARI values of 59.7 in a multilingual BART (mBART) model and 29.74 in a pre-trained mBART model for the Spanish summary generation task. They also achieved improved readability of the original texts according to the Inflesz scale.
Keywords: lexical simplification, Spanish, language models, Spanish, multilingual BART.

Resumen: La alfabetización sanitaria se ha convertido en una habilidad cada vez más importante para que los ciudadanos tomen decisiones sobre su salud en las sociedades modernas. Para ayudar a las personas a comprender la información sobre su estado de salud, es necesaria una tecnología que facilite la accesibilidad de los textos. Este artículo presenta un enfoque de transfer learning implementado con BART (Bidirectional AutoRegressive Transformers), una técnica sequence-to-sequence que se entrena como un autoencoder de eliminación de ruido. Para llevar a cabo esta tarea, se han ajustado modelos preentrenados para simplificar textos en español. Dado que el ajuste de los modelos lingüísticos requiere datos de muestra para adaptarlos a una nueva tarea, en este artículo también se presenta el proceso de creación de un conjunto de datos paralelos sintéticos de textos en español relacionados con la salud. Los resultados en el conjunto de prueba de los modelos afinados alcanzaron valores SARI de 59,7 en un modelo multilingual BART (mBART) y 29,74 en un modelo mBART pre-entrenado para la tarea de generación de resumenes en español. Además lograron mejorar la legibilidad de los textos originales según la escala de Inflesz.
Palabras clave: Simplificación léxica, modelos del lenguaje, Español, BART multilingüe.

1 Introduction

Nowadays, the average citizen has access to much more information through the Internet than at any other time in history with a high impact on most people's daily lives. However, this information may have been written in a form that makes the content hard to understand (Saggion et al., 2015). Difficulty with texts on the Internet can affect a wide range of people such as deaf people, illiterate

ISSN 1135-5948 DOI 10.26342/2023-70-9

people, second language learners, and people with intellectual disabilities, among others (Moreno, Alarcon, and Martínez, 2020).

One of the most difficult content to understand is health-related content because of the excessive use of abbreviations, incomplete sentences, and specific terminology. Poor health literacy is a limiting factor that prevents patients from making well-informed health decisions, which can result in high costs for both healthcare institutions and the patient (Kauchak, Apricio, and Leroy, 2022).

Following this issue, there are standards and guidelines (UNE, 2018; Plainlanguage, 2017b; Plainlanguage, 2017a) that provide accessibility requirements and criteria to make the textual content more cognitively accessible, through the application of easy-to-read and plain language guidelines. For example, these requirements indicate that a text must be written in an active voice, use everyday words and/or use short sentences as much as possible. All these requirements and criteria are defined to provide a familiar and simple vocabulary used in texts in Plain Language. Nonetheless, this issue is difficult to address.

There are ways to follow these directives and manually deal with this problem. For instance, some websites offer simplified versions of their original content oriented to their target users [1][2]. However, this is a time consuming task. Therefore, over the years, different proposals to provide an automatic solution to this problem have emerged, the most prominent of which is Natural Language Processing (NLP) techniques (Alarcon, Moreno, and Martínez, 2021; Alarcón García, 2022).

This article proposes a transfer learning method to simplify Spanish texts with medical content. To achieve this, a state-of-the-art approach is presented, by fine-tuning multilingual BART models (Tang et al., 2020) with parallel data to lexical simplification of Spanish health-related content. This strategy was chosen because it has achieved state-of-the-art results in a diverse set of generation tasks (Martin et al., 2020) and outperforms Text-to-Text transfer transformers (T5) models of comparable size (Lewis et al., 2019).

[1]https://plenainclusionmadrid.org/blog/etapa-educativa-inclusion/

[2]https://plenainclusionmadrid.org/blog/reclutador-discapacidad-intelectual/

The contributions of this paper can be outlined as follows:

- Creation of a Spanish synthetic parallel resource for the training and validation of simplification methods in the health domain. This resource contains pairs of original sentences related to simplified ones.
- Proposal of fine-tuning two mBART models for text-to-text generation, with the aim of simplifying Spanish health-related texts.

2 Related Work

Text simplification is the process of lexically and/or syntactically modifying a text to produce a simple version of the original text (Al-Thanyyan and Azmi, 2021), preserving its original meaning. Text simplification could benefit a wide range of people, to mention a few, may include second language learners (Paetzold and Specia, 2016b) or people with some type of disability, such as autism (Barbu et al., 2015), dyslexia (Wilkens, Oberle, and Todirascu, 2020) or some type of intellectual disability (Saggion et al., 2015; Alarcon, Moreno, and Martínez, 2021).

Over the years, resources to support training and/or evaluation of automatic text simplification algorithms have been shared. These resources belong either in a general domain such as resources with content from Wikipedia articles (Yimam et al., 2018; Ferrés and Saggion, 2022a) or other resources with a specific domain, such as resources with a medical vocabulary (Campillos Llanos et al., 2022). Additionally, there have been evaluation campaigns aimed at providing a solution to this task in a modular way (Truică, Stan, and Apostol, 2022), such as workshops that aimed to foster research on the detection of unusual words in a given text (Paetzold and Specia, 2016a; Yimam et al., 2017), others that focused in ranking words according to their complexity (Shardlow et al., 2021) and competitions that aimed to propose replacements for unusual words or phrases (McCarthy and Navigli, 2007). Other works presented strategies using parallel resources, as in the work of (Zhu, Bernhard, and Gurevych, 2010) who proposed a complex word identification trans-

	EASIER	
	Sentence	**Substitutes**
EASIER	El tabaquismo constituye el principal problema de salud pública prevenible en los países desarrollados siendo un factor determinante de numerosas **patologías**. (Smoking is the main preventable public health problem in developed countries and is a determining factor in numerous **pathologies**.)	enfermedades (diseases), dolencias (afflictions), trastornos (disorders)
	Sentence	**Simple version**
Paralell Instance	El tabaquismo constituye el principal problema de salud pública prevenible en los países desarrollados siendo un factor determinante de numerosas **patologías**. (Smoking is the main preventable public health problem in developed countries and is a determining factor in numerous **pathologies**.)	El tabaquismo constituye el principal problema de salud pública prevenible en los países desarrollados siendo un factor determinante de numerosas **enfermedades**. (Smoking is the main preventable public health problem in developed countries and is a determining factor in numerous **diseases**.)
Paralell Instance	El tabaquismo constituye el principal problema de salud pública prevenible en los países desarrollados siendo un factor determinante de numerosas **patologías**. (Smoking is the main preventable public health problem in developed countries and is a determining factor in numerous **pathologies**.)	El tabaquismo constituye el principal problema de salud pública prevenible en los países desarrollados siendo un factor determinante de numerosas **dolencias**. (Smoking is the main preventable public health problem in developed countries and is a determining factor in numerous **afflictions**.)
Paralell Instance	El tabaquismo constituye el principal problema de salud pública prevenible en los países desarrollados siendo un factor determinante de numerosas **patologías**. (Smoking is the main preventable public health problem in developed countries and is a determining factor in numerous **pathologies**.)	El tabaquismo constituye el principal problema de salud pública prevenible en los países desarrollados siendo un factor determinante de numerosas **transtornos**. (Smoking is the main preventable public health problem in developed countries and is a determining factor in numerous **disorders**.)
	EASY-DPL	
	Sentence	**Target Word - Substitutes**
Easy-DPL	En pacientes con esquizofrenia la incidencia de **acatisia** fue de 6,2% para aripiprazol y de 3,0% para placebo. (In patients with schizophrenia, the incidence of **akathisia** was 6.2% for aripiprazole and 3.0% for placebo.)	acatisia (akathisia) - incapacidad de quedarse quieto (inability to remain still)
Easy-DPL	lteraciones gastrointestinales: Frecuente (1% y <10%): dolor abdominal, diarrea, **dispepsia**, náuseas y vómitos. (Gastrointestinal alterations: Frequent (1% and <10%): abdominal pain, diarrhea, **dyspepsia**, nausea and vomiting.)	dispepsia (dyspepsia) - enfermedades del estómago (diseases of the stomach)
	Sentence	**Simple version**
Paralell Instance	En pacientes con esquizofrenia la incidencia de **acatisia** fue de 6,2% para aripiprazol y de 3,0% para placebo. (In patients with schizophrenia, the incidence of **akathisia** was 6.2% for aripiprazole and 3.0% for placebo.)	En pacientes con esquizofrenia la incidencia de **incapacidad de quedarse quieto** fue de 6,2% para aripiprazol y de 3,0% para placebo. (In patients with schizophrenia, the incidence of **inability to stay still** was 6.2% for aripiprazole and 3.0% for placebo.)
Paralell Instance	Alteraciones gastrointestinales: Frecuente (1% y <10%): dolor abdominal, diarrea, **dispepsia**, náuseas y vómitos. (Gastrointestinal alterations: Frequent (1% and <10%): abdominal pain, diarrhea, **dyspepsia**, nausea and vomiting.)	..Alteraciones gastrointestinales: Frecuente (1% y <10%): dolor abdominal, diarrea, **enfermedades del estómago**, náuseas y vómitos. (Gastrointestinal alterations: Frequent (1% and <10%): abdominal pain, diarrhea, **diseases of the stomach**, nausea and vomiting.)

Table 1: EASIER and EASY- DPL corpora substitutes dataset examples.

lation method with a tree-based simplification model trained on a parallel Wikipedia and simple Wikipedia dataset. A prominent project for the Spanish language is the Simplext project (Saggion et al., 2015), where a parallel resource was generated in Spanish to reduce the syntactic complexity of texts.

A recent competition is CLEF-2022, where three tasks focused on the automatic simplification of scientific texts were proposed. Of these tasks we can highlight tasks 2 and 3, where the teams with the best results were based on the use of language models in their strategies. Task 2 consisted in the detection of terms in a text that requires an explanation for the whole text to be understood (Ermakova et al., 2022a). Participants obtained a train set of 453 examples annotated with difficulty scales and a test set of 116763 sentences, where each participant had to determine a score for the difficulty of the term in the target text. Approaches based on IDF (Inverse Document Frequency) term weighting (Mostert et al., 2022), approaches based on semantic similarity complemented by different lexical and syntactic features (Huang and Mao, 2022), and finally, methods based on transfer learning (Talec-Bernard, 2022) were presented. Task 3 aimed at generating simplified versions of scientific texts (Ermakova et al., 2022b). Participants were given 648 parallel sentences to develop their architectures, and to validate them, they obtained a test set of 116724 instances to be evaluated by the organizers. Transfer learning-based approaches were presented, where models were fine-tuned with the task data and other existing English language corpora (Monteiro, Aguiar, and Araújo, 2022). A similar approach to the one proposed in this paper was described in (Rubio and Martínez, 2022) by fine-tuning a BART model to simplify sentences. This method was highlighted by the task organizers as it showed that tasks 2 and 3 of the competition are largely related.

In addition, in the Shared Task on Lexical Simplification (TSAR 2022) (Saggion et

al., 2023) for English, Portuguese and Spanish languages, given a sentence/context and a complex target word, participants had to generate up to 10 possible substitutes ordered by simplicity. To perform this task, the organizers shared different resources for the training and/or validation of systems. ALEXSIS (Ferrés and Saggion, 2022a) dataset, which contains open domain terms, was used in the case of Spanish language. Prior to the publication of the task, the authors of this work experimented with this resource to rank substitutes for target words, achieving an accuracy score of 0.51 (Alarcón García, 2022). However, since the objective of this work is to simplify medical terms, a specific medical domain resource is proposed to train/validate the methods described in this work.

Research with BART out of competitions has been also published recently, as (Cumbicus-Pineda et al., 2022) outperforms other approaches in three different English datasets using several language models, trained with complex sentences to predict simple sentences and others trained with simple sentences to predict complex sentences, achieving higher values in the SARI metric than other similar approaches. (Chamovitz and Abend, 2022) described a BART-based method that also defines a series of simplification operations based on cognitive simplification guidelines, improving the performance compared to a baseline model in a dataset for the English language. Some of these operations consisted of ambiguity reduction, rephrasing, summarizing, reordering or deleting paragraphs. The work of (Štajner, Sheang, and Saggion, 2022) presented a sentence simplification approach by experimenting with transformer models for text simplification such as BART and T5 combined with control mechanisms, achieving results comparable to other previous systems.

This paper is based on metrics and methods from BART's previously described work and presents a text-to-text generation approach by fine-tuning two mBART models for the task of text simplification. To accomplish this task, this paper also describes the process of creating a synthetic Spanish resource containing lexical modifications to original sentences.

3 Datasets

This Section briefly describes the data used to fine-tune the BART language models. These data are obtained from the EASIER[3] and EASY-DPL[4] (Segura-Bedmar and Martínez, 2017) corpora.

3.1 EASIER

The EASIER corpus was created to support Complex Word Identification (CWI) and Substitute Generation/Selection (SG/SS) tasks, two important processes in lexical simplification, targeting an audience with intellectual disabilities. With this objective, linguistic experts in easy-to-read and simple language guidelines have annotated 260 news documents on various topics, including health news. Currently, this resource has gathered 8155 complex words and 7894 proposed substitutes.

For the purpose of text simplification, data from the SG/SS dataset were used (Alarcon, 2021). EASIER corpus contains simple alternative substitutes to existing complex words. To create the instances of the tuning process, parallel versions are created by taking the original sentences, the target complex word, and the proposed substitutes. As a result, 7894 instances were obtained where for each instance there is a code, original sentence, and the same original sentence where one or more words have been replaced. Table 1 shows examples of the original content of the EASIER corpus dataset and the content of the generated parallel versions. The datasets of this resource are available in csv formats.

3.2 EasyDPL

The remaining data used for the experiments in this article come from the Easy-DPL corpus (easy drug leaflets). This corpus was annotated by three annotators trained for their task, where they annotated the adverse effects section of 306 medical leaflets, resulting in 1400 adverse reactions detected along with their simplest synonym. Table 1 shows examples of the original content and the generated parallel versions. This resource is available in XML and BRAT formats.

[3]Easier Corpus Repository github.com/LURMORENO/EASIER_CORPUS

[4]https://github.com/isegura/EasyDPL

3.3 EASIER-EasyDPL dataset

A Spacy model[5] in Spanish was used to generate the parallel dataset to eliminate duplicate instances, tokenizing, and sentence splitting, among other operations. For this version of the resource, possible errors in grammatical forms were ignored when substituting a target word in the original sentence. Table 2 shows some statistics between the resources described above.

	Number of instances	%
EASIER	7894	86.5
Easy-DPL	1230	13.5
Total	9124	100

Table 2: Number of instances for the EASIER and Easy-DLP resources.

4 Methods and system description

This Section describes the proposal, which is based on fine-tuning two pre-trained multilingual BART models from HuggingFace. The first model (MBART-50)[6] is 12 layers multilingual sequence-to-sequence model trained on 50 different languages, while the second model (MBART-ESP)[7] is a 12 layer Spanish language fine-tuned version of the first model (Tang et al., 2020) with the wiki_lingua dataset[8] for the summarization task. The hypothesis behind the choice of these models is to determine whether the model fine-tuned to the Spanish language is better at the simplification task than the base model because it was trained to better understand the Spanish language.

BART (Bidirectional AutoRegressive Transformers), (Lewis et al., 2019), is a sequence-to-sequence strategy trained as a denoising autoencoder. This technique resembles BERT and GPT as it uses a standard sequence-to-sequence Neural Machine Translation architecture (transformer) with a bidirectional encoder (Devlin et al., 2018) and a left-to-right decoder (Radford et al., 2018). This model could be fine-tuned to the simplification problem by taking a text sequence as input and producing a text sequence as output. Given a complex text 'x' and its references 'y', a model in inference time is used to select the simplification that maximizes this probability (e.g. $argmax_y p(y|x)$). To train a BART model, a bidirectional encoder similar to BERT is used, where spaces are masked from the input text (adding "noise"). Also, autoregressive decoder such as GPT is used, which reconstructs the original input, using the output of the encoder and the previous unmasked tokens.

For the experimentation of this work, the training data set described in Section 3 was used to fine-tune the models. The inputs to the process are the source sentence and the simplified sentence. Each model tokenizes each sentence and obtains the embeddings of the inputs for the transformers. With a transformer encoder, it is not necessary to pass each word individually through the input embedding, all words in the sentence are passed simultaneously and the word embeddings are simultaneously determined.

5 Experiments and results

Different experiments were performed with the data described in Section 3. These data were randomly divided into three sets with the help of the sklearn library, a training set (80%), a dev set (10%), and a test set (10%). The experiments and resources described in this article can be found in a public repository[9]. The objective of fine-tuning with this data is to create models capable of generating simplifications as close as possible to those provided by taking into account the lexical modifications of the synthetic parallel versions.

The evaluation metrics are the following:

- SARI: Measures the goodness of words that are added, deleted, and kept by the predictions. This metric was widely used in lexical simplification tasks (Xu et al., 2016).
- ROUGE: Measures the number of matching n-grams between the model-generated text and the dataset´s references. Because of using generative models in this work ROUGE is proposed as an evaluation metric. Although mBART models were fine-tuned with

[5]https://spacy.io/models/es

[6]https://huggingface.co/facebook/mbart-large-50

[7]https://huggingface.co/eslamxm/MBART-finetuned-Spanish

[8]https://huggingface.co/datasets/wiki_lingua

[9]https://github.com/ralarcong/BART_for_simplification

Perspicuity	Inflesz
0-40	Very difficult
40-55	Somewhat difficult
55-65	Normal
65-80	Easy enough
80-100	Very easy

Table 3: Interpretation of Inflesz Scale.

parallel data with lexical changes, they sometimes seek to reorder the content of a sentence, especially the MBART-ESP model, which was previously fine-tuned for the task of text summarization (Lin, 2004).

- Inflesz Scale: It was chosen to measure readability levels of the original texts, the target texts, and those predicted by the models. This metric, adapted to today's average Spanish reader, measures perspicuity, which refers to the level of clarity and comprehensibility of a text. Formula 1 shows the calculation of this metric where S represents the number of syllables, P the number of words and F the number of sentences. This metric can be used for any text domain, although it has initially been used in the healthcare domain to assess the readability of informed consent, package leaflets, and health education materials (Barrio-Cantalejo et al., 2008). Table 5 describes the interpretation for every range of values.

$$I = 206.835 - \frac{62.3S}{P} - \frac{P}{F} \qquad (1)$$

To train each model, different values of hyperparameters had to be explored. Fortunately, the Fast.ai library[10] helped by choosing a learning rate appropriate to the configuration set in each minibatch (Smith, 2018). By defining a "learner" object, the library is able to test between different learning rate values and plot the loss values. Figure 1 shows an example of this, where the learning rate was chosen before it diverges.

Table 5 shows the experimentation with the other hyperparameters. It was observed that the optimal number of epochs for this experiment was 4 since with more epochs the model started to overfit the data to the training data. Figure 2 shows an example of

[10]https://docs.fast.ai/

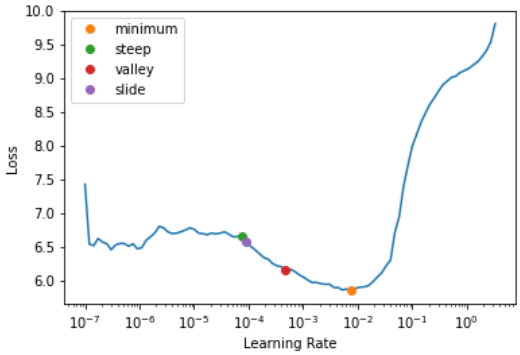


Figure 1: Loss value vs learning rate.

the MBART-ESP model, showing the loss in training and validation at 7 epochs, were at a higher epoch than the optimum the loss in training is reduced but the loss in validation is increased.

Hyperparam.	Value	Best
# epochs	[1,2,3,4,5,6,7]	[4]
Batch size	[1,2,3]	[1]
Max length, Min length	[(10,30),(10,40), (15,30),(10,50)]	(10,50)
# beams	[3][4][5]	[4]

Table 4: List of tested hyperparameters along with the best choice for the experiment.

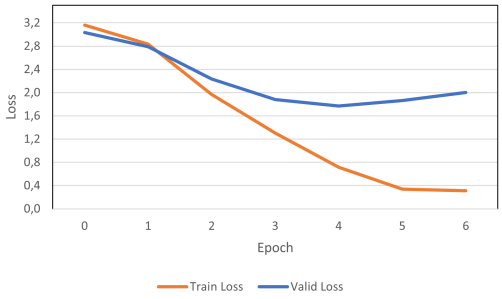


Figure 2: Loss value vs epochs.

When testing batch sizes, it was found that the best results were achieved with the length of 1, by reducing the potential noise that arises with increasing length. Also, increasing the batch length demanded more memory space, so it was decided to give preference to memory usage. On the other hand, when experimenting with the minimum and maximum output lengths, there was a noticeable change in the results when reducing the lengths, so the decision was made to keep the maximum optimal length at 50 words. Finally, when experimenting with the num-

ber of beams, it was decided to keep the default value of 4, since increasing the number of beams dramatically increased the training time without obtaining better results.

Once the optimal configuration was explored, the results shown in Table 5 were obtained with the train set data. The MBART-ESP model reached Rouge 1, Rouge 2, and Rouge L scores of 0.622, 0.477, and 0.573 respectively, and a SARI score of 43.68 points. While the MBART-50 model reached Rouge 1, Rouge 2, and Rouge L scores of 0.859, 0.82, and 0.858 respectively, and a SARI score of 67.3 points.

Additionally, these models were validated with the other two sets. Table 5 shows the results where it can be seen that in the dev set the MBART-ESP model reached scores of 0.682, 0.548, 0.635 in the Rouge 1, Rouge 2, and Rouge L metrics respectively, and a score of 29.175 points in the SARI metric. While the MBART-50 model reached scores of 0.928, 0.883, and 0.928 in the Rouge 1, Rouge 2, and Rouge L metrics respectively, and a score of 58.555 points in the SARI metric. In the test set the MBART-ESP model reached scores of 0.675, 0.535, and 0.627 in the Rouge 1, Rouge 2, and Rouge L metrics respectively, and a score of 29.749 points in the SARI metric. While the MBART-50 model reached scores of 0.926, 0.881, and 0.925 in the Rouge 1, Rouge 2, and Rouge L metrics respectively, and a score of 59.777 points in the SARI metric.

The results on these datasets suggest that because the MBART-ESP model was previously trained for the summarization task, in addition to attempting to perform lexical substitutions it attempts to summarize the content, thus scoring lower on the Rouge and SARI metrics than the MBART-50 model, which has been trained only for the text simplification task. Although ROUGE is not the most appropriate metric for this simplification task, it was used since it allowed the detection of the difference in the predictions of both models, being MBART-50 the one that better performed the necessary lexical replacements, as could be seen with the SARI metric.

An important feature of these fine-tuned models is that they perform the lexical substitutions for which they were trained. Furthermore, it additionally takes into account substitutions from other instances and attempts to modify complex words in the entire target sentence. Appendix A, Table A shows examples of the models input, target, and prediction. In the first example the target word to be replaced is expectación (expectation), however, both models predict a sentence where the target word and the word suscitas (suscitas) are replaced by a simpler substitute.

Different scenarios occur with the second example. The MBART-50 model performs the desired lexical substitutions as in example 1, but in some instances the MBART-ESP model attempts to summarize the content (example 2.1), as it was the model was previously trained for the task of summarization. Therefore, it is concluded that for this specific experimentation, the MBART-50 model is more appropriate, since it focuses on the simplification task to which it was trained (example 2.2).

Finally, to evaluate the readability of the predictions, in each dataset, the Inflesz metric was calculated along the original sentences (Source), the simplified sentences (Reference), and the predictions of the models (Prediction). Table 5 shows this score on every set, where it can be seen that both models improved the readability levels of the original sentences (Source), and in some cases surpassed the readability level of the simplified parallel sentences (Reference), such is the case of the predictions of the MBART-ESP model in the Train and Test sets, obtaining a score of 41.41 and 44.3 respectively. This is due to the fact that this model tends to summarize, and the predictions are shorter, thus obtaining a better score than the MBART-50 model that only performs lexical modifications.

Since the EASIER-EasyDPL dataset is introduced in this paper, there is no direct way of comparison with other approaches. However, Table 5 shows a comparison of our best result with other works for the English language with the SARI metric in task 3 of the SimpleText@CLEF-2022 workshop. As can be seen, the results are comparable to those present in the state of the art, as in (Monteiro, Aguiar, and Araújo, 2022) where they used a T5 model to perform the text simplification task reaching 31.26 SARI values on the workshop's dataset. Another approach to this competition presented the tuning of a BART model for the English text simplifica-

MBART-ESP						
Epoch	**Train Loss**	**Valid Loss**	**Rouge 1**	**Rouge 2**	**Rouge L**	**SARI**
0	3.687307	3.640873	0.309830	0.121615	0.244606	-
1	2.434027	2.700396	0.416706	0.219104	0.345381	-
2	1.119833	1.905820	0.571923	0.406836	0.510689	-
3	0.650932	1.837321	0.622281	0.477601	0.573849	43.6808
MBART-50						
Epoch	**Train Loss**	**Valid Loss**	**Rouge 1**	**Rouge 2**	**Rouge L**	**SARI**
0	1.021594	0.891353	0.679494	0.595874	0.667983	-
1	0.573852	0.563872	0.805591	0.754661	0.802736	-
2	0.371475	0.387605	0.853029	0.812000	0.852026	-
3	0.212017	0.366373	0.859541	0.820185	0.858498	67.3065

Table 5: Train dataset results (4 epochs with optimal configuration).

	Dev			
Fine-tuned model	**Rouge 1**	**Rouge 2**	**Rouge L**	**SARI**
MBART-ESP	0.6821	0.5484	0.6354	29.175
MBART-50	0.9287	0.8837	0.9281	58.555
	Test			
Fine-tuned model	**Rouge 1**	**Rouge 2**	**Rouge L**	**SARI**
MBART-ESP	0.6756	0.5358	0.6276	29.749
MBART-50	0.9261	0.8816	0.9251	59.777

Table 6: Dev and Test datasets results (model trained with optimal configuration).

	Src	**Ref**	**Pred M-ESP**	**Pred M-50**
Train	38.75	40.24	41.41	39.82
Dev	39.21	40.90	39.05	39.73
Test	38.63	40.81	44.30	39.75

Table 7: Inflesz scale results across the datasets.

System	**SARI**
Our approach	**59.7**
HULAT@CLEF	47.8
PortLinguE@CLEF	38.1
CLARA-HD@CLEF	37.4

Table 8: SARI values for the English dataset in the SimpleText workshop.

tion task reaching SARI values of 47.83 (Rubio and Martínez, 2022). In the same competition, the approach of (Menta and Garcia-Serrano, 2022) presented a transfer learning method where they combined control tokens such as word length, paraphrasing or syntactic complexity to help in the predictions of the COVID-SciBERT model, reaching SARI values of 37.4 in the workshop's dataset.

6 Conclusions

This paper presented the process of fine-tuning two mBART pre-trained models for text simplification for the Spanish language. Because this technique requires sample data for its execution, a new synthetic resource that includes data from two corpora oriented to the simplification of Spanish texts containing health-related terminology is also introduced. This resource was divided into three subsets for training, adjustment, and validation of the different fine-tuned models. In the training and fine-tuning phase, different configurations were experimented with in order to capture the best similarity to the target sentences of the sets.

The results in the training dataset shown the great difference between each pre-trained model. Similarly, the results of these models in the dev and test sets showed a great difference. Therefore, the predictions of both fine-tuned models were analyzed, where it was observed that both models lexically modified the target words in a sentence and also modified the learned words in other examples, optimizing the simplification task. But also the pre-trained model for the summarization task in some cases tended to reduce the sentence length instead of performing the lexi-

cal modifications, resulting in lower ROUGE and SARI scores, but improving on the Inflesz readability metric. In addition, these fine-tuned models showed comparable results in the SARI metric to approaches in a similar task for the English language.

As future work, it is planned to incorporate new resources to the training/fine-tuning/validation sets containing substitutes to target words with health-oriented content, such as the IULA resource (sentences of clinical cases in Spanish)(Marimon, Vivaldi, and Bel Rafecas, 2017) and also to extend the domain of the models with news resources such as the ALEXSIS dataset (Ferrés and Saggion, 2022b). More resources with plain an easy-to-read texts written by experts are also necessary to obtain models with better performance.

Moreover, as shown in this research, the tuning process was only performed on two embedding models, so it would be interesting to experiment with other multilingual models of different sizes and/or fine-tuned for other tasks.

Acknowledgments

This work is part of the R&D&i ACCESS2MEET (PID2020-116527RB-I0) project financed by MCIN AEI/10.13039/501100011033/, and the "Intelligent and interactive home care system for the mitigation of the COVID-19 pandemic" project (PRTR-REACT UE) awarded by CAM. CONSEJERÍA DE EDUCACIÓN E INVESTIGACION.

References

Al-Thanyyan, S. S. and A. M. Azmi. 2021. Automated text simplification: A survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Alarcon, R. 2021. Dataset of sentences annotated with complex words and their synonyms to support lexical simplification, March.

Alarcon, R., L. Moreno, and P. Martínez. 2021. Lexical simplification system to improve web accessibility. *IEEE Access*, 9:58755–58767.

Alarcón García, R. 2022. Lexical simplification for the systematic support of cognitive accessibility guidelines. *https://doi.org/10.1145/3471391.3471400*.

Barbu, E., M. T. Martín-Valdivia, E. Martínez-Cámara, and L. A. Urena-López. 2015. Language technologies applied to document simplification for helping autistic people. *Expert Systems with Applications*, 42(12):5076–5086.

Barrio-Cantalejo, I. M., P. Simón-Lorda, M. Melguizo, I. Escalona, M. I. Marijuán, and P. Hernando. 2008. Validación de la escala inflesz para evaluar la legibilidad de los textos dirigidos a pacientes. In *Anales del Sistema Sanitario de Navarra*, volume 31, pages 135–152. SciELO Espana.

Campillos Llanos, L., A. R. Terroba Reinares, S. Zakhir Puig, A. Valverde, and A. Capllonch-Carrión. 2022. Building a comparable corpus and a benchmark for spanish medical text simplification.

Chamovitz, E. and O. Abend. 2022. Cognitive simplification operations improve text simplification.

Cumbicus-Pineda, O. M., I. Gutiérrez-Fandiño, I. Gonzalez-Dios, and A. Soroa. 2022. Noisy channel for automatic text simplification.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Ermakova, L., I. Ovchinnikov, J. Kamps, D. Nurbakova, S. Araújo, and R. Hannachi. 2022a. Overview of the clef 2022 simpletext task 2: Complexity spotting in scientific abstracts.

Ermakova, L., I. Ovchinnikov, J. Kamps, D. Nurbakova, S. Araújo, and R. Hannachi. 2022b. Overview of the clef 2022 simpletext task 3: Query biased simplification of scientific texts.

Ferrés, D. and H. Saggion. 2022a. Alexsis: a dataset for lexical simplification in spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3582–3594.

Ferrés, D. and H. Saggion. 2022b. ALEXSIS: A dataset for lexical simplification in Spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3582–3594, Mar-

seille, France, June. European Language Resources Association.

Huang, J. and J. Mao. 2022. Assembly models for simpletext task 2: Results from wuhan university research group.

Kauchak, D., J. Apricio, and G. Leroy. 2022. Improving the quality of suggestions for medical text simplification tools. In *AMIA Annual Symposium Proceedings*, volume 2022, page 284. American Medical Informatics Association.

Lewis, M., Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

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

Marimon, M., J. Vivaldi, and N. Bel Rafecas. 2017. Annotation of negation in the iula spanish clinical record corpus. *Blanco E, Morante R, Saurí R, editors. SemBEaR 2017. Computational Semantics Beyond Events and Roles; 2017 Apr 4; Valencia, Spain. Stroudsburg (PA): ACL; 2017. p. 43-52.*

Martin, L., A. Fan, É. de la Clergerie, A. Bordes, and B. Sagot. 2020. Muss: multilingual unsupervised sentence simplification by mining paraphrases. *arXiv preprint arXiv:2005.00352*.

McCarthy, D. and R. Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pages 48–53.

Menta, A. and A. Garcia-Serrano. 2022. Controllable sentence simplification using transfer learning. *Proceedings of the Working Notes of CLEF*.

Monteiro, J., M. Aguiar, and S. Araújo. 2022. Using a pre-trained simplet5 model for text simplification in a limited corpus. *Proceedings of the Working Notes of CLEF*.

Moreno, L., R. Alarcon, and P. Martínez. 2020. Easier system. language resources for cognitive accessibility. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–3.

Mostert, F., A. Sampatsing, M. Spronk, and J. Kamps. 2022. University of amsterdam at the clef 2022 simpletext track. *Proceedings of the Working Notes of CLEF*.

Paetzold, G. and L. Specia. 2016a. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

Paetzold, G. and L. Specia. 2016b. Unsupervised lexical simplification for non-native speakers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Plainlanguage. 2017a. Plain english- free guides (co.uk).

Plainlanguage. 2017b. Plain language action and information network (plain).

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

Rubio, A. and P. Martínez. 2022. Hulat-uc3m at simpletext@ clef-2022: Scientific text simplification using bart. *Proceedings of the Working Notes of CLEF*.

Saggion, H., S. Štajner, S. Bott, S. Mille, L. Rello, and B. Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):1–36.

Saggion, H., S. Štajner, D. Ferrés, K. C. Sheang, M. Shardlow, K. North, and M. Zampieri. 2023. Findings of the tsar-2022 shared task on multilingual lexical simplification. *arXiv preprint arXiv:2302.02888*.

Segura-Bedmar, I. and P. Martínez. 2017. Simplifying drug package leaflets written in spanish by using word embedding. *Journal of biomedical semantics*, 8(1):1–9.

Shardlow, M., R. Evans, G. H. Paetzold, and M. Zampieri. 2021. SemEval-2021 task 1:

Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online, August. Association for Computational Linguistics.

Smith, L. N. 2018. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*.

Talec-Bernard, T. 2022. Is using an ai to simplify a scientific text really worth it. *Proceedings of the Working Notes of CLEF*.

Tang, Y., C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Truică, C.-O., A.-I. Stan, and E.-S. Apostol. 2022. Simplex: a lexical text simplification architecture. *Neural Computing and Applications*, pages 1–16.

UNE. 2018. Une 153101:2018 ex easy to read. guidelines and recommendations for the elaboration of documents.

Wilkens, R., B. Oberle, and A. Todirascu. 2020. Coreference-based text simplification. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 93–100.

Xu, W., C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Yimam, S. M., C. Biemann, S. Malmasi, G. H. Paetzold, L. Specia, S. Štajner, A. Tack, and M. Zampieri. 2018. A report on the complex word identification shared task 2018. *arXiv preprint arXiv:1804.09132*.

Yimam, S. M., S. Stajner, M. Riedl, and C. Biemann. 2017. Multilingual and cross-lingual complex word identification. In *RANLP*, pages 813–822.

Zhu, Z., D. Bernhard, and I. Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361.

Štajner, S., K. C. Sheang, and H. Saggion. 2022. Sentence simplification capabilities of transfer-based models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12172–12180, Jun.

A Prediction examples

Original	**Target**
(1) En todos ellos suscitas gran expectación en la comunidad científica, con la que colaboras en el departamento de medicina deportiva del Instituto Nacional de Educación Física. (In all of them you arouse great expectation in the scientific community, with which you collaborate with the sports medicine department of the National Institute of Physical Education.)	(1) En todos ellos suscitas gran curiosidad en la comunidad científica, con la que colaboras en el departamento de medicina deportiva del Instituto Nacional de Educación Física. (In all of them you arouse great curiosity in the scientific community, with whom you collaborate in the sports medicine department of the National Institute of Physical Education.)
(2) Además de todo lo anterior y como curiosidad, un 3% reconoció que quería conocer a sus nietos que habían nacido durante el confinamiento y a los que todavía no habían podido ver, y un 2% quería celebrar la boda de sus hijos o la Comunión de sus nietos que había sido suspendida. (In addition to all of the above and as a curiosity, 3% recognized that they wanted to meet their grandchildren who had been born during confinement and whom they had not yet been able to see, and 2% wanted to celebrate the wedding of their children or the Communion of their grandchildren that had been suspended.)	(2) Además de todo lo anterior y como curiosidad, un 3% reconoció que quería conocer a sus nietos que habían nacido durante el confinamiento y a los que todavía no habían podido ver, y un 2% quería celebrar la boda de sus hijos o la Comunión de sus nietos que había sido suspendida. (In addition to all of the above and as a curiosity, 3% recognized that they wanted to meet their grandchildren who had been born during isolation and whom they had not yet been able to see, and 2% wanted to celebrate the wedding of their children or the Communion of their grandchildren that had been suspended.)
Prediction	
(1) En todos ellos provocan gran interés en la comunidad científica, con la que colaboras en el departamento de medicina deportiva del Instituto Nacional de Educación Física. (In all of them, they provoke great interest in the scientific community, with which you collaborate in the sports medicine department of the National Institute of Physical Education.)	
(2.1) Aunque su madre falleció con el fin de conseguirlo han sido todos los nietos y mostrar los nietos que nos permitan a los niños y niñas, o incluso con quien lo sufre. (Although his mother passed away in order to get it have been all the grandchildren and show grandchildren that allow us to children, or even with the one who suffers it.) (2.2) Además de todo lo anterior y como curiosidad, un 3% reconoció que quería conocer a sus nietos que habían nacido durante el confinamiento y a los que todavía no habían podido ver, y un 2% quería celebrar la boda de sus hijos o la Comunión de sus nietos que había sido suspendida. (In addition to all of the above and as a curiosity, 3% recognized that they wanted to meet their grandchildren who had been born during isolation and whom they had not yet been able to see, and 2% wanted to celebrate the wedding of their children or the Communion of their grandchildren that had been suspended.)	

Table 9: Examples of model's predictions.

Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 123-133 recibido 15-12-2022 revisado 03-02-2023 aceptado 12-02-2023

Anticipating the Debate: Predicting Controversy in News with Transformer-based NLP

Anticipando el Debate: Prediciendo la Controversia en Noticias con PLN basado en Transformers

Blanca Calvo Figueras, Asier Gutiérrez-Fandiño, Marta Villegas
Barcelona Supercomputing Center
blanca.calvo@bsc.es

Abstract: Controversy is a social phenomenon that emerges when a topic generates large disagreement among people. In the public sphere, controversy is very often related to news. Whereas previous approaches have addressed controversy detection, in this work, we propose to predict controversy based on the title and content of a news post. First, we collect and prepare a dataset from a Spanish news aggregator that labels the news' controversy in a community-based manner. Next, we experiment with the capabilities of language models to learn these labels by fine-tuning models that take both title and content, and the title alone. To cope with data unbalance, we undergo different experiments by sampling the dataset. The best model obtains an 84.72 micro-F1, trained with an unbalanced dataset and given the title and content as input. The preliminary results show that this task can be learned by relying on linguistic and social features.
Keywords: nlp, controversy prediction, news, spanish.

Resumen: La controversia es un fenómeno social que ocurre cuando un tema genera desacuerdo entre los ciudadanos. En la esfera pública, la controversia se encuentra a menudo relacionada con las noticias de actualidad. Mientras que trabajos anteriores investigaron la detección de la controversia, en este trabajo nos proponemos predecirla basándonos en el título y el contenido de una noticia. En primer lugar, recogemos y curamos un conjunto de datos de un agregador de noticias en castellano que etiqueta las noticias según su controversia mediante las interacciones de la comunidad. Entonces, experimentamos con las capacidades de los modelos de lenguaje para aprender la categoría de controversia mediante el fine-tuneado de modelos que tienen el título y el contenido como contenido de entrada, y también con solo el título. Para lidiar con el desbalanceo de los datos, realizamos experimentos de sampleado de los datos. El mejor modelo obtiene una micro-F1 de 84.72, entrenado con un conjunto de datos desbalanceado y con el título y el contenido como entrada. Los resultados preliminares muestran que esta tarea puede ser aprendida mediante características lingüísticas y sociales.
Palabras clave: pln, predicción de la controversia, noticias, castellano.

1 Introduction

With the rise of digital media, public opinion has increasingly become a political actor (Kshetri and Voas, 2017). In digital spaces, citizens are able to publicly denote their stances and collectively define what topics do not offer consensus. To comprehend these opinion flows, researchers have focused on detecting controversy (Popescu and Pennacchiotti, 2010).

Controversial topics have been defined as topics that generate strong disagreement among large groups of people (Dori-Hacohen, Yom-Tov, and Allan, 2015). Controversial news should not be confused with fake news. While fake news are messages that carry false information (Kshetri and Voas, 2017), controversial news can not be proved false, although its content is being disputed by some. Controversy is a subjective category that lies between what a large group of people might consider true and others false, where there is no objective proof to deny either position. Operationally, whatever people conceive as controversial is controversial (Dori-Hacohen,

ISSN 1135-5948 DOI 10.26342/2023-70-10

Yom-Tov, and Allan, 2015).

From the natural language processing perspective (NLP), controversy detection has been approached as a text classification task (Dori-Hacohen, Jensen, and Allan, 2016). Previous work has focused on the identification of controversy by using edition features (Bykau et al., 2015), interaction features (Coletto et al., 2017), or comment-based features (Hessel and Lee, 2019). While detecting that something has been controversial is a potentially useful task for the study of news, in this work we propose predicting the controversy instead. By predicting we mean forecasting there will be a controversy before the controversy itself has even happened by using the only information we have available as soon as the news are out: its content. We believe this task can be very useful for raising alarms about potentially troublesome news. Our goal in this work is to investigate if controversy in news can be predicted using only the title and the content of the piece of news.

To achieve our goal, we gather a dataset from the news aggregation platform *Menéame*.[1] This website is driven by its community and moderated by senior users. Feedback mechanisms are in place to prevent false or wrong information from being distributed through the platform. While false information is removed, controversial news stay, although they are labeled with the tag *controversial*, with the goal of promoting a critical reading from users. We employ these tags as the annotations of our dataset and we experiment with fine-tuning different language models for the task. We also try different balancing strategies and explore the decisions taken by our best model.

The main contributions of this work are:

- We present a new approach for developing a controversy prediction dataset that matches our operational definition of controversy based on the algorithm of *Menéame*.
- We show that it is possible to predict the forthcoming controversy using mainly the title and the content of a news post.
- We investigate the relevance of linguistic features for the controversy prediction model.
- Finally, we make the best model available for the natural language community.[2]

In the following sections, we review the previous work on controversy detection (Section 2), we present our dataset and the labeling methodology (Section 3), we explain the models that have been trained on it (Section 4), and we display the primary results and an analysis based on explicability techniques (Section 5). Finally, we reflect on the need for this kind of work and propose future work to be done with these novel resources (Section 6).

2 Previous Work

Popescu and Pennacchiotti (2010) were the first to propose the detection of controversial events on social media. This idea was followed by other researchers, who modeled controversy through social media interactions (Coletto et al., 2017), sentiment analysis, and word matching (Sriteja, Pandey, and Pudi, 2017).

Other approaches investigated controversy in a collaboratively edited database (namely Wikipedia), by relying on the back-and-forth substitutions of content embedded within a similar context (Bykau et al., 2015). This challenge has been addressed as a clustering task (Dori-Hacohen, Jensen, and Allan, 2016) and as a classification task (Jang et al., 2016).

Controversy in the newswire domain was first approached by (Rethmeier, Hübner, and Hennig, 2018), who labeled user comments using up and down votes from other users, collecting 20.5k comments. More recent approaches have gone further and have used the labeled comments to predict the controversy of the post they are commenting on, using manually-labeled data from public forums (Hessel and Lee, 2019; Zhong et al., 2020). Finally, (Kim, 2019) created an explainable model by providing a descriptive sentence of the controversial topic, automatically generated from the comments.

Overall, controversy detection has been overlooked when compared to other NLP tasks in the area of information systems. Our work differentiates from all the previous approaches in both its data collection design and the inputs that we use to train the model. As in

[1] `https://www.meneame.net/`

[2] The model: `https://huggingface.co/PlanTL-GOB-ES/Controversy-Prediction`
The code: `https://github.com/PlanTL-GOB-ES/controversy-detection-model`

previous work, our work relies on community-labeled data, which is essential to identify a social phenomenon such as controversy. Furthermore, our work focuses on predicting controversy (as opposed to detecting it), for which reason we just provide the model with the title and the content of the piece of news.

3 Dataset

Given that there is no dataset for controversy detection in Spanish, we create our own by using available data and sampling it. We also analyze the dataset and give details on its statistics.

3.1 Nature of the Dataset

In this work, we have gathered a dataset of news posts from the platform *Menéame*. The internal design of this website,[3] which is conceived as a social network to promote healthy debate by allowing different views to converge and discuss, provides us with the possibility of labeling our data in an automated community-driven way. The gross dataset that we collected as of February 18th of 2022 contains a total of 236,969 posts.

Menéame is a news aggregator that compiles Spanish news based on users' suggestions. Users can publish the pieces of news they found interesting, and the rest of the users can vote and comment to decide if it is interesting enough to get them into the front page. To prevent the dissemination of fake news, spam, or other issues, the users can also report if there is a problem with the piece of news. The reactions of the users can trigger a warning algorithm that raises the alert sign. If the reports are well-grounded, moderators or the publishers themselves can decide to remove the content. In a middle ground between posts that are reported and posts that are finally removed, we find controversial posts. This is a temporary tag that the website gives to promote further consideration from the readers. After 30 hours, if the post is not removed, it just stays in the historical data of the platform but is marked as *controversial*.

The warning algorithm works in the following way.[4] One hour after the piece of news has been published, the algorithm starts checking the reactions to the post, and marks it as *controversial* if both of the following conditions are met:

- There are more than 4 negative votes or the negative votes represent more than 62.5% of the overall votes. This percentage keeps decreasing over time and is 10% after 6 hours.
- The average karma[5] of the users who voted negatively is higher than the average karma of the users who voted positively multiplied by 0.625. This ratio keeps decreasing over time and the average positive karma is multiplied by 0.1 after 6 hours.

We collected this boolean feature along with relevant metadata and the title and summary of the news documents. Some examples of controversial news are:

- *La pasta podría ser considerada verdura en los comedores escolares de EEUU.*
  Pasta could be considered a vegetable in the school canteens in the US.
  Negative votes: 24
  Positive votes: 129
- *Los ateos, mucho más inteligentes que los creyentes*
  Atheists are way smarter than believers.
  Negative votes: 40
  Positive votes: 331
- *Entramos en el caos de los test Covid a los profesores de Madrid: "Nos llevan como ovejas al matadero"*
  We get into the chaos of teacher's Covid tests in Madrid: "They are carrying us like sheep to the slaughter."
  Negative votes: 32
  Positive votes: 89

[3]The source code of this website is fully open-source in `https://github.com/Meneame/meneame.net`

[4]The code of the algorithm can be found here: `https://github.com/Meneame/meneame.net/blob/60fc5935e46fb72c47945abc63cd062803d030a8/www/libs/link.php#L1441`

[5]The karma of the users is a metric of the overall reliability of each user in the platform. Actions that raise someone's karma are: positive votes to their proposed posts, positive votes to news that are finally published, negative reports to news that are finally deleted, and positive votes to their comments. Actions that decrease someone's karma are: negative votes to their proposed news, negative votes to news that are not deleted after 30 hours, and negative votes to their comments. More detailed information can be found here: `https://github.com/Meneame/meneame.net/wiki/Karma`

Collection	Set	Total	Label
All	-	236,969	5,584 (C)
			231,385 (NC)
Sampled	-	20,386	5,584 (C)
			14,802 (NC)
Unbal.*	Train	18,270	4,950 (C)
			13,320 (NC)
	Valid	1,058	317 (C)
			741 (NC)
Balanced*	Train	9,900	4,950 (C)
			4,950 (NC)
	Valid	634	317 (C)
			317 (NC)
Test*	Test	1,058	317 (C)
			741 (NC)

Table 1: Dataset split counts. Collections marked with '*' are taken from "Sampled". "Sampled" comes from "All". C stands for controversy and NC for Not Controversy.

3.2 Sampling and Splitting

Table 1 shows the instance counts of the whole dataset. A total of 231,385 non-controversial posts and 5,584 controversial posts have been collected. The label imbalance of the whole dataset made the modeling difficult, as it ends up mainly predicting the label of the majority of the instances (non-controversial). To address this problem we undersampled the data and created the *Sampled* dataset, from which we defined a train, valid, and a test set (*Unbalanced* in Table 1). Additionally, we created a balanced set for train and valid, with the same number of instances in the two classes. The test set is shared among datasets.

3.3 Statistics

While most of our corpus comes from news of the general press, we also have instances from social networks, specialized press, blogs, satirical press, sports press, and fact-checking websites. Remarkably, general press has the lowest ratio of controversial posts, while social networks, satirical, and fact-checks are more often controversial. The data can be seen in Figure 1.

We obtained the top ten sources of news in Table 2. The first two sources are social networks, which exhibit that posts without publishing control are more likely to end up in controversy.

Although the collected data has abundant metadata, such as tags, topic, positive votes, negative votes, users' comments, clicks, promoting votes,[6] and karma, we only rely on the title and the summary of the content of the piece of news for our classification.[7] The reason behind this decision is to develop a model that can be used over any set of news, coming straight from the source. In this sense, we do not want to use any platform-specific feature for the prediction, but rather common features that are present on all kinds of news sites. Nevertheless, as an interesting insight, we discuss the relations between some of these metadata and our labels.

Source	Controv.	Total	Ratio
twitter.com	326	2,300	14.17%
youtube.com	306	5,955	5.13%
eldiario.es	293	7,533	3.88%
publico.es	190	5,804	3.27%
20minutos.es	118	6,048	1.95%
elconfidencial	117	5,188	2.25%
elplural.com	103	1,590	6.47%
huffingtonpost.es	85	786	10.81%
elmundo.es	81	6,191	1.30%
elespanol.com	81	1,995	4.06%

Table 2: Top ten sources, sorted by the number of controversial posts.

In Table 3, we show the ten most frequent tags given by the users to controversial news posts along with the total number of occurrences on the whole dataset. Remarkably, the most frequent labels are those related to politics. Some of them (i.e. "ayuso" and "vox") show a notably higher percentage than others.

Regarding the topics, we perform a similar analysis and show the distribution in Table 4. We decided to obviate the tenth topic as it is "rude language" and yet negligible in terms of controversy. Remarkably, "current issues" and "issues of social interest" are on the top of the list.

A point-biserial correlation was run to determine the relationship between the number of comments and the controversy label. There was a positive correlation between the variables ($r_{pb}29.90$, $p < 0.05$).

Finally, we also analyzed the number of likes and dislikes, as well as the difference between the likes and dislikes with respect

[6] What the website calls *meneos*.

[7] We use the summary of the piece, which is given by the users. In shorter posts, this is usually the whole content. In traditional news articles, this is almost always the first paragraph of the article. This is useful for training, as we have a length limitation given by the language model.

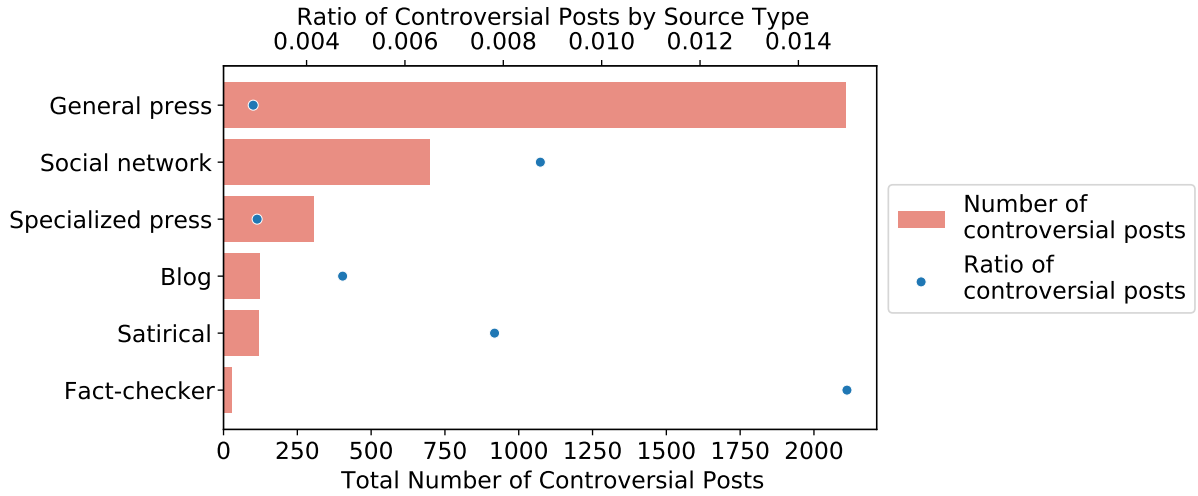


Figure 1: Controversial posts for each source type. The blue dots indicate the ratio between the number of controversial posts and the overall number of posts of each source type.

Tag	Explanation	Controversial	Total	Ratio
madrid	Spanish city/county	231	4,644	4.97%
pp	Political party	226	8,036	2.81%
españa	Spain	190	7,185	2.64%
vox	Political party	181	919	19.69%
podemos	Political party	173	1,515	11.41%
humor	Humour	170	3,726	4.56%
coronavirus	Coronavirus	153	2,596	5.89%
ayuso	Political leader	109	523	20.84%
psoe	Political party	84	2,505	3.35%
covid	Covid	74	1,174	6.30%

Table 3: The ten most common tags, sorted by the number of controversial posts.

to the controversy label. We run the point-biserial experiments and the results are the following:

- Positive votes: There is no correlation between the positive votes and the controversy label ($r_{pb}3.86$, $p < 0.05$).
- Negative votes: There is a strong positive correlation between the negative votes and the controversy label ($r_{pb}80.97$, $p < 0.05$).
- Positive-Negative difference: There is a negative correlation between the positive-negative difference and the controversy label ($r_{pb}10.51$,$p < 0.05$).

4 *Experimental setup*

To train a baseline for this task we selected the Spanish RoBERTa-base (Gutiérrez-Fandiño et al., 2022), as it has been trained on a large and clean corpus and it is the best performing model in Spanish to date.

Given that the training model only supports up to 512 input tokens, we used a truncation strategy for our data. When fine-tuning the model with title only, the truncation strategy does not apply, as titles are never long enough. In contrast, using the title and the summary concatenated,[8] the input data ends very often truncated.

We trained all the dataset combinations for 5 epochs, using a batch size of 4 per Graphical Processing Unit (GPU), a warmup of 0.06, a weight decay of 0.01, and a learning rate of 1e−5. For the optimizer, we chose Adam (Kingma and Ba, 2015), as it has been proved by the community to offer strong results.

The models were trained on our HPC premises on a machine with 2 IBM Power9 8335-GTH @ 2.4Ghz processors, 512GB of

[8]We concatenate with two light horizontal line symbols ("- -").

Topic	Explanation	Controversial	Total	Ratio
actualidad	current issue	2,671	59,673	4.47%
mnm	social interest	1,207	133,961	0.90%
ocio	leisure	616	7,378	8.34%
cultura	culture	577	21,492	2.59%
politica	politics	252	2,040	12.35%
tecnología	technology	170	8,779	1.93%
ciencia	science	18	1,156	1.55%
Podemos	political party	18	44	40.90%
Hemeroteca	news archive	12	96	12.50%

Table 4: The nine most common topics sorted by controversy.

Random Access Memory, and 4 NVIDIA V100 GPUs with 16GB of HBM2 memory.

5 Results and Analysis

The results of our fine-tuning experiments are shown in Table 5. They are displayed by dataset and by training setting.

Overall, our best model achieves a micro-F1 of 84.72 and an accuracy of 76.65, proving that the task can be effectively learned by a model with only the title and the summary. The addition of the summary does not provide much improvement, since using only the title in the unbalanced dataset already gives reasonably positive results. By contrast, the best model trained on the unbalanced dataset is around 4 points better on F1 than the model trained on the balanced one, showing that the model profits from more negative examples. We also experiment with other language models, such as mBERT (Devlin et al., 2019) and BETO (Cañete et al., 2020), using the same experimental setup and getting similar results.

Explainability analysis. To further analyze our results, we compute SHAP values (Lundberg and Lee, 2017), which assign contribution values to the tokens of each input.[9] We use the best model obtained, set up a SHAP explainer with it, and feed it a balanced set of 11k posts. Then, we aggregate the SHAP values for each token by part of speech (POS) and look at the open categories, namely: verbs, nouns, proper nouns, adjectives, and adverbs.[10] The used model for POS tagging shares the same vocabulary as our controversy model to ensure the tokenization is the same. We aggregate the positively and negatively contributing tokens for each POS tag and observe that proper nouns are, on average, the most contributing tokens to both the controversy and the non-controversy classes (Table 6).

Table 8 shows the 10 most influential words as per part of speech. When looking at proper nouns, we find that the most divergent parties of Spain (*Vox* and *Podemos*) correspond to the most controversial proper nouns of the dataset. Followed by *Ayuso* and *Pablo Iglesias*, two Spanish politicians well-known for having a large number of followers and haters and for being popular targets of memes (Paz, Mayagoitia-Soria, and González-Aguilar, 2021). On the other side of the table, contributing to the non-controversy category, we find technological companies, such as *Google*, *Linux*, and *Microsoft*, former Spanish presidents, like *Rajoy* and *Zapatero*, and geographical entities, such as *Reino Unido*, *Europa* and *Estados Unidos*. In sum, the use of proper nouns related to current politicians and parties is highly related to controversy, while the mention of companies, countries, or former politicians contributes negatively to the controversy class. Although this kind of information would also be highly valuable for humans trying to predict a controversy, it is very dependent on knowledge about the current cultural and political reality of Spain. To observe more enduring patterns, we train another model dropping the proper nouns.

Removing proper nouns. We use the same POS tagger to remove all proper nouns from the dataset and substitute them by the token *PROPN*. Then, we train a model with the exact same experimental setup as our best model: a Roberta-base model fine-tuned with the Unbalanced dataset and the

[9] Positive values mean that the token is contributing to the label "Controversy", while negative values mean a contribution towards Not Controversy. The furthest the value is from 0, the stronger the contribution.

[10] The part-of-speech model we are using can be found in `https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne-capitel-pos`.

Model	Dataset	Training setting	micro-F1	Accuracy	Time (s)
Roberta-base	Balanced	Title	0.7026	0.6295	1653
		Title + Summary	0.8093	0.7353	1267
	Unbalanced	Title	0.8197	0.7268	2631
		Title + Summary	**0.8472**	**0.7665**	2615
BETO	Unbalanced	Title	0.8398	0.7533	2568
		Title + Summary	0.8361	0.7429	2562
mBERT	Unbalanced	Title	0.8309	0.7287	3221
		Title + Summary	0.8347	0.7448	3277

Table 5: Model results by dataset and training setting.

POS	Positive SHAP	Negative SHAP
PROPN	0.0155	-0.0233
VERB	0.0053	-0.0122
ADV	0.0038	-0.0089
NOUN	0.0054	-0.0134
ADJ	0.0058	-0.0125

Table 6: SHAP values of the tokens, aggregated per part-of-speech.

POS	Positive SHAP	Negative SHAP
PROPN-token	0.0054	-0.0119
VERB	0.0060	-0.0120
ADV	0.0044	-0.0086
NOUN	0.0063	-0.0147
ADJ	0.0067	-0.0134

Table 7: SHAP values of the tokens aggregated per part-of-speech for the model without proper nouns. The PROPN-token value corresponds to the substitution token we used.

Title+Summary input. The obtained results are remarkably good, with an F1 of 83.53 and an accuracy of 74.76. The aggregated SHAP values in Table 7 show that the impact of the PROPN-token is much lower than the aggregated proper nouns were, and the rest of the POS categories have increased in relevance. These results suggest that controversy can be predicted by relying mainly on linguistic features.

We identify some linguistic patterns in the table with the top influencing tokens by POS for this new model (Table 9). When looking at verbs, we observe that actions in the third singular person of the perfect tense (e.g. *ha hecho*, *ha sido*, *ha convertido*, *ha respondido*, etc.) are often associated with controversial posts. Instead, verbs in the simple present tense (e.g. *es*, *hay*, *pide*, *tienen*, etc.) are associated with non-controversial posts. Looking at the posts of our dataset, we identify that while the third singular person of the perfect tense is often used to speak about people in a rather informal tone, like in the sentence "*La portavoz adjunta de Compromís Mónica Oltra ha vuelto a lucir esta mañana en el primer pleno ordinario del nuevo periodo de sesiones de Les Corts una de sus famosas camisetas.*" (This morning Monica Oltra has worn again another of her famous t-shirts in the Parliament); the present simple is used for more factual information, like in the sentence "*Los trabajadores de RTVE rodean la mesa de edición con carteles que dicen: #Vergüenza #Vergonya*" (The workers from RTVE surround the edition office with signs that say: #Shame).

Finally, we observe some other linguistic patterns that seem to indicate controversy, such as the use of adverbs at the beginning of the sentences, and the use of adjectives indicating political positioning (e.g. *feminist* or *independentist*).

6 Conclusion and Future Work

In this work, we have built a dataset for controversy prediction in Spanish and we have characterized it in many dimensions. Controversy has been labeled in a community-driven manner, which matches the operational definition of controversy itself, given in the introduction.

With this dataset, we have experimented creating two different collections: a balanced one and an unbalanced one. In this particular experiment, we have shown that the amount of samples is more important than balancing the labels.

The models we trained have provided positive results and have shown that they can effectively capture the insights of the title (and the summary). These models can be used

as predictors of the controversy generated in news. In the explainability analysis, we have shown that this model is capturing differences in the linguistic register of controversial posts, as well as the social and political reality of Spain. We have been able to highlight some characteristics of the controversial register, such as using the third singular person of the perfect tense.

In the future, this model can be used for media monitoring, by trying to understand how controversy evolves in media as a function of time. It can also be used for media analysis, by running it against online media to observe editorial inclinations toward controversy. Additionally, one could analyze how the perception of controversy evolves in Spanish society, as what generates controversy today might not be controversial in the future, or the other way around. This dataset will continue to grow, as the community behind it is still highly active. We set as a future goal to keep expanding it to capture possible shifts in the perception of a controversy.

Finally, the model could also be used to highlight disputed posts as soon as they are published, as this has been suggested as a mitigation strategy for the impact of disinformation (Dori-Hacohen, 2015). In this line, previous work has indicated certain relation between highly disputed news and fake news (Shu et al., 2019). We suggest executing this model against a fake news database to study this phenomenon.

Acknowledgements

This work has been funded by the Spanish State Secretariat for Digitalization and Artificial Intelligence (SEDIA) within the framework of the Plan-TL, and the IBERIFIER project funded by the European Union (action number 2020-EU-IA-0252).

References

Bykau, S., F. Korn, D. Srivastava, and Y. Velegrakis. 2015. Fine-grained controversy detection in Wikipedia. In *2015 IEEE 31st International Conference on Data Engineering*, pages 1573–1584, April. ISSN: 2375-026X.

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*.

Coletto, M., K. Garimella, A. Gionis, and C. Lucchese. 2017. Automatic controversy detection in social media: A content-independent motif-based approach. *Online Social Networks and Media*, 3-4:22–31, October.

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.

Dori-Hacohen, S. 2015. Controversy Detection and Stance Analysis. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, page 1057, New York, NY, USA, August. Association for Computing Machinery.

Dori-Hacohen, S., D. Jensen, and J. Allan. 2016. Controversy Detection in Wikipedia Using Collective Classification. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, SIGIR '16, pages 797–800, New York, NY, USA, July. Association for Computing Machinery.

Dori-Hacohen, S., E. Yom-Tov, and J. Allan. 2015. Navigating Controversy as a Complex Search Task. page 5.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, and M. Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68(0):39–60.

Hessel, J. and L. Lee. 2019. Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features. 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 1648–1659, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Jang, M., J. Foley, S. Dori-Hacohen, and J. Allan. 2016. Probabilistic Approaches to Controversy Detection. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 2069–2072, New York, NY, USA, October. Association for Computing Machinery.

Kim, Y. a. 2019. Unsupervised Explainable Controversy Detection from Online News. *Proceedings of the European Conference on Information Retrieval*.

Kingma, D. P. and J. Ba. 2015. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kshetri, N. and J. Voas. 2017. The Economics of "Fake News". *IT Professional*, 19:8–12, November.

Lundberg, S. M. and S.-I. Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., pages 4765–4774.

Paz, M. A., A. Mayagoitia-Soria, and J.-M. González-Aguilar. 2021. From Polarization to Hate: Portrait of the Spanish Political Meme. *Social Media + Society*, 7(4):205630512110629, October.

Popescu, A.-M. and M. Pennacchiotti. 2010. Detecting controversial events from twitter. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1873–1876, New York, NY, USA, October. Association for Computing Machinery.

Rethmeier, N., M. Hübner, and L. Hennig. 2018. Learning Comment Controversy Prediction in Web Discussions Using Incidentally Supervised Multi-Task CNNs. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 316–321, Brussels, Belgium, October. Association for Computational Linguistics.

Shu, K., L. Cui, S. Wang, D. Lee, and H. Liu. 2019. dEFEND: Explainable Fake News Detection. *In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 395–405, July.

Sriteja, A., P. Pandey, and V. Pudi. 2017. Controversy Detection Using Reactions on Social Media. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 884–889, November. ISSN: 2375-9259.

Zhong, L., J. Cao, Q. Sheng, J. Guo, and Z. Wang. 2020. Integrating semantic and structural information with graph convolutional network for controversy detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 515–526, Online, July. Association for Computational Linguistics.

POS	Word	Explanation	SHAP	Word	Explanation	SHAP
	explica	explains	0.40	es	is	-7.10
	desmonta	dismantles	0.34	son	are	-3.15
	responde	answers	0.32	hay	there are	-2.94
	muestra	shows	0.31	tiene	has	-2.77
VERB	explicando	explaining	0.24	era	was	-2.25
	habla	talks	0.24	pide	asks	-2.22
	Desmontando	Dismantling	0.19	tienen	have	-2.00
	recuerda	remembers	0.18	dice	says	-1.68
	voy	coming	0.15	está	is	-1.60
	ha respondido	has answered	0.14	hacer	do/make	-1.57
	Vox	political party	6.05	Google		-2.45
	Podemos	political party	5.96	PP	political party	-2.43
	Pablo Iglesias	politician	5.76	Zapatero	ex-politician	-1.82
	Ayuso	politician	5.30	Reino Unido	UK	-1.69
PROPN	Madrid		3.88	Rajoy	ex-politician	-1.54
	VOX	political party	1.65	Linux		-1.51
	Pedro Sánchez	politician	1.28	SGAE	institution	-1.45
	Isabel Díaz Ayuso	politician	1.07	Europa		-1.32
	Pablo Casado	politician	0.79	Estados Unidos	United States	-1.30
	Ada Colau	politician	0.65	Microsoft		-1.28
	Así	Like this	1.26	no	no	-21.78
	Cómo	How	0.76	más	more	-15.17
	Cuando	When	0.31	hoy	today	-3.68
	Además	Morover	0.30	también	also	-2.34
ADV	Aquí	Here	0.16	muy	very	-2.18
	literalmente	literally	0.09	después	after	-1.88
	Anoche	Last night	0.07	ahora	now	-1.82
	consecuentemente	consequently	0.06	ya	already	-1.72
	brutalmente	brutally	0.06	donde	where	-1.54
	sexualmente	sexually	0.06	casi	almost	-1.44
	vídeo	video	3.80	años	years	-7.93
	derecha	right	0.80	millones	milions	-5.74
	mentiras	lies	0.60	Gobierno	Government	-5.07
	bulo	fakes	0.54	presidente	president	-3.56
NOUN	discurso	discourse	0.50	personas	people	-3.16
	ultraderecha	far-right	0.44	euros	euros	-2.87
	izquierda	left	0.44	ministro	minister	-2.73
	respuesta	answer	0.43	mundo	world	-2.57
	tuit	tweet	0.41	juez	judge	-2.48
	monarquía	monarchy	0.38	Policía	Police	-2.38
	feminista	feminist	0.67	nuevo	new	-2.31
	extrema	extreme	0.44	gran	big	-1.93
	independentista	independentist	0.37	mayor	older/higher	-1.85
	independentistas	independentists	0.28	nueva	new	-1.71
ADJ	ultraderechista	far-rightist	0.27	pasado	past	-1.48
	española	Spanish	0.26	Nacional	National	-1.44
	morada	purple	0.25	grandes	big	-1.42
	ultraderechistas	far-rightists	0.24	últimos	last	-1.31
	mediática	media	0.23	mejor	better	-1.21
	política española	Spanish politics	0.22	general	general	-1.18

Table 8: Top 10 influencing words per Part-of-Speech with no named entities.

POS	Word	Explanation	SHAP	Word	Explanation	SHAP
	muestra	shows	1.04	es	is	-5.25
	desmonta	dismantles	0.49	hay	there is	-2.61
	responde	answers	0.47	pide	asks	-1.86
	ha hecho	has done	0.45	tienen	have	-1.62
VERB	explica	explains	0.42	son	are	-1.61
	ha sido	has been	0.35	hace	does	-1.51
	ha convertido	has converted	0.34	era	was	-1.41
	ha respondido	has answered	0.28	pagar	pay	-1.32
	ha declarado	has declared	0.27	Fallece	Dies	-1.24
	ha publicado	has published	0.26	tiene	has	-1.24
	Así	Like this	1.84	no	no	-16.52
	Además	Morover	1.11	más	more	-12.99
	Cómo	How	0.76	hoy	today	-5.26
	Sin	Without	0.54	ahora	now	-1.41
ADV	Cuando	When	0.41	ayer	yesterday	-1.23
	Ahora	Now	0.33	muy	very	-1.21
	No	No	0.25	antes	before	-1.17
	No obstante	Nevertheless	0.25	menos	less	-1.09
	Sin embargo	Nevertheless	0.24	sólo	only/just	-1.09
	Después	After	0.16	casi	almost	-1.00
	vídeo	video	4.31	años	years	-6.83
	partido	party	2.12	ministro	minister	-5.51
	respuesta	answer	1.70	millones	milions	-4.66
	líder	leader	1.51	Gobierno	Government	-4.36
NOUN	formación	formation	1.22	ciudad	city	-3.40
	bulo	fake	0.98	mundo	world	-3.35
	discurso	discourse	0.96	personas	people	-2.73
	mensaje	message	0.86	países	countries	-2.71
	pandemia	pandemic	0.83	país	country	-2.64
	redes	networks	0.82	presidente	president	-2.42
	feminista	feminist	0.85	mayor	older	-2.01
	independentista	independentist	0.67	nuevo	new	-1.83
	morada	purple	0.50	nueva	new	-1.68
	oficial	official	0.45	Europea		-1.55
ADJ	ultraderechista	far-rightist	0.44	gran	big	-1.50
	falso	false	0.43	grandes	big	-1.37
	independentistas	independentists	0.42	general	general	-1.35
	extrema	extreme	0.37	Civil	Civil	-1.17
	madrileña	from Madrid	0.36	Nacional	National	-1.15
	madrileño	from Madrid	0.33	nuevas	new	-1.07

Table 9: Top 10 influencing words per Part-of-Speech in the model with no proper names.

Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 135-144 recibido 15-12-2022 revisado 31-01-2023 aceptado 10-02-2023

The state of end-to-end systems for Mexican Spanish speech recognition

El estado de los sistemas end-to-end *para el reconocimiento de voz del Español de México*

Carlos Daniel Hernández-Mena[1], **Ivan Vladimir Meza Ruiz**[2]
[1]Language and Voice Laboratory, Reykjavík University.
[2]Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas,
Universidad Nacional Autónoma de México.
carlosm@ru.is, ivanvladimir@turing.iimas.unam.mx

Abstract: Current end-to-end speech recognizer systems report an excellent performance for Spanish. However, this is not reported for specific variants. Moreover, it is unclear if there would be a benefit in creating a fine-tuned version for a particular variant. To investigate these aspects, particularly for Mexican Spanish, we evaluate four different of-the-shelf speech recognizers (one commercial and three open-source); additionally, we fine-tune two systems for Mexican Spanish. We evaluate read and spontaneous speech, present an error analysis and show that fine-tuning for a variant decreases the error rate. As a result of our experimentation, we build two new systems available to the community.
Keywords: speech recognition, acoustic models, mexican spanish.

Resumen: El desempeño actual de los reconocedores de voz se reporta como notablemente bueno para el español, sin embargo, no se especifica el desempeño para variantes especificas, y sobretodo no se establece si existe un beneficio de crear una versión ajustada explicitamente a una variante particular. Para investigar estos aspectos, y especificamente para el español de México, nuestro trabajo evalua el desempeño de cuatro sistemas de reconocimiento de voz (uno comercial y tres de código abierto); adicionalmente creamos dos versiones especificas al español de México mediante la técnica de *fine-tuning*. Se evaluan los sistemas en voz leída y espontanea, presentamos un análisis de error y mostramos que ajustando los sistemas actuales con la variante todavía se puede reducir el error. Como resultado de la experimentación se obtuvieron dos nuevos sistemas que se hacen disponibles a la comunidad.
Palabras clave: reconocimiento de voz, modelos acústicos, español de México.

1 Introduction

Recent progress in end-to-end speech recognition has shown that the performance of Spanish is among the best ones (Radford et al., 2022). In addition, current practices in sharing models and their associated systems have made this technology accessible to a larger pool of potential users. However, it is crucial to notice that this advancement has been reached through the extensive use of private resources and the surge of multilingual and self-supervision settings. This situation makes it difficult to understand the nuances of the field's current state, particularly for individual languages and their variants. It also complicates researching possible improvements to the existing approaches. In this work, we shed light on the current state of Mexican Spanish speech recognition. We use several available language resources and fine-tune well-established speech recognition models to understand the current state better.

We focus on Mexican Spanish, one of the language's most spoken variants, which is spoken predominately in Mexico (Hernández-Mena et al., 2017; Pineda et al., 2010b) and extensively in the United States. It makes

ISSN 1135-5948 DOI 10.26342/2023-70-11

use of 24 phonemes. Table 1 shows the phonetic repertoire of Mexican Spanish in terms of the points of articulation for consonant and vowel sounds. In particular, two phonemes in Mexican Spanish are characteristic of the variant. They are related to the influence of the Nahuatl language (indigenous language of Central Mexico), and they are used in everyday speech (Hernández-Mena and Herrera-Camacho, 2014; Cuétara Priede and others, 2004; Hernández-Mena, 2019): /ʃ/ as in "Xolos" (or "shadow" in English), /t͡l/ at the end of words as "Popocatépetl" (with no counterpart in English). In comparison, the Spain variant of the Spanish language also uses 24 phonemes. The two phonemes that are different are /θ/ and /ʎ/ (for more details see (Quilis, 1993)).

The contribution of this research is to present an evaluation of four different of-the-shelve systems and two fine-tuned versions to quantify the current speech recognition performance for the Mexican Spanish variant. These fine-tuned systems are made freely available. Our evaluation focuses on reading and spontaneous speech; for a general evaluation of Spanish, we include other variants, which include Latin-American, Spain and US-based dialects. First, section 2 presents the work done for Mexican Spanish speech resources and the current rise of end-to-end Speech recognition. Next, in section 3 we present the corpora used in fine-tuning and our evaluation. In section 4, we present the experimentation, results and error analysis to highlight the different system behaviors. Finally, section 5 presents our main findings.

2 Previous work

Since the second half of the nineties, there have been corpora that included or consisted of Mexican Spanish speech recordings. These resources have been commonly used to model the acoustic properties of speech; the resulting model is frequently referred to as *acoustic model*. There have been three primary strategies behind the efforts to collect the data for the acoustic models:

1. To explicitly record samples of Latin-American speakers for speaker identification or speech recognition.
2. To focus only on Mexican speakers.
3. Projects that collected spoken audio in Spanish and later were processed to work as data for speech recognition.

Table 2 list the corpora that was reported and available originally in an academic context: HUB4-NE, a Spanish broadcast news corpus (Consortium, 1997; Fiscus et al., 2001); VoxForge, a corpus of reading speech collected through Internet volunteers (Voxforge.org, 2006); DIMEx100, a phonetic balanced reading speech corpus by Mexican Speakers (Pineda et al., 2010a); DIMEx100 niños, is a version of DIMEx100 where the speakers are children (Moya et al., 2011); Golem-Universum, contains spontaneous interactions of children with a dialogue system system (Venegas-Brione, Meza-Ruiz, and Pineda, 2011); LATINO-40, a corpus of Spanish reading news (Bernstein et al., 1995); West Point Heroico, a corpus of spontaneous speech from Mexican and non-native Spanish speakers (Morgan, 2006); Fisher Spanish, a collection of spontaneous telephone calls (Graff et al., 2010); Hispanic, a collection of reading recordings (Byrne et al., 2014); CIEMPIESS, a spontaneous Mexican Spanish Corpus (Hernández-Mena and Herrera, 2015); CIEMPIES Light, an updated version of CIEMPIESS corpus (Hernández-Mena and Herrera, 2017); CIEMPIESS Balance, a corpus to gender balance CIEMPIESS (Hernández-Mena, 2018); CIEMPIESS experimentation, a version of CIEMPIESS to develop speech recognition systems, it includes CIEMPIES Test for testing Mexican Spanish spontaneous speech (Hernández-Mena, 2019a); LibriVox[1], corpus based on book readings (Hernández-Mena, 2020);Wikipedia grabada[2], corpus of readings of Wikipedia articles (Hernández-Mena and Ruiz, 2021); TEDx, collection of TED talks in Spanish (Hernández-Mena, 2019b). Table 2 summarizes the sizes and availability of the different corpora[3].

Acoustic resources, speech recordings and transcriptions became more relevant with the advent of end-to-end systems, which avoided two traditional sources of informa-

[1]LibriVox website `https://librivox.org/` (last visited April 2022).

[2]Wikipedia *grabada* website `https://es.wikipedia.org/wiki/Wikiproyecto:Wikipedia_grabada` (last visited April 2022).

[3]For further detail about these corpora see (Hernández-Mena et al., 2017) and (Mena and Meza-Ruiz, 2022).

	Points of articulation						
Manners of articulation	**Consonants**	**Labial**	**Labiodental**	**Dental**	**Alveolar**	**Palatal**	**Velar**
	Voiceless Stop	p		t			k
	Voiced Stop	b		d			g
	Voiceless Affricate					t͡ʃ	
	Voiceless Fricative		f		s	ʃ	x
	Voiced Fricative					ʝ	
	Nasal	m			n	ɲ	
	Rhotic				ɾ/ r		
	Lateral				l	t͡l	
	Vowels				**Front**	**Central**	**Back**
	Close				i		u
	Mid				e		o
	Open					a	

Table 1: Phonetic repertoire of Mexican Spanish (Hernández-Mena et al., 2014).

Corpora	**Hours**	**Year**	**Av.**	**Modality**	**Variants**
LATINO-40	6.8h	1995	Cost	Read	Latin-American
HUB4-NE	31h	1997	Cost	Spontaneous	US
CALLHOME Spanish	13*h*	1997	Cost	Spontaneous	Latin-American
DIMEx100	6.1h	2004	Req.	Read	Mexican
VoxForge	50h	2006	Free	Read	Mix
West Point Heroico	16.6h	2006	Cost	Both	North-American
Fisher	163h	2010	Cost	Spontaneous	Latin-American
DIMEx100 niños	8h	2011	Unk.	Read	Mexican
Golem-Universum	0.2h	2011	Unk.	Read	Mexican
CIEMPIESS	17h	2015	Free	Spontaneous	Mexican
CHM150	1.6h	2016	Free	Spontaneous	Mexican
CIEMPIESS Light	18h	2017	Free	Spontaneous	Mexican
CIEMPIESS Balance	18h	2018	Free	Spontaneous	Mexican
CIEMPIESS Experimentation	40h	2019	Free	Spontaneous	Mexican
TEDx Spanish	24h	2019	Free	Spontaneous	Mix
LibriVox Spanish	73h	2020	Free	Read	Mix
Wikipedia Spanish	25h	2021	Free	Read	Mix
Mozilla Common Voice Spanish	320h	2022	Free	Read	Mix

Table 2: Corpora that include Mexican Spanish for the development of speech recognizers (in bold, those that only focus on Mexican Spanish; Av., Availability; Unk., unknown; Req., by request).

tion required by previous approaches: pronunciation dictionaries and language models. Pronunciation dictionaries are a list of words with other corresponding computer-based phonetic transcription; language models are probabilistic models that determine the probability of a sequence of words. Although these last ones nowadays are heavily used to re-score the output of end-to-end systems (Wang, Wang, and Lv, 2019). In particular, end-to-end speech recognition relies on deep neural networks to relate segments of the acoustic signal with the character sequence of transcription (Hannun et al., 2014) and on the CTC loss function (Graves et al., 2006), which collapse the sequence of characters and compare it to the correct transcription during training.

Another recent advancement in the field consisted of using self-supervision settings to train models that rely on the acoustic signal to create better sound representations than what traditional acoustic models can reach (Schneider et al., 2019). In addition, self-supervision allows reaching good performance in multilingual settings. In this setting, first, a model is self-trained with speech recordings from multiple languages without the need for transcriptions; for instance, it is trained to predict the next segment of the

audio signal; later, this model gets fine-tuned using an end-to-end setting to perform speech recognition (Conneau et al., 2020). This arrangement was the backbone during the creation of the Whisper system, which became state-of-the-art in the field (Radford et al., 2022). As a result, Whisper reaches a 5.4% word error rate performance for Spanish, the best-reported performance for the language up to this moment.

3 Systems and datasets

For our experiments, we selected four out-of-the-shelf systems:

Google speech recognizer[4] This is a commercial system widely adopted in Mexico which supports Latin America variants. However, its parts and training are not publicly shared.

Quarztnet[5]: It is an implementation of a Quarztnet architecture (Kriman et al., 2020) on the NeMo platform developed by NVIDIA (Kuchaiev et al., 2019; Fidjeland et al., 2009). This model was first trained with several English corpora to be later fine-tuned using the Spanish Common Voice Mozilla corpus. This is an example of transfer learning using a pre-trained model.

Wav2vec[6]: Model based on the XLSR-53 system (Conneau et al., 2020) train in a multilingual setting. In particular, this model was also fine-tuned with the Spanish Common Voice Mozilla corpus.

Whisper[7]: Model trained on 680,000 hours of several languages recordings, including Spanish (Radford et al., 2022). There are five pre-trained versions of this system which vary in size: *tiny*, *base*, *small*, *medium* and *large*.

In addition to the pre-trained systems, we fine-tuned two new models:

Quartznet fine-tuned[8] based on a Spanish Quarztnet model described above[9].

Wav2vec fine-tuned[10] based on a XLSR wav2vec large model[11].

Both fine-tuned systems were further trained with 944h of predominantly Mexican Spanish. For the Mexican Spanish, we use the corpora: CIEMPIESS Light, CIEMPIESS Balance, CIEMPIES FEM, CHM150, TEDx Spanish, DIMEX100, DIMEX100 niños, Golem-Universum, VoxForge, LIBRIVOX Spanish, WIKIPEDIA Spanish, Spanish Mozilla Common Voice 10.0, West Point Heroico, LATINO-40, CALLHOME Spanish, HUB4NE Spanish, FISHER Spanish. Additionally, we also incorporate two private collections called *Tele con Ciencia* (28h16m) and extra recordings from another private collection of Mexican recordings (118h22m), which can not be shared given their copyright status. At the fine-tuning stage, we also added Spanish variants corpora: the Spanish portion of MediaSpeech (10h) citemediaspeech2021, Spanish form Spain (6h40m) (Garrido et al., 2013), Chilean (7h08m), Colombian (7h34m), Peruvian (9h13m), Argentinian (8h01m) and Puerto Rican (1h00m) all of these corpora are part of the project *Crowdsourcing Latin American Spanish* (Guevara-Rukoz et al., 2020).

During testing we use Mozilla Common Voice Speech (MCVS) since given its accessibility, results on MCVS are commonly reported; however, since their modality is read speech, the performance tends to be better than expected for spontaneous speech. To contrast, we have evaluated performance in three spontaneous speech corpora: the HUB4-NE, CALLHOME, and CIEMPIESS. We also isolated the Mexican speakers from the MCVS and evaluated performance on this segment of this corpus. We also report

[4]Website describing system: `https://cloud.google.com/speech-to-text/` (last visited November 2022).

[5]Website for the pre-trained STT Es Quartznet15x5 model: `https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_es_quartznet15x5` (last visited November 2022).

[6]Website for Fine-tuned XLSR-53 large model for speech recognition in Spanish: `https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-spanish` (last visited November 2022).

[7]Website for whisper model: `https://github.com/openai/whisper` (last visited November 2022).

[8]Fine-tuned model available at: `https://huggingface.co/carlosdanielhernandezmena/stt_es_quartznet15x5_ft_ep53_944h` (last visited November 2022).

[9]Description of the model: `https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_es_quartznet15x5` (last visited November 2022).

[10]Fine-tuned model available at: `https://huggingface.co/carlosdanielhernandezmena/wav2vec2-large-xlsr-53-spanish-ep5-944h` (last visited November 2022).

[11]Description of the model: `https://huggingface.co/facebook/wav2vec2-large-xlsr-53`

our development results on the MCVS and CALLHOME which include this partition.

4 Experiments and results

Table 4 shows the Word Error Rate (*WER*) results of ten versions of the four systems (the lower the result, the better). As it can be seen, the *wav2vec* system performs the best in six of the eight testing scenarios, while *Whisper large* performs the best in the rest (two). Notice that both versions of *wav2vec* are fine-tuned, *W2V* originally was fine-tuned using the Mozilla Spanish Corpus (includes read Mexican Spanish). In contrast, our fine-tuned version was trained with the collection of predominately Mexican Spanish corpora (spontaneous Mexican Spanish). The benefit of this fine-tuning can be appreciated within the results for the CIEMPIESS test corpus, which consists of Mexican Spanish recordings; this system reaches the best performance for spontaneous speech: 11.17 of WER. Also, it reaches a new best score for the HUB4-NE with 7.48. On the other hand, the *Whisper* system consistently gets a competitive performance and scores two of the best performances for the Mozilla Common Voice Speech test corpora. Remember, this is a large model which relies on more than half a million training hours in a multilingual setting, including Spanish speech.

Another aspect to consider with the new *Whisper* system is the longer time it takes to transcribe. As expected, the larger the model, the more parameters and time to transcribe. This can be seen in Table 3, which records more than 7 hours for the large version of the system. As a point of comparison, the other systems took no longer than 30 minutes on the same amount of data. Their performance was so consistent among themselves that we did record them. However, when Whisper took too long, we started to record it.

In Table 4, we also notice some drawbacks when fine-tuning a model. While it would be logical that fine-tuning the model using data closest to the target will improve the performance, this does not always happen since the performance of the original model could be degraded. For instance, *Quartznet* had an excellent performance for the Mozilla Common Voice, but this became worst with the fine-tuning. This effect has been noticed previously (Huang et al., 2020). Sometimes, the fine-tuning degrades the performance from a previously scored performance. However, the positive impact can be noticed with the rest of the testing corpora in which the fine-tuned version produces fewer mistakes. We hypothesize that this is related to the "closeness" of the variant. By fine-tuning, it stops being close to reading Mexican speech and gets closer to the spontaneous version.

Another interesting aspect is the difficulty associated with the CALLHOME corpus (associated with the worst performances). Our experience points out that this is a problematic corpus. A preliminary analysis of the transcriptions shows a challenging setting where it is common to find overlapping speech among speakers. Additionally to this, there is no suitable transcription protocol for such cases.

4.1 Error analysis

Word Error Rate (WER)is based on edit-distance operations: *insert*, *delete*, and *replace*. WER quantifies the percentage of operation to transform the expected output (reference transcription) into the system's transcription (hypothesis). Figure 1 shows the percentage of *insertions* per word and normalized by the occurrence of such term in the system transcription. These percentages were ranked from larger to lower. A sound system should start transitioning from words for which all occurrences were inserted (1.0) to terms for which a low percentage of the occurrences were inserted. The *wav2vec* system has fewer insert operations in this figure (green line); it is followed by Whisper medium and large (grey and light pink lines). This can be interpreted as these systems being less eager to propose words. Insertions could be viewed as an acoustic hallucination (the system "listen" to a word which is not there). Table 5 shows some examples of hallucinated words and their frequencies (between parentheses). In total hallucinations are in the hundreds, but most words get inserted only once. One source of error is the insertion of single letters, which is expected partly because these systems are end-to-end and allow bits of the signal to relate to a bit of transcription, even though it does not relate to a word.

On the other hand, Figure 2 shows the ranking of the *deletions* per word and normalized by the reference corpus. Similarly to

Corpus	Tiny	Base	Small	Medium	Large
MCVS dev	1h47m	1h41m	2h48m	5h23m	7h51m
MCVS test	1h47m	1h46m	2h40m	5h25m	8h2m

Table 3: Whisper time to transcribe 15k audios from the Mozilla Common Voice Speech corpora *test* and *dev* portions. The runs were done using NVidia GeForce GTX Titan X GPU.

System	Go.	QN	W2V	Whisper					fine-tuned	
Corpora				Tn.	Bs.	Sm.	Med.	Lg.	QN	W2V
HUB4-NE	17.79	22.87	12.84	29.27	22.84	15.63	11.92	10.82	14.48	**7.48**
MCVS dev	17.68	12.85	**4.12**	33.75	20.89	10.27	6.49	5.86	15.97	8.02
MCVS Mex dev	17.81	11.72	**4.69**	32.0	20.32	9.90	6.66	5.87	14.96	7.59
MCVS test	19.59	14.89	8.70	37.02	23.32	11.73	7.58	**6.80**	17.99	9.20
MCVS Mex test	21.84	14.29	8.04	33.82	21.75	11.7	7.92	**6.89**	16.33	8.93
CALLHOME dev	52.93	78.68	61.45	91.76	74.52	53.37	44.42	41.44	56.34	**40.39**
CALLHOME test	51.92	78.07	60.29	86.43	70.18	50.32	41.91	39.25	55,43	**39.12**
CIEMPIESS test	18.19	36.69	23.16	28.59	22.17	18.28	15.10	15.25	18.57	**11.17**

Table 4: Word error rate for evaluation corpora (lower the better; Go., Google QN, Quartznet; W2V, wav2vec; Tn, tiny;Bs, base; Sm., small; Med., medium; Lg., Large).

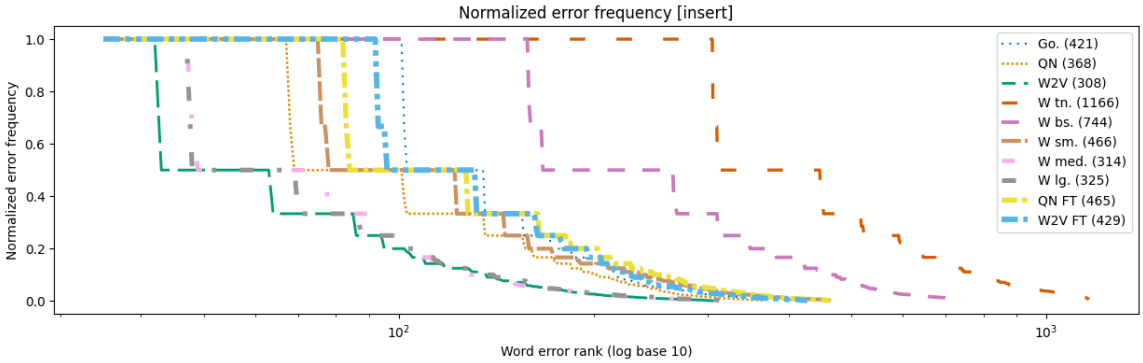


Figure 1: Ranked normalized *insertions* frequency per word (lower the better; Go., Google QN, Quartznet; W2V, wav2vec; FT, fine-tuned; Tn, tiny;Bs, base; Sm., small; Med., medium; Lg., Large).

the *insertions*, the faster the system transitions from a high percentage (1.0) to lower, the better performance. Here we can see that both fine-tuned systems have the fewer deletions, *wav2vec* (blue line) and *QuartzNet* yellow line.

Figure 3 shows the normalized frequencies per word replaced. This operation is harder to normalize because it can be interpreted as a combination of a deletion and an insertion and is anchored to two words, the deleted and the inserted. To normalize, we use the higher count of any of the words: inserted or deleted. This is related to replacement operations being higher than insertions and deletions. In order to gain further understanding regarding the different systems, we calculate the number of operations per word; this is a proxy to learn how different the words in the hypothesis transcription are compared to the references. Figure 4 shows the histograms of several edits on each word to replace. As can be noticed, most terms need to be replaced by a word very close in spelling and only different in one edit. The *W. Lg.* is the system that better performs with 7,735 replacements. However, remember this system implies more significant transcription times, so it might not be a reasonable cost-effective compromise.

Reflecting on the performance of the systems, the best ones *Whisper Large* and *W2V fine-tuned* performance are very similar, but they have different behaviours. *W2V fine-tuned* takes more risks proposing words (higher insertion error), but the proposed ones are usually correct (lowest deletion error). On the other hand, the *Whisper Large*

System	Total	Exaples
Go.	90	post(2), auto(2), 16(2), refuerzo(2), 70(2), reorganizar(1), pos(1),...
QN	127	l(5), n(3), s(3), auto(3), digamo(2), dy(2), qu(2), tam(2), d(2), ...
W2V	86	l(6), n(4), mas(2), pos(2), d(2), puras(2), fracean(1), your(1), metodo(1), ...
W Tn.	565	os(6), auto(4), p(4), estero(4), gabriel(2), s(2), tr(2), l(2) ...
W Bs.	287	auto(3), transcribe(3), estimar(2), os(2), juris(2), pura(2), tango(2), s(2), ...
W Sm.	180	qte(5), auto(3), post(2), agarró(2), pura(2), mecánicas(2), extra(2), eh(2), ...
W Med.	72	high(3), pura(2), post(1), método(1), lacktut(1), delan(1), ow(1), ...
W Lg.	110	agarró(2), pura(2), manny(2), auto(1), método(1), hertz(1), papito(1), ...
QN FT	164	n(8), s(4), piro(4), l(4), d(3), g(2), bum(2), payz(2), epaíses(1), escribier(1), ...
W2V FT	116	s(8), h(8), n(5), pos(3), eh(3), l(3), pura(2), d(2), método(1), cl(1), seso(1), ...

Table 5: Example of words transcribed by the systems but not present in the reference transcription.

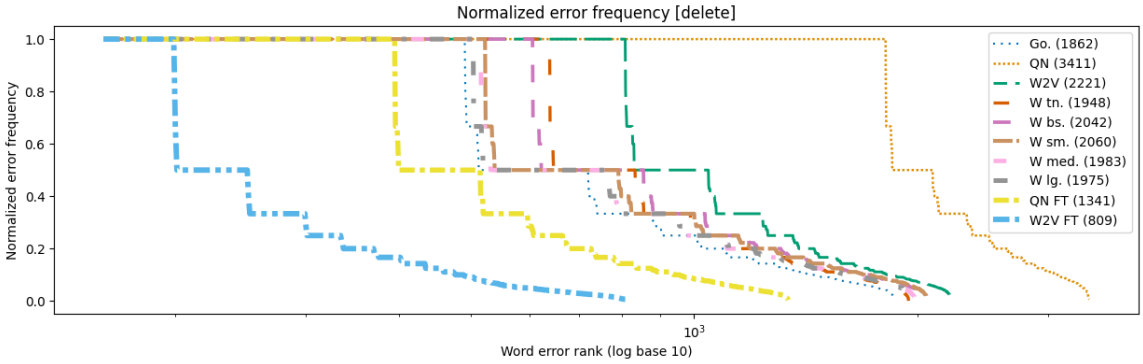


Figure 2: Ranked normalized *deletions* frequency per word (lower the better; Go., Google QN, Quartznet; W2V, wav2vec; FT, fine-tuned; Tn, tiny;Bs, base; Sm., small; Med., medium; Lg., Large).

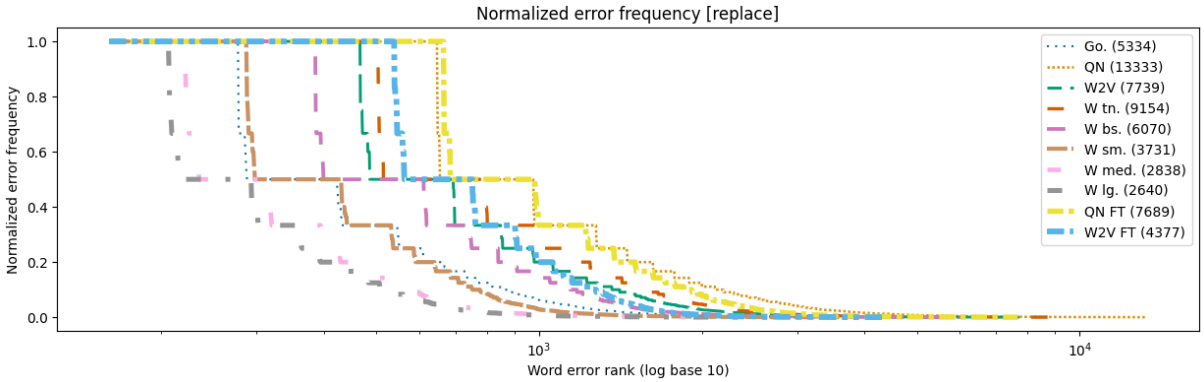


Figure 3: Ranked normalized *replace* frequency per word (lower the better; Go., Google QN, Quartznet; W2V, wav2vec; FT, fine-tuned; Tn, tiny;Bs, base; Sm., small; Med., medium; Lg., Large).

is shier when proposing words, so it omits several words (higher deletion error, but lowest insertion error). This characteristic of taking more risks when proposing words is also observed in the replacement, on which *W2V fine-tuned* also gets a higher rate of errors. However, it seems to be a good strategy for a speech recognizer since it performs better. The code for the analysis is freely available[12].

5 *Conclusion*

We have presented an evaluation of speech recognisers for the Mexican variant of Spanish. For several decades there has been

[12]Code for the error analysis in speech transcription: `https://github.com/ivanvladimir/speech_transcriptions_analysis` (last visited March 2023)

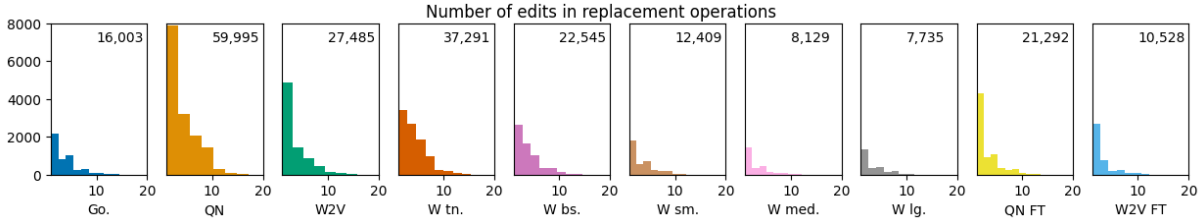


Figure 4: Histograms of Levenshtein distances between replaced words (the lower, the better; a sum of all the distances is in the upper right corner; it follows the same colours than 3).

an effort to support the creation of language resources focused on this variant. Although new methods such as end-to-end speech recognition facilitate the construction of multilingual settings and have helped increase the performance of the current systems, in this work, we show that fine-tuning for a specific variant still has its benefits. In our experience, we will continue to recommend the collection of language resources for specific variants and the fine-tuning based on pre-trained models.

On the other hand, there are still open questions regarding these adaptations. First, our observations must be confirmed on new multilingual general models recently released, such as *Wav2vec XLR* or new *Whisper* versions, which unfortunately rely on much more computer power. Second, we believe it would be important to the development of speech technology to have more diversity of variants with their corresponding resources. However, at this moment, there is no clear answer on how to mix these variants to reach a good performance for most speakers without losing performance during the fine-tuning process. We also would like to create fine-tuned versions of specific variants without including other ones to quantify the effect of variants and sub-variants and the support the rest of the variants can provide.

Acknowledgments

Authors thank Mónica Alejandra Ruiz López for verifying and correcting the transcriptions of the CIEMPIESS Test corpus. Carlos Hernández-Mena thanks the support from Language and Voice Laboratory from Reykjavik University in the realization of this manuscript.

References

Bernstein, J., B. Grundy, E. Rosenfeld, A. Najmi, and P. Mankoski. 1995. Latino-40 spanish read news ldc95s28. CD.

Byrne, W., E. Knodt, J. Bernstein, and F. Emami. 2014. Hispanic-english database ldc2014s05. CD.

Conneau, A., A. Baevski, R. Collobert, A. Mohamed, and M. Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Consortium, L. D. 1997. 1997 spanish broadcast news speech (hub4-ne) ldc98s74. Web download.

Cuétara Priede, J. O. et al. 2004. *Fonética de la ciudad de México. Aportaciones desde las tecnologías del habla*. Tesis-UNAM.

Fidjeland, A. K., E. B. Roesch, M. P. Shanahan, and W. Luk. 2009. Nemo: a platform for neural modelling of spiking neurons using gpus. In *2009 20th IEEE international conference on application-specific systems, architectures and processors*, pages 137–144. IEEE.

Fiscus, J. G., J. S. Garofolo, M. Przybocki, W. Fisher, and D. Pallett. 2001. 1997 hub4 broadcast news evaluation non-english test material ldc2001s91. Web download.

Garrido, J. M., D. Escudero, L. Aguilar, V. Cardeñoso, E. Rodero, C. De-La-Mota, C. González, C. Vivaracho, S. Rustullet, O. Larrea, et al. 2013. Glissando: a corpus for multidisciplinary prosodic studies in spanish and catalan. *Language resources and evaluation*, 47(4):945–971.

Graff, D., S. Huang, I. Cartagena, K. Walker, and C. Cieri. 2010. Fisher spanish speech ldc2010s01. CD.

Graves, A., S. Fernández, F. Gomez, and J. Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent

neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Guevara-Rukoz, A., I. Demirsahin, F. He, S.-H. C. Chu, S. Sarin, K. Pipatsrisawat, A. Gutkin, A. Butryna, and O. Kjartansson. 2020. Crowdsourcing Latin American Spanish for Low-Resource Text-to-Speech. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 6504–6513, Marseille, France, May. European Language Resources Association (ELRA).

Hannun, A., C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition.

Hernández-Mena, C. D. 2019. The CIEMPIESS proper-names pronouncing dictionary. In *Corpus presented at OpenCor 2019 Conference, Guanajuato City, Mexico. Available online at https://opencor.gitlab.io/corpora-list/*, page 1.

Hernández-Mena, C. D. and J.-A. Herrera-Camacho. 2014. Ciempiess: A new open-sourced mexican spanish radio corpus. In *LREC*, volume 14, pages 371–375.

Hernández-Mena, C. D., I. Meza-Ruiz, J. Herrera-Camacho, et al. 2017. Automatic speech recognizers for Mexican Spanish and its open resources. *Journal of Applied Research and Technology*, 15(3):259–270.

Hernández-Mena, C. D. 2018. Ciempiess balance ldc2018s11.

Hernández-Mena, C. D. 2019a. Ciempiess experimentation ldc2019s07.

Hernández-Mena, C. D. 2019b. TEDx Spanish Corpus. Audio and transcripts in Spanish taken from the TEDx Talks; shared under the CC BY-NC-ND 4.0 license. Web Download.

Hernández-Mena, C. D. 2020. Librivox spanish ldc2020s01.

Hernández-Mena, C. D. and A. Herrera. 2015. Ciempiess ldc2015s07.

Hernández-Mena, C. D. and A. Herrera. 2017. Ciempiess light ldc2017s23.

Hernández-Mena, C. D. and I. V. M. Ruiz. 2021. Wikipedia spanish speech and transcripts ldc2021s07.

Huang, J., O. Kuchaiev, P. O'Neill, V. Lavrukhin, J. Li, A. Flores, G. Kucsko, and B. Ginsburg. 2020. Cross-language transfer learning, continuous learning, and domain adaptation for end-to-end automatic speech recognition. *arXiv preprint arXiv:2005.04290*.

Kriman, S., S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang. 2020. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6124–6128. IEEE.

Kuchaiev, O., J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook, et al. 2019. Nemo: a toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*.

Mena, C. D. H. and I. Meza-Ruiz. 2022. Creating mexican spanish language resources through the social service program. In *Proceedings of the 2nd Workshop on Novel Incentives in Data Collection from People: models, implementations, challenges and results within LREC 2022*, pages 20–24.

Morgan, J. 2006. West point heroico spanish speech ldc2006s37. Web Download.

Moya, E., M. Hernandez, L. Pineda, and I. Meza. 2011. Speech recognition with limited resources for children and adult speakers. In *2011 10th Mexican International Conference on Artificial Intelligence*, pages 57–62. IEEE.

Pineda, L. A., H. Castellanos, J. Cuétara, L. Galescu, J. Juárez, J. Llisterri, P. Pérez, and L. Villaseñor. 2010a. The corpus dimex100: transcription and evaluation. *Language Resources and Evaluation*, 44(4):347–370.

Pineda, L. A., H. Castellanos, J. Cuétara, L. Galescu, J. Juárez, J. Llisterri, P. Pérez, and L. Villaseñor. 2010b. The Corpus DIMEx100: transcription and evaluation. *Language Resources and Evaluation*, 44(4):347–370.

Quilis, A. 1993. *Tratado de fonología y fonética españolas*, volume 2. Gredos Madrid.

Radford, A., J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *OpenAI Blog*.

Schneider, S., A. Baevski, R. Collobert, and M. Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *INTERSPEECH*.

Venegas-Brione, E., I. Meza-Ruiz, and L. A. Pineda. 2011. Evaluation of a dialogue system for children based on an interaction-oriented cognitive architecture. *Procesamiento del lenguaje natural*, pages 113–120.

Voxforge.org. 2006. Free speech... recognition (linux, windows and mac) - voxforge.org. `http://www.voxforge.org/`.

Wang, D., X. Wang, and S. Lv. 2019. An overview of end-to-end automatic speech recognition. *Symmetry*, 11(8):1018.

Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 145-155 recibido 16-12-2022 revisado 31-01-2023 aceptado 07-02-2023

Widaug. Data augmentation for named entity recognition using Wikidata

Widaug. Aumento de datos para el reconocimiento de entidades nombradas usando Wikidata

Pablo Calleja, Alberto Sánchez, Oscar Corcho
Ontology Engineering Group
Universidad Politécnica de Madrid
{p.calleja,o.corcho}@upm.es

Abstract: The current state of the art of Natural Language Processing models are based on the use of a big amount of data to be trained. The more, the better. However, this is quite a limitation in the creation of datasets for specific natural language processing tasks such as Named Entity Recognition, which involves one or more annotators to read, understand and annotate those required named entities along a corpus. Currently, there are many good general domain corpora for the English language. However, particular domains or scenarios and other non-English languages are still not so represented in the research community. Thus, data augmentation techniques are explored to create synthetic data similar to the originals to enrich the training process of the models. On the other hand, knowledge graphs contain a lot of valuable information that is not being used to help in the data augmentation process. This work proposes a data augmentation method based on the Wikidata knowledge graph which is tested in a Spanish corpus for a Named Entity Recognition challenge.
Keywords: Named Entity Recognition, data augmentation, Wikidata.

Resumen: El estado del arte actual de los modelos de Procesamiento de Lenguaje Natural se basa en el uso de una gran cantidad de datos para ser entrenados. Cuantos más, mejor. Sin embargo, esto es una gran limitación en la creación de conjuntos de datos para tareas específicas de procesamiento de lenguaje natural, como el reconocimiento de entidades nombradas, que involucra a uno o más anotadores para leer, comprender y anotar las entidades nombradas requeridas a lo largo de un corpus. Actualmente, hay bastantes corpus buenos de dominio general para el inglés. Sin embargo, los dominios o escenarios particulares y otros idiomas distintos del inglés aún no están tan representados en la comunidad de investigación. Por ello, se exploran técnicas de aumento de datos para crear datos sintéticos similares a los originales para luego enriquecer el proceso de entrenamiento de los modelos. Por otro lado, los grafos de conocimiento contienen muchísima información valiosa que no se está utilizando para ayudar en el proceso de aumento de datos. Este trabajo propone un método de aumento de datos basado en el grafo de conocimiento de Wikidata que es evaluado en un corpus español para un desafío de reconocimiento de entidades nombradas.
Palabras clave: Reconocimiento de Entidades Nombradas, aumento de datos, Wikidata.

1 Introduction

Named Entity Recognition (NER) is a Natural Language Processing (NLP) task that consists of identifying named entities in the text and classifying them. Traditionally, those name entities are proper names of persons, locations, organizations or miscellaneous proper names (Grishman and Sundheim, 1996; Tjong Kim Sang, 2002). Nowadays, the original classification groups are extended and adapted to particular domains or to scenarios of interest. For instance, the biomedical domain has defined its own relevant classification groups such as diseases, chemical compounds, DNA, etc. (Perera,

ISSN 1135-5948 DOI 10.26342/2023-70-12

Dehmer, and Emmert-Streib, 2020; Asghari, Sierra-Sosa, and Elmaghraby, 2022). Moreover, recent challenges and corpora propose to identify new classification groups, such as drugs (Li, Zhang, and Zhou, 2020), complex named entities (Malmasi et al., 2022) or software mentions (Schindler et al., 2021). Although the classification groups are highly different between them and the instances could not be proper names, the task is still based on the identification of particular terms or nouns that belong to a particular classification group.

Currently, the best results obtained in Named Entity Recognition are usually based on pretrained language models[1] which have been fine-tuned for the task. Although the fine-tuning process requires less data to achieve good results, models still need a training corpora as large as possible to learn from it. However, the creation of annotated corpora is a process that requires a huge human effort that involves reading, understanding, and annotating particular entities in the text. Also, more than one annotator per document is usually required to achieve a good quality dataset, in which the inter-annotation agreement is validated.

The lack of data in certain domains makes it impossible to create efficient models for a NER task. An example of this is the case of the SmarTerp project (Rodriguez et al., 2021), a project to help interpreters in simultaneous interpretation contexts in the European Parliament. This project required the implementation of a named entity recognition system for 'important' words within a specific scope. Entities such as locations or dates are required but also specific terms about the topic being discussed, such as 'types of soil for cultivation' or 'types of metal forging procedures'. This type of 'important' named entities were specific for each European Parliament session related to the topic which was discussed and in different European languages. However, due to the ambiguity of the problem and the lack of prepared annotated corpora, the development of a specific annotated corpus was one of the most difficult tasks.

Beside the problem of lack of data in certain scenarios, the language problem is also added. Other non-English languages are underrepresented in terms of corpora and models. A common approach to solve this gap is the proposal of challenges and shared tasks for different languages. For instance, in the last years the Spanish community have released datasets for challenges in particular domains. Two of the most important ones have been LivingNER challenge (Farré-Maduell et al., 2022) which is focused on the identification of living entities providing a corpus of 2000 annotated clinical cases for training, development and testing, and CANTEMIST (Miranda-Escalada, Farré, and Krallinger, 2020), which is focused on the identification of tumor morphology providing a corpus of 1301 annotated oncological clinical cases for the same three tasks. Both challenges have been an important improvement on the Spanish language community, but they are still small compared to others that English language models use in other scenarios.

Data augmentation is a technique that is used to generate new synthetic data based on the modification of the original data in those cases where there are not enough samples to train a machine learning model and to achieve better results. This technique has been used and adapted in different research areas. In the NLP context, data augmentation is usually based on adding noise (removing/adding words) to original sentences, adding synonyms or moving words to other positions (Erd et al., 2022; Dai and Adel, 2020). However, there are not so many works that exploit knowledge graphs to acquire structured data to enrich the data augmentation process.

The objective of the work consists in the creation of a method named Widaug using information extracted from the common well-known Wikidata knowledge graph. The target of this method is to cover scenarios such as the Smarterp project in which just a few sentences could be annotated by experts, making it impossible to create a representative corpus. Our hypothesis is that the proposed data augmentation method can improve the performance of the training model better than other traditional techniques, specifically for small corpora, relying on the knowledge provided by Wikidata.

For the evaluation of the method, a set of experiments have been performed on the second task of ProfNER's challenge (Miranda-Escalada et al., 2021) which proposes a na-

[1] https://paperswithcode.com/sota/token-classification-on-conll2002

med entity recognition problem for the recognition of 'professions' within tweets related to the COVID-19 pandemic. The code and the experiments can be found in our public GitHub repository.[2]

The paper is structured as follows: Section 2 details the state-of-the-art of data augmentation and Section 3 presents the approach of the method, the use case and the experiments performed. Section 4 evaluates the results and, finally, Section 5 achieves the conclusions and future work.

2 State of the art

Data augmentation has been a widely research area that has been involved in many different tasks in which machine learning models have significance such as computer vision or, in this particular, for NLP tasks. The basic idea is to take partial data or related data as seeds to create a larger dataset to train a machine learning model. With more data, the model will be able to generalize better and obtain better results.

In NLP there are some common techniques such as replacement of words. Replacement is one of the first techniques used in data augmentation. External resources such as Wordnet (Zhang, Zhao, and LeCun, 2015) or Word2vec (Wang and Yang, 2015) are used for synonym replacement. Recent approaches (Wu et al., 2019) used pretrained language models to replace words that are suitable in the position of the word in the original sentence, by doing masks. Other works do mention replacement; in the training data the entities are replaced for others from a manually created dictionary that contained entities that were not part of the training data (Liu et al., 2020).

The Easy data augmentation techniques (EDA) were presented for text classification (Wei and Zou, 2019), which are based on synonym replacement, random insertion, random swap and random deletion. Other work extended EDA techniques for NER tasks using the UMLS knowledge base (Kang et al., 2021). Currently, there are libraries such as NLPAug[3] or TextAugment (Marivate and Sefara, 2020) that facilitate the implementation process for most of these techniques.

Back translation is also a common technique which is based on the retranslation of content from a target language back to its source language. The purpose in data augmentation is to get similar sentences with changes in some words that have been modified in the translation process. It can be used for topic classification (Xie et al., 2020) or sentiment analysis classification (Luque, 2019). Additionally, this approach has been tested for NER tasks (Yaseen and Langer, 2021). In the Spanish language, there are works that have presented back translation techniques for text classification problems (Luo, 2021; Guzman-Silverio, Balderas-Paredes, and López-Monroy, 2020)

In addition, text generation or sentence generation is a technique in which new synthetic sentences are created using language models or generative models to extend the original data. There are works that use this technique on text classification tasks (Bayer et al., 2022) or NER tasks (Ding et al., 2020).

Related to knowledge graphs, works that have explored the use of Wikidata (Raiman and Miller, 2017) have proposed a general scheme to do mention replacement for the general types (person, location, dates, etc.) for a Question Answering task using the instances of the general types represented in Wikidata. Other works (Kim, Kim, and Kang, 2022) used Wikidata to extract aliases of named entities with the label *Also known as* which are used for mention replacement.

The method proposed in this work combines the sentence generation approach using the information represented in the Wikidata concepts, using its labels and the most important relations to create new sentences in combination with sentences extracted from the Wikipedia pages of the concepts. Moreover, back translation and mention replacement approaches are tested.

3 Methodology

This section presents the use case in which the data augmentation method has been tested, the proposed data augmentation method using Wikidata, the experiments design and how the models have been fine-tuned with the augmented data.

3.1 Use case

The method for data augmentation has been tested for the ProfNER challenge (Miranda-Escalada et al., 2021). This challenge is part of the Social Media Mining for Health

[2] https://github.com/oeg-upm/widaug

[3] https://github.com/makcedward/nlpaug

(SMM4H), an initiative that seeks the application of machine learning methods for the extraction of information in social networks and its use in the health sector. The ProfNER-ST challenge, in particular, seeks to identify professions and occupations on social networks in Spanish within the healthcare field.

This challenge is particularly relevant in the context of the COVID-19 pandemic and its repercussions on mental health. Therefore, the annotated data were obtained using a web crawler on Twitter using keywords such as 'Covid-19', 'epidemic' or 'confinement'. The main aim is to use Natural Language Processing to identify vulnerable occupations in this context. This vulnerability can be both direct (health professionals in the first line of contact) and indirect (professions such as drivers, guards, carers, etc.).

Specifically, the use case has focused on the second task of the challenge, which seeks the identification and classification of professions in tweets related to the COVID-19 pandemic. The challenge provides the training, development and test corpus sets. The train set contains around 6,000 annotated tweets and the validation set contains around 2,000 annotated tweets.[4] As the gold annotations from the test set are not released, the development set is used for the evaluation of the data augmentation method.

The types of the named entities are: *PROFESION* (profession) which are entities referred to a profession that provides a salary such as 'doctor' or 'driver', *SITUACION_LABORAL* (employment situation) which are entities referred to an specific working condition such as 'worker' or 'self-employed', *ACTIVIDAD* (activity) which are unpaid works such as 'volunteer' and *FIGURATIVA* (figurative) which are used to mention metaphoric works such as 'joker' or 'pseudo journalist', usually used as sarcasm or jokes.

In total, there are 2,597 entities classified as PROFESSION (2,163), WORKING SITUATION (349), ACTIVITY (61) and FIGURATIVE (24). This work has focused only on those entities of type 'PROFESSION', which is the most representative named entity type in the corpus. The other types have not been considered by their ambiguity and their under-representation in the corpus. Of the 2,161 profession named entities, 1,532 are single words such as 'president' and 631 are multiword terms that refer to a profession such as 'bus driver' or 'national policeman'.

Moreover, the training and development corpus have been cleaned in order to avoid emojis, urls, hashtags and other characters outside of the scope of the utf-8 chars recognized by the tokenizer of the language model. Also, sentences with less than four tokens without any named entity annotated are removed.

3.2 Proposed method

The general approach of the Widaug method is presented in Figure 1. The method needs a corpus with annotated named entities, the language of the corpus and the target type of the named entities that will be augmented. The method performs the following tasks.

First, the method extracts the entities from the corpus that belong to the target type. Then, the method queries Wikidata to obtain instances of the target type. This search is performed by querying for instances of the concept in the graph that corresponds to the target type. The relation of the instance is represented by the relation '*instanceOF*' (wdt:P31). The instances are stored and tagged as candidate named entities.

The next step is to perform a filter in order to obtain only those named entity candidates close to the domain of the original annotated named entities extracted in the first step. Wikidata represents a huge amount of information that even the instances of a similar class could not have the same semantic meaning for the target domain of the corpus. The filter is carried out with a word embedding method, in which a candidate has to be up to 70 % similar, using the cosine similarity, for more than one original named entity. Capturing entities with more than one 70 % similarity can represent a trend or partial topic for similar named entities and avoids outliers. This value has been considered based on previous studies (Rekabsaz, Lupu, and Hanbury, 2017) and a preliminary study in which clearly unwanted terms over 50 % of similarity (not actual works such as prefect of Rome or controversial works such as sexual works) are analyzed to be below the selected threshold.

As word embeddings pretrained models do

[4] https://zenodo.org/record/4563995#.Y5cuPuzML0o

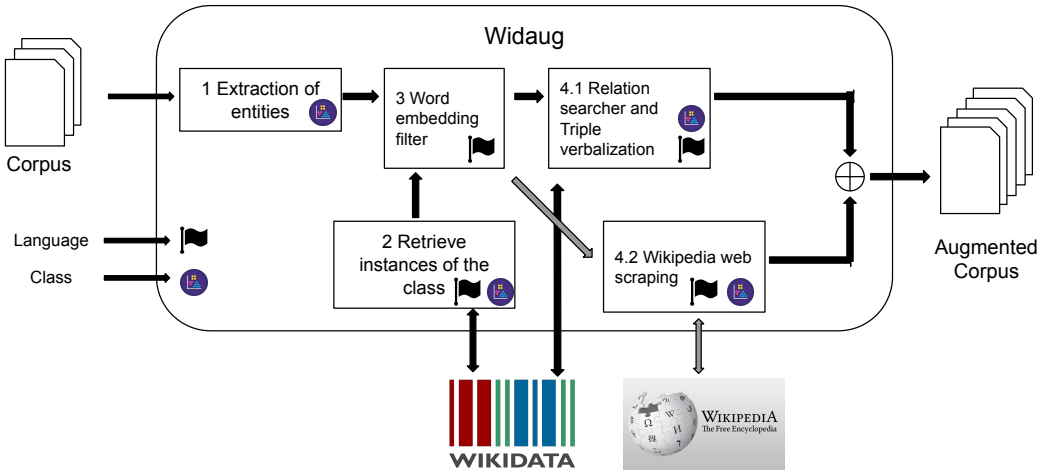


Figura 1: Overview of the data augmentation method Widaug.

not represent multi-word terms, the embedding filter is performed with the principal noun of the term of the candidate over the rest of principal nouns of the original named entities.

Once the candidates have been filtered, the proposed method exploits the information from the Wikidata concept, based on its properties and relations, and verbalizes it. The property to be found is *schema:description*, which provides a short description of the concept in natural language, and the following relations: P279 (subclass_of), P1056 (produce), P2283 (use) and P425 (field of occupation). The description property and the related concepts require to have their label property in the target language to be extracted. For instance, if a concept contains descriptions but not in the required language, the property is not extracted. The selection of these three relations apart from subclass_of is based on a previous study to find the three most repeated relations with labels in the required language in all the candidates. This method could be extended for other named entity types.

Then the properties and relations are verbalized. An example of this is shown in Table 1 with the Wikidata concept of baker for the English language. For the description, the verb 'to be' is used to join the named entity with its description. The named entity is annotated with the tag of the target type. In the rest of the relations, more than one element can be represented. For instance, in the example of the relation 'use' for baker, three elements are retrieved (heat, oven and bakery). The verbalization is used using the same verbs that represent the relation and the concept is also annotated. In the case of subclass of, the elements retrieved are also tagged with the same type (because they are subclasses). At this moment, the developed method covers the verbalization of sentences for the English and Spanish languages.

Finally, the last task consists of generating sentences based on web scrapping. This approach consists in capturing sentences from the Wikipedia page of each candidate (as most of the concepts have one) in order to create well-structured natural language sentences. Only sentences that contain the named entity are captured. Finally, those sentences are tokenized and the named entity is labeled according to its classification.

3.3 Experiment design

For the evaluation of the method, different experiments have been performed. First, the original training corpus has been randomly sampled at 10, 30, 50 and 100 %. The idea is to evaluate the performance of the method with different subsets of the original data and with the complete corpus such as Erd's evaluation (Erd et al., 2022). Then, four data augmentation methods are performed over the prepared corpora: the proposed augmentation method based on Wikidata (Widaug), a mention replacement method, a back translation method and finally, a combination of Widaug and back translation. All the augmented corpora are used to fine-tune a Spanish language model for the Named Entity Recognition task (a.k.a. token classification

	Examples								
Description	baker	is	the	persons	who	prepares	or	sells	bread
	B	O	O	O	O	O	O	O	O
Subclass of	baker	is	a	type	of	artisan	-	-	-
	B	O	O	O	O	B	-	-	-
Produce	baker	produces	bread	-	-	-	-	-	-
	B	O	O	-	-	-	-	-	-
Use	baker	uses	heat	oven	and	bakery	-	-	-
	B	O	O	O	O	O	-	-	-
Field of occupation	baking	is	the	field	of	occupation	of	baker	-
	O	O	O	O	O	O	O	B	-

Tabla 1: Example of verbalization of relations of the Wikidata concept 'baker' in English language. Tag 'B' represents the label B-Profession.

task). Moreover, the corpus without augmentation is used as a base line. The target of the experiments is to show how the methods perform augmenting the data for training and to check if it is possible to reach higher performance rather than without them. Figure 2 shows the overview of the experiments.

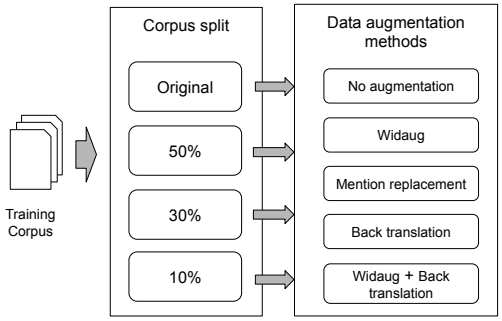


Figura 2: Experiments performed.

The proposed Widaug method has been adapted as follows. First, the target language needed is Spanish. Then, the target classification group is 'profession' which is represented in Wikidata as the concept wd:Q28640, which has 2,720 instances. The filtering process is performed with a FastText Spanish word embedding model (Bojanowski et al., 2016). After the filtering process, the candidate named entities are reduced to 852. The process of augmenting data from Wikidata generated 532 new annotated sentences and the generation of sentences from Wikipedia 855 new sentences.

The mention replacement method has been configured as follows. The 852 previously filtered candidates have been used to replace the mentions (named entities) in the corpus to generate new sentences, based on the work of Jonathan Raiman (Raiman and Miller, 2017). All the sampled corpus (10 %, 30 % and 50 %) have been augmented until the size of the original corpus (100 %). For the 100 % corpus, it has been augmented a 50 % more.

The back translation method is based on the library *BackTranslation*[5]. For each sentence of the corpus that contains an annotated named entity, a new sentence is generated translating it to English language and back to Spanish language. Finally, a combination of the Widaug method and the back translation method is proposed to evaluate the performance of one of the most representative methods for Spanish language in combination with the proposed one.

3.4 Fine-tuning process

For the fine-tuning process, the *Google Collaboratory* platform has been used to use GPUs for acceleration. Usually, a Tesla T4 GPU is given to train the models. The different trainings have been carried out for six epochs, which has been seen to be the point in which the original training corpus does not improve more.

The language model used is the MarIA (Gutiérrez-Fandiño et al., 2022) model developed at the Barcelona Supercomputing Center (BSC) with the database of the *Biblioteca Nacional Española* (National Library of Spain). Currently, MarIA models are the

[5]https://pypi.org/project/BackTranslation/

best models in terms of performance publicly available for Spanish language. For instance, the Rigoberta model (Serrano et al., 2022) claims to outperform MarIA results, but is not public.

This model is based on the RoBERTa architecture and the dataset contains 570 GB of cleaned training data. Although there are several different models, the model to use would be the *base*[6] model, which has 12 layers, 768 hidden layers and 125M parameters.

Hyperparameters of the training model have been 16 of batch size, 500 warm-up steps, 0.01 of weight decay and 1e-4 of learning rate and the models have been trained with the Huggingface transformers library. All models are evaluated with the validation corpus of the challenge.

4 Evaluation

This section details the evaluation results obtained from the experiments. As a traditional Named Entity Recognition task or token classification problem, the metrics used in the evaluation are precision, recall and f-measure. In addition, a discussion and an analysis error are performed.

4.1 Obtained results

Table 2 shows the results obtained for each portion of the original corpus (10, 30, 50 and 100 %) and for each data augmentation approach (mention replacement, back translation, Widaug and the combination of Widaug and back translation). Additionally, the results of the training corpus without data augmentation are presented as a baseline. The results presented for each combination of corpus and approach is the best result obtained within the 6 epochs of training.

The results obtained for the not augmented data (No Aug) for all created corpus (10, 30 and 50 %) have the lower values of all experiments in terms of the F-measure. However, none of the data augmentation methods have improved the results obtained for the original training corpus. Data augmentation methods have added noise to the training process.

In contrast, the data augmentation methods (Mention Replacement (MR), Back Translation (BT) and Widaug) have improved all the results over the baseline for all corpus, being Widaug the best performance method over the rest, followed by back translation. However, the final experiment in which Widaug is combined with back translation (BT) has not improved the overall results; only in the 30 % corpus is slightly improved (0.02).

4.2 Discussion and error analysis

The results obtained confirm the hypotheses of the work, that the proposed data augmentation method can improve the performance of the training model better than other traditional methods, specifically for small corpora such as the 10 % corpus which has an improvement of 0.11 in the F-measure. However, an error analysis has been performed and studied to understand the behaviour of the trained models.

First, the corpus presents some limitations. The size of the original training corpus is quite small (12,707 sentences), which has been reduced in the cleaning process to 11,050 sentences, with entities tended to repeat themselves a lot such as *sanitario* (sanitary) and *guardia civil* (civil guard). Also, the corpus is comprised of sentences of tweets, which contains typos and unstructured information (e.g., several mentions to public charges and services to advise them). Therefore, the training process with a generic language model with this corpus has a limitation.

Moreover, the validation corpus, which has been used for testing, has named entities out of the scope of common knowledge and that are not present either in the training corpus. For instance, the named entity 'tcae' is annotated and means *Técnico en Cuidados Auxiliares de Enfermería* (Nursing Auxiliary Care Technician). All the models fail in the recognition of these kinds of named entity and the results are never improved more than the 78 % of F-Measure.

Also, it is important to highlight that the mention replacement method has not achieved the expected results presented in other works of the state-of-the-art. Analyzing the context of the use case, we have consider than mention replacement needs from more representative natural language sentences with named entities to be switched. The original corpus is comprised of short tweets that contain profession named entities, sometimes, without any context. On the contrary, the back translation method has achieved good

[6] https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne

	10 %			30 %			50 %			100 %		
	p	r	F	p	r	F	p	r	F	p	r	F
No Aug	0.792	0.471	0.591	0.761	0.685	0.721	0.782	0.695	0.736	0.829	0.746	**0.785**
MR	0.640	0.685	0.661	0.732	0.692	0.711	0.752	0.728	0.740	0.752	0.728	0.740
BT	0.738	0.654	0.693	0.739	0.709	0.724	0.775	0.690	0.730	0.788	0.756	0.771
Widaug	0.756	0.667	**0.709**	0.833	0.647	**0.728**	0.810	0.695	**0.748**	0.819	0.718	0.765
Widaug+BT	0.733	0.670	0.700	0.787	0.698	**0.740**	0.787	0.698	0.740	0.797	0.748	0.772

Tabla 2: Evaluation results measuring Precision (p), Recall (r) and F-measure (F) for the four corpus (10, 30, 50 and 100 %). Each row corresponds to a data augmentation method: no augmentation (No Aug), mention replacement (MR), back translation (BR), the proposed method (Widaug) and the combination of Widaug and back translation (Widaug+BT).

results for two main reasons. The first one is that some named entities have changed their gender in the process of translating to English language. Words that in Spanish were feminine, come back as masculine, doing a good data augmentation process. The second one is that some named entities of the corpus have changed their original language; there are cases of Catalan mentions that are back translated to Spanish or cases in which the English terms are also accepted in Spanish language such as *animadora* (cheerleader). Even though back translation have achieve close results to Widaug, we cannot generalize that could be similar for other scenarios without the properties of this particular corpus.

5 Conclusions and future lines

This paper has shown a simple data augmentation approach based on the use of Wikidata as a source of information. Knowledge graphs represent and link concepts and information already validated by humans, and this type of resource has not been exploited at all in the generation of new synthetic data for data augmentation.

The results show that there is a significant improvement for small datasets. For instance, the improvement of the 10 % corpus has been up to 0.11 more. Therefore, this method will cover scenarios in which it is difficult to find annotated data without involving human effort. Moreover, one of the benefits of using Wikidata and Wikipedia as an external resource to generate new data is that the new sentences are still human-readable and do not contain language errors, as approaches such as synonym replacement or random deletion may produce.

However, the method does not reflect a significant improvement over the full dataset. In these particular experiments we have discovered, as the discussion section presents, that the difficulties presented in the validation corpus make difficult to achieve better results over the training with the original corpus.

Analyzing the results and the discussion lines, the method has different future lines which should be explored. Some of them are:

- The use of a generic knowledge graph such as Wikidata does not allow the application of this method to more specific scenarios. For instance, for the identification of diseases within a biomedical domain. Wikidata does not contain specific domain information such as MESH or SNOMED-CT. Therefore, the use of knowledge graphs of other specific domains could be explored. This could allow for a higher level of detail in the generation of new synthetic data, which could lead to better quality results.
- Wikidata contains a huge amount of heterogeneous information. This is why many of the entities extracted, although correct, are far from the target domain and should be filtered. An example of this can be seen in the context of the 'profession' itself, where some of the retrieved entities were far from the context of the use case (for instance: 'chess referee', 'esperantologist', 'primatologist', etc). The word embedding filter is a key process to get better results and not add noise to the training corpus. However, pretrained word embedding models do not represent n-gram terms. Thus, pro-

fessions with n-gram terms are not being filtered correctly (e.g., *guardia civil* (civil guard)). So, new approaches will focus on the identification of the correct vectors for those terms, using, for instance, sentence embeddings which are based on language models.

- Moreover, Wikidata has a gender bias. Most of the concepts are presented with male gender and the new synthetic data are created in the same way. The next step is to identify the impact of generating instances in the female gender.
- Synthetic data generation with language adaptation of the original sentences. In this particular use case, we have found that the syntactic structures of natural language sentences generated or extracted are far different from the original ones that are tweets. Thus, a better adaptation to the original style, in terms of terminology and syntactic structure, should be done.

Acknowledgements

Financed by the European Union-NextGenerationEU (UP2021-035), by the SmarTerp Project (EIT-Digital-21184) and by the project HCommonK (RTC2019-007134-7, funded by MCIN/AEI/ 10.13039/501100011033).

References

Asghari, M., D. Sierra-Sosa, and A. S. Elmaghraby. 2022. Biner: A low-cost biomedical named entity recognition. *Information Sciences*, 602:184–200.

Bayer, M., M.-A. Kaufhold, B. Buchhold, M. Keller, J. Dallmeyer, and C. Reuter. 2022. Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers. *International Journal of Machine Learning and Cybernetics*, pages 1–16.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2016. Enriching word vectors with subword information.

Dai, X. and H. Adel. 2020. An analysis of simple data augmentation for named entity recognition. *arXiv preprint arXiv:2010.11683*.

Ding, B., L. Liu, L. Bing, C. Kruengkrai, T. H. Nguyen, S. Joty, L. Si, and C. Miao. 2020. Daga: Data augmentation with a generation approach for low-resource tagging tasks. *arXiv preprint arXiv:2011.01549*.

Erd, R., L. Feddoul, C. Lachenmaier, and M. J. Mauch. 2022. Evaluation of data augmentation for named entity recognition in the german legal domain. In *AI4LEGAL-KGSUM 2022 Artificial Intelligence Technologies for Legal Documents and Knowledge Graph Summarization 2022*, number 3257 in CEUR Workshop Proceedings.

Farré-Maduell, E., G. González Gacio, S. Lima, A. Miranda-Escalada, and M. Krallinger. 2022. LivingNER Guidelines: Named entity recognition, normalization & classification of species, pathogens and food, April. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

Grishman, R. and B. M. Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, and M. Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68(0):39–60.

Guzman-Silverio, M., Á. Balderas-Paredes, and A. P. López-Monroy. 2020. Transformers and data augmentation for aggressiveness detection in mexican spanish. In *IberLEF@ SEPLN*, pages 293–302.

Kang, T., A. Perotte, Y. Tang, C. Ta, and C. Weng. 2021. Umls-based data augmentation for natural language processing of clinical research literature. *Journal of the American Medical Informatics Association*, 28(4):812–823.

Kim, J., Y. Kim, and S. Kang. 2022. Weakly labeled data augmentation for social media named entity recognition. *Expert Systems with Applications*, 209:118217.

Li, X., H. Zhang, and X.-H. Zhou. 2020. Chinese clinical named entity recognition

with variant neural structures based on bert methods. *Journal of biomedical informatics*, 107:103422.

Liu, Q., P. Li, W. Lu, and Q. Cheng. 2020. Long-tail dataset entity recognition based on data augmentation. In *EEKE@ JCDL*, pages 79–80.

Luo, H. 2021. Emotion detection for spanish with data augmentation and transformer-based models. In *IberLEF@ SEPLN*, pages 35–42.

Luque, F. M. 2019. Atalaya at tass 2019: Data augmentation and robust embeddings for sentiment analysis. *arXiv preprint arXiv:1909.11241*.

Malmasi, S., A. Fang, B. Fetahu, S. Kar, and O. Rokhlenko. 2022. Semeval-2022 task 11: Multilingual complex named entity recognition (multiconer). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1412–1437.

Marivate, V. and T. Sefara. 2020. Improving short text classification through global augmentation methods. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 385–399. Springer.

Miranda-Escalada, A., E. Farré, and M. Krallinger. 2020. Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. *IberLEF@ SEPLN*, pages 303–323.

Miranda-Escalada, A., E. Farré-Maduell, S. Lima-López, L. Gascó, V. Briva-Iglesias, M. Agüero-Torales, and M. Krallinger. 2021. The profner shared task on automatic recognition of occupation mentions in social media: systems, evaluation, guidelines, embeddings and corpora. In *Proceedings of the Sixth Social Media Mining for Health (SMM4H) Workshop and Shared Task*, pages 13–20.

Perera, N., M. Dehmer, and F. Emmert-Streib. 2020. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in cell and developmental biology*, page 673.

Raiman, J. and J. Miller. 2017. Globally normalized reader. *arXiv preprint arXiv:1709.02828*.

Rekabsaz, N., M. Lupu, and A. Hanbury. 2017. Exploration of a threshold for similarity based on uncertainty in word embedding. In *Advances in Information Retrieval: 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings 39*, pages 396–409. Springer.

Rodriguez, S., R. Gretter, M. Matassoni, A. Alonso, O. Corcho, M. Rico, and F. Daniele. 2021. SmarTerp: A CAI system to support simultaneous interpreters in real-time. In *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 102–109, Held Online, July. INCOMA Ltd.

Schindler, D., F. Bensmann, S. Dietze, and F. Krüger. 2021. Somesci-a 5 star open data gold standard knowledge graph of software mentions in scientific articles. *arXiv preprint arXiv:2108.09070*.

Serrano, A. V., G. G. Subies, H. M. Zamorano, N. A. Garcia, D. Samy, D. B. Sanchez, A. M. Sandoval, M. G. Nieto, and A. B. Jimenez. 2022. Rigoberta: A state-of-the-art language model for spanish. *arXiv preprint arXiv:2205.10233*.

Tjong Kim Sang, E. F. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Wang, W. Y. and D. Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using % petpeeve tweets. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2557–2563.

Wei, J. and K. Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Wu, X., S. Lv, L. Zang, J. Han, and S. Hu. 2019. Conditional bert contextual augmentation. In *International conference*

on computational science, pages 84–95. Springer.

Xie, Q., Z. Dai, E. Hovy, T. Luong, and Q. Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268.

Yaseen, U. and S. Langer. 2021. Data augmentation for low-resource named entity recognition using backtranslation. *arXiv preprint arXiv:2108.11703*.

Zhang, X., J. Zhao, and Y. LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 157-170 recibido 16-12-2022 revisado 31-01-2023 aceptado 10-02-2023

Lessons learned from the evaluation of Spanish Language Models

Conclusiones de la evaluación de Modelos del Lenguaje en Español

Rodrigo Agerri, Eneko Agirre
HiTZ Center - Ixa, University of the Basque Country UPV/EHU
rodrigo.agerri@ehu.eus, e.agirre@ehu.eus

Abstract: Given the impact of language models on the field of Natural Language Processing, a number of Spanish encoder-only masked language models (aka BERTs) have been trained and released. These models were developed either within large projects using very large private corpora or by means of smaller scale academic efforts leveraging freely available data. In this paper we present a comprehensive head-to-head comparison of language models for Spanish with the following results: (i) Previously ignored multilingual models from large companies fare better than monolingual models, substantially changing the evaluation landscape of language models in Spanish; (ii) Results across the monolingual models are not conclusive, with supposedly smaller and inferior models performing competitively. Based on these empirical results, we argue for the need of more research to understand the factors underlying them. In this sense, the effect of corpus size, quality and pre-training techniques need to be further investigated to be able to obtain Spanish monolingual models significantly better than the multilingual ones released by large private companies, specially in the face of rapid ongoing progress in the field. The recent activity in the development of language technology for Spanish is to be welcomed, but our results show that building language models remains an open, resource-heavy problem which requires to marry resources (monetary and/or computational) with the best research expertise and practice.
Keywords: Masked Language Models, Text Classification, Sequence Labelling, Natural Language Processing.

Resumen: Actualmente existen varios modelos del lenguaje en español (también conocidos como BERTs) los cuales han sido desarrollados tanto en el marco de grandes proyectos que utilizan corpus privados de gran tamaño, como mediante esfuerzos académicos de menor escala aprovechando datos de libre acceso. En este artículo presentamos una comparación exhaustiva de modelos de lenguaje en español con los siguientes resultados: (i) La inclusión de modelos multilingües previamente ignorados altera sustancialmente el panorama de la evaluación para el español, ya que resultan ser en general mejores que sus homólogos monolingües; (ii) Las diferencias en los resultados entre los modelos monolingües no son concluyentes, ya que aquellos supuestamente más pequeños e inferiores obtienen resultados más que competitivos. El resultado de nuestra evaluación demuestra que es necesario seguir investigando para comprender los factores que subyacen a estos resultados. En este sentido, es necesario seguir investigando el efecto del tamaño del corpus, su calidad y las técnicas de preentrenamiento para poder obtener modelos monolingües en español significativamente mejores que los multilingües ya existentes. Aunque esta actividad reciente demuestra un creciente interés en el desarrollo de la tecnología lingüística para el español, nuestros resultados ponen de manifiesto que el desarrollo de modelos de lenguaje sigue siendo un problema abierto que requiere conjugar recursos (monetarios y/o computacionales) con los mejores conocimientos y prácticas de investigación en PLN.
Palabras clave: Modelos de Lenguaje, Clasificación de Textos, Etiquetado Secuencial, Procesamiento del Lenguaje Natural.

ISSN 1135-5948 DOI 10.26342/2023-70-13

1 Introduction

Deep Learning has changed the application and research landscape in Natural Language Processing (NLP). The field has experienced a paradigm shift that has rendered previous techniques obsolete for many tasks, and nowadays large companies such as Google or Meta rely on deep learning techniques to develop NLP applications. Central to these developments lay large pre-trained language models, which are trained on gigantic corpora (e.g. crawls of the entire Web) requiring costly hardware. The cost of developing and training such models is so high that most recent innovations come from such large companies and focus on English. Thus, the best available language models for English have been released to the public by large companies. Furthermore, in some cases large language models that are currently being used are not even released, but offered instead as a pay-per-use API.

A natural question arises regarding languages other than English, as the same large companies have published multilingual versions of these models with support for 100 languages, such as multilingual BERT and XLM-RoBERTa (Devlin et al., 2019; Conneau et al., 2020). While these multilingual models excel in many NLP tasks involving high-resourced languages such as English, their performance is not always as good as monolingual models. In fact, recent studies seem to suggest that a careful training design and appropriate corpora selection results in better models for each specific language (Martin et al., 2020; Agerri et al., 2020; Agerri, 2020). Although several language model architectures exist, most efforts building monolingual models have focused on encoder-only masked language models (e.g. BERT and variants) (Devlin et al., 2019; Liu et al., 2019), so we will leave decoder-only causal language models (e.g. GPT) and encoder-decoder models (e.g. T5) for future analysis (Brown et al., 2020; Zhang et al., 2022; Scao et al., 2022; Raffel et al., 2020; Xue et al., 2021).

Thus, following previous work comparing monolingual and multilingual models (de Vries et al., 2019; Virtanen et al., 2019; Martin et al., 2020; Agerri, 2020; Tanvir, Kittask, and Sirts, 2021; Armengol-Estapé et al., 2021), in this paper we are going to focus on Spanish, for which several encoder-only masked language models have been trained and released (Cañete et al., 2020; Gutiérrez-Fandiño et al., 2022; De la Rosa et al., 2022). The models have been developed either in heavily-subsidized projects with very large corpora or in smaller scale academic efforts on more limited, freely available corpora. In order to compare the quality of the language models, we follow usual practice and perform a downstream evaluation where all language models are treated equally and applied to a large set of Spanish NLP evaluation datasets, including common tasks such as part-of-speech tagging, named-entity recognition, natural language inference, semantic textual similarity, question answering, paraphrasis or metaphor detection. However, unlike previous evaluations for Spanish, we do include in our evaluation widely used multilingual models such as XLM-RoBERTa and mDeBERTa (Conneau et al., 2020; He, Gao, and Chen, 2021).

Our comprehensive head-to-head comparison yields surprising results: (i) Considering the previously ignored XLM-RoBERTa and mDeBERTa substantially change the evaluation landscape of language models in Spanish, as they happen to fare better than their monolingual counterparts. In particular, our results show that XLM-RoBERTa-large, released by Meta in 2020 (Conneau et al., 2020) obtains the best results in the majority of the tasks. Furthermore, mDeBERTa (He, Gao, and Chen, 2021), a smaller base-size model, performs second overall. (ii) Despite claims to the contrary (Gutiérrez-Fandiño et al., 2022), results among the monolingual models are quite close, and supposedly smaller and inferior models such as IXABERTesv2[1] obtaining similar or better results with respect to the the MarIA RoBERTa-bne models; (iii) In addition to downstream evaluation, the effect of corpus size, corpus quality and pretraining techniques need to be further investigated (Martin et al., 2020; Artetxe et al., 2022) to advance current state-of-the-art in language models; (iv) despite the strong results obtained by evaluating the language models, for some tasks they remain well below the state-of-the-art. Code and data is publicly available to facilitate research on this topic and reproducibility of results[2].

[1] http://www.deeptext.eus/eu/node/3

[2] https://github.com/ragerri/evaluation-spanish-language-models

Based on this findings, we argue for more research to understand the factors underlying the results and to be able to obtain Spanish monolingual models significantly better than the multilingual ones released by large private companies. While this recent activity building models bodes well the development of language technology for Spanish, our results show that building language models remains an open, resource-heavy problem which requires to marry resources (monetary and/or computational) with the best research expertise and practice.

The rest of the paper is structured as follows. Next section discusses related work on monolingual and multilingual language models. Section 3 provides details of the language models for Spanish that will benchmarked in Section 5 following the experimental setup of Section 4. In Section 6 we will go over the lessons learned quite throughly and we will finish with some concluding remarks.

2 Related Work

The release of encoder-based masked language models (MLMs) for English caused a paradigm-shift in Natural Language Processing (NLP) research. After the original BERT model (Devlin et al., 2019), many variations and improvements were quickly developed (Liu et al., 2019; He, Gao, and Chen, 2021). At the same time, large multilingual models such as multilingual BERT and XLM-RoBERTa, trained to work on 100 languages, were published, with extraordinary results both monolingual and, especially, on multilingual and cross-lingual settings (Pires, Schlinger, and Garrette, 2019; Wu and Dredze, 2020; Conneau et al., 2020). The availability of such multiligual models posed the question whether they were the optimal solution for other languages different to English. This in turn caused the appearance of a large body of research studying the performance of such multilingual models on specific languages, often in comparison to monolingual counterparts specifically tailored to the target language (Nozza, Bianchi, and Hovy, 2020).

Recent studies suggest that while the multilingual models excel in many NLP tasks involving high-resourced languages such as English, their performance is not usually as good as monolingual models. Thus, previous work on monolingual models for languages such as Basque or French suggest that a careful training design and appropriate corpora selection results in better models for each specific language (Martin et al., 2020; Agerri et al., 2020).

Other studies focused on the quality of the corpus itself (Virtanen et al., 2019; Tanvir, Kittask, and Sirts, 2021) while for other languages such as Basque or Catalan, in addition to developing language models, a large effort on generating new datasets for benchmarking was also put in place (Armengol-Estapé et al., 2021; Urbizu et al., 2022). Finally, recent research has empirically demonstrated that, while size is important, carefully studying the pre-training method and auditing the quality of the corpus is crucial to understand the performance of language models on downstream tasks (Kreutzer et al., 2022; Artetxe et al., 2022).

In any case, most of the previous work shows that monolingual models perform in general better than the multilingual ones, also with respect to XLM-RoBERTa (Martin et al., 2020; Armengol-Estapé et al., 2021). However, for Spanish the situation is slightly different because the largest evaluation of language models for Spanish does not include XLM-RoBERTa or the more recent mDeBERTa (Gutiérrez-Fandiño et al., 2022). In this work we will address this issue by including them in the evaluation of language models for Spanish.

3 Spanish Language models

Spanish has been quite a newcomer in the Transformer-based language model fever, which was hard to understand given that Spanish is the fourth most spoken language in the world. Thus, while the number of language-specific models proliferated at a vertiginous rhythm for many world languages, BETO (Cañete et al., 2020) remained the only language model for a surprisingly large period of time. BETO follows a BERT-base architecture and was released around the end of 2019 by researchers at the University of Chile[3]. The model was trained on a collection of corpora which included the Spanish Wikipedia and the OPUS Spanish corpus (Tiedemann and Thottingal, 2020) and it was evaluated on the GLUES (short for GLUE in Spanish) dataset[4], compar-

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

[4] https://github.com/dccuchile/glues

Model	corpus	#words	L	H	A	V	#params
Multilingual BERT	Wiki	0.7B	12	768	12	110K	110M
BETO	Opus,Wiki	3B	12	768	12	30K	110M
IXABERTesv1	Gigaword,Wiki	5.7B	12	768	12	50K	110M
ixambert	Wiki	0.7B	12	768	12	119K	110M
IXABERTesv2	OSCAR	25B	12	768	12	50K	125M
XLM-RoBERTa-base	CC-100	9.3B	12	768	12	250K	270M
XLM-RoBERTa-large	CC-100	9.3B	24	1024	16	250K	550M
Electricidad	Opus,Wiki	3B	12	768	12	31K	110M
BERTIN	mC4-es	47B	12	768	12	50K	125M
RoBERTa-base-bne	BNE	135B	12	768	12	50K	125M
RoBERTa-large-bne	BNE	135B	24	1024	16	50K	350M
mDeBERTa	CC-100	9.3B	12	768	12	250K	198M

Table 1: Spanish Language Models (in approximate order of creation). L: layer size; H: hidden size; A: attention heads; V: vocabulary.

ing favourably with respect to multilingual BERT.

However, once started, language models for Spanish quickly proliferated. In 2020 two models, based on BERT and RoBERTa-base (IXABERTesv1 and v2), were released[5] by the Ixa Group of the University of the Basque Country. This group also published that year a multilingual model for Basque, Spanish and English, ixambert, following the BERT-base architecture (Otegi et al., 2020).

One year later, a community-based effort coordinated within the Flax/Jack Community Week organized by HuggingFace released BERTIN[6] a RoBERTa-base model (De la Rosa et al., 2022). This model was trained on the Spanish portion of the mC4 dataset (Xue et al., 2021). Some of the BERTIN developers also released an Electra-base Spanish model: Electricidad[7].

Concurrently, a team from the Barcelona Supercomputing Center funded by the Spanish Government released under the MarIA project[8] two models, RoBERTa-base-bne and RoBERTa-large-bne, trained on a large corpus based on crawling data from the Spanish National Library (BNE corpus). The MarIA models were compared with respect to BETO, BERTIN, Electricidad and multilingual BERT (Gutiérrez-Fandiño et al., 2022). Results from other commonly-used multilingual models such as XLM-RoBERTa (both base and large) or mDeBERTa were not included in the evaluation.

All language models have been trained on publicly available corpora, except the BNE corpus[9]. Public availability is important, as many features and biases of the language models depend on the corpora where they have been trained. Furthermore, public availability is required to guarantee reproducibility of results. It also allows researchers, companies and users to examine those corpora and thus assess the impact that the features of the corpora will have in their research and products.

3.1 Models details

Table 1 shows the most important details of the language models we will use in our study, including the corpus type and size on which they were trained, and technical pre-training details such as the number of layers, the hidden size, number of attention heads, the vocabulary and the number of parameters. In the rest of this section we will comment other relevant aspects to interpret the results reported in Section 5.

BETO, IXABERTesv1 and ixambert are BERT-base models pre-trained with both Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) (Devlin et al., 2019). BETO performed 2M steps in two different stages: 900K steps with a batch

[5] http://www.deeptext.eus/eu/node/3

[6] https://huggingface.co/bertin-project/bertin-roberta-base-spanish

[7] https://huggingface.co/mrm8488/electricidad-base-discriminator

[8] https://github.com/PlanTL-GOB-ES/lm-spanish

[9] In the paper the MarIA authors mention that it will be released soon, although at the time of writing the corpus is not available.

size of 2048 and maximum sequence length of 128, and the rest of the training with batch size of 256 and maximum sequence length of 512. Both IXABERTesv1 and ixambert were trained by executing 1M steps with 256 of batch size and 512 sequence length.

The language models using RoBERTa-base (IXABERTesv2, BERTIN and RoBERTa-base-bne) and large (RoBERTa-large-bne) are based on the BERT architecture but (i) trained only on the MLM task, (ii) on larger batches (iii) on longer sequences and (iv), with dynamic mask generation. While IXABERTesv2 performed 120.500 steps with 2048 batch size and sequence length 512, BERTIN was trained on 250K steps divided in two steps: 230k steps with sequences of length 128 and batch size 2048, and the rest of the training with 512 sequence length and 384 of batch size. Thus, both IXABERTesv2 and BERTIN roughly follow the RoBERTa approach to pre-training (Liu et al., 2019). However, the MarIA models opted instead for a batch of 2048 and 512 sequence length, but reducing the training to one epoch only with no dropout (Komatsuzaki, 2019).

With respect to the multilingual models, multilingual BERT was trained with a batch size of 256 and 512 sequence length for 1M steps, using both the MLM and NSP tasks. Regarding XLM-RoBERTa, both versions were trained over 1.5M steps with batch 8192 and sequences of 512 length. Finally, mDeBERTa (He, Gao, and Chen, 2021) is based on RoBERTa but incorporating disentangled attention, gradient-disentagled embedding sharing and, most importantly, replacing the MLM task with replaced token detection (RTD), originally proposed by ELECTRA (Clark et al., 2020); mDeBERTa was trained following the XLM-RoBERTa procedure but reducing the steps from 1.5M to 500K.

Thus, the specific pre-training details and the corpora used to generate the language models substantially differ across the monolingual and the multilingual models. However, as we will see in the next section, the fine-tuning performed to evaluate the models on downstream tasks will follow the same methodology.

4 Experimental setup

Our experimental setup follows the one proposed by MarIA (Gutiérrez-Fandiño et al., 2022), with the caveat that we include 6 more language models in our evaluation and two extra datasets. Thus, the 12 models listed in Table 1 are evaluated on 8 tasks and 11 datasets: For POS tagging the UD and Capitel datasets (Taulé, Martí, and Recasens, 2008; Porta and Espinosa-Anke, 2020); for NER we use CoNLL 2002 (Tjong-Kim-Sang, 2002), Capitel (Porta and Espinosa-Anke, 2020) and Ancora 2.0 (Taulé, Martí, and Recasens, 2008); the Semantic Text Similarity dataset is based on the data by Agirre et al. (2014) and Agirre et al. (2015); MLDoc (Schwenk and Li, 2018) for document classification; paraphrase identification with PAWS-X (Yang et al., 2019), XNLI for Natural Language Inference (Conneau et al., 2018), Question Answering with the SQAC data (Gutiérrez-Fandiño et al., 2022) and CoMeta (Sanchez-Bayona and Agerri, 2022) for metaphor detection.

For comparison purposes, we use the same data splits as in the MarIA paper. For the two datasets added for this paper, Ancora 2.0 NER and CoMeta, we make public the splits we created. Both Ancora 2.0 and CoMeta are publicly available and we thought that they were a good addition to the benchmark. In this sense, it should be noted that every dataset is public except the Capitel POS and NER corpora. We are not particularly fond of using data which is not publicly available, at least for research, because it makes reproducibility impossible thereby hindering the progress of scientific research. However, we decided to include them to make it a more comprehensive comparison with previous work on benchmarking language models in Spanish.

For fine-tuning the models we use the same scripts used by Gutiérrez-Fandiño et al. (2022) as available in their Github repository[10] with minor modifications. For each task, a single linear layer is added on top of the model being fine-tuned. In the case of sentence/paragraph-level classification tasks, the `[CLS]` token is used for BERT models, and the `<s>` token in the case of RoBERTa models. We use maximum sequence length of

[10] `https://github.com/PlanTL-GOB-ES/lm-spanish`

	Spanish Base					*Multilingual Base*				*Large*	
Dataset	Beto	Bertin	Elect.	MarIA	IXAes	IXAm	mBERT	XLM-R	mDeB3	MarIA **L**	XLM-R**L**
PoS UD	99.00	98.98	98.18	99.07	99.03	98.90	99.01	99.02	99.05	99.04	**99.11**
PoS Capitel	98.36	98.47	98.16	98.46	98.55	98.32	98.39	98.47	98.56	98.56	**98.63**
NERC CoNLL	87.59	88.35	79.54	88.51	88.70	87.85	86.91	88.11	88.73	88.23	**89.02**
NERC Ancora	92.46	92.15	85.66	93.34	**93.57**	92.58	92.58	92.47	93.02	92.45	93.13
NERC Capitel	87.72	88.56	80.35	89.60	89.83	88.65	88.10	88.55	89.86	**90.51**	90.19
STS	81.59	79.45	80.63	**85.33**	83.82	83.09	81.64	83.47	83.61	84.11	84.04
MLDoc	**97.14**	96.68	95.65	96.64	96.78	96.70	96.17	96.30	96.62	97.02	97.05
PAWS-X	89.30	89.65	90.45	90.20	89.99	88.06	90.00	89.82	91.90	91.50	**91.93**
XNLI	81.30	78.90	78.78	80.16	82.40	79.40	78.76	81.14	84.85	82.63	**84.95**
SQAC	79.23	76.78	73.83	79.23	78.91	77.38	75.62	77.28	80.78	82.02	**84.10**
CoMeta	64.28	61.52	61.18	63.08	64.79	62.04	61.77	63.82	**67.46**	62.02	67.44
Average	87.09	86.32	83.86	87.60	87.85	86.63	86.27	87.13	88.59	88.01	**89.05**
Average*	89.37	88.80	86.12	90.05	90.16	89.09	88.72	89.46	90.70	90.61	**91.22**
Wins group	1.5		1	2.5	6	1			10	2	9
Wins all	1			1	1				1	1	**6**

Table 2: Results with models grouped according to: Spanish base-size, multilingual base-size, and large-size (one Spanish and one multilingual). Best result per group with underline, best result overall in **bold**. We report average across datasets, average* without the metaphor dataset CoMeta, wins in each group and wins overall (ties are scored as $1/n$ where n is systems tied). Metric F1 micro except for MLDoc and XNLI (accuracy); STS is evaluated on the official *combined score*. For space reasons we only report results from one Ixa monolingual model: IXAes = IXABERTesv2.

512. A grid search of hyperparameters is performed to pick the best batch size (8, 16, 32), weight decay (0.01, 0.1) and learning rate (1e-5, 2e-5, 3e-5, 5e-5). We pick the best model on the development set over 5 epochs. We keep a fixed seed to ensure reproducibility of results. The experiments have been implemented using the HuggingFace Transformers API (Wolf et al., 2020). Code and data splits are publicly available[11].

5 Results

Table 2 shows the results for each model in each dataset. Results already reported by Gutiérrez-Fandiño et al. (2022) are included here verbatim. The rest of the results have been obtained by fine-tuning the models following the method described in the previous section. The average across datasets and the number of datasets where one method wins over the rest allow to set a clear picture.

First, among Spanish-only base models, the best results are obtained by IXAes, which performs better than MarIA (the second best) in both average and wins in datasets. They are followed by BETO, BERTIN and finally Electricity. This result is interesting as IXAes is trained with a much smaller public corpus.

Second, if we look at the multilingual base models, mDeBERTa is the clear winner, followed by XLM-RoBERTa and ixambert which perform quite similarly.

Third, if we compare monolingual and multilingual base models, the monolingual IXAes outperforms the best comparable multilingual model, XLM-RoBERTa. However, the newer mDeBERTa yields the best results overall. It should be noted that all the Spanish models were produced before the DeBERTa v3 architecture was introduced, which may perhaps explain their lower results.

Fourth, regarding the largest models, XLM-RoBERTa outperforms MarIA large in 9 out of 11 datasets, and obtains a better average performance. In fact, even mDeBERTa obtains slightly better results than MarIA large. Moreover, the pre-existing XLM-RoBERTa model works for 99 additional languages, allowing also to perform cross-lingual transfer. The only single disadvantage is that the size of XLM-RoBERTa is larger, mostly due to its larger vocabulary size, but the cost in running time (Flops) is comparable for both.

Overall, results demonstrate that XLM-RoBERTa-large is the best model across the board, including the newer mDeBERTa. The DeBERTa team have not reported results

[11] `https://github.com/ragerri/evaluation-spanish-language-models`

or released a large DeBERTa multilingual model, but given the strong results of the English DeBERTa large model (He, Gao, and Chen, 2021), it can be assumed that its results may be superior to those obtained by XLM-RoBERTa-large.

Finally, it should be noted that for the task of metaphor detection the results are significantly lower across the board. This is not entirely surprising, as the state-of-the-art in metaphor detection is in general quite low. In any case, and motivated by this fact, we also calculated the average* without taking into account the metaphor detection results. As it can be seen, while the results get slightly higher, the trends discussed still hold.

6 Discussion

According to the results, the following lessons can be drawn.

Which model should I use according to my computing budget? If the user is interested in best results at inference, XLM-RoBERTa-large is nowadays the best option, at the cost of requiring more time and GPU memory. mDeBERTa would be the next best choice for smaller memory and runtime budgets. For a more modest solution, IXAes would be a good choice.

Which model should I use according to my task? In this work we cover a broad but limited number of datasets. If your target task is similar to one of the datasets, then you might want to use the model that excels at this task and that meets your budget requirements (in terms of the GPU hardware that it can be afforded). For most tasks XML-RoBERTa-large is the best option, with the additional benefit from cross-lingual transfer. For smaller budgets we recommend to check the underlined results in the different groups in Table 2. For the cases where your target task is not covered, the safest option is to take the best overall model according to your budget.

Is there an explanation for the lower performance of some models? Larger models are expected to perform better. Furthermore, the mDeBERTa results are not particularly surprising. However, in the case of models with the same architecture and size, it would be good to be able to pinpoint the causes for the disappointing performance of some models.

An important factor could be the **corpora** used. In principle the MarIA models use the largest and, according to their authors, the cleanest corpus for Spanish ever produced. However, it turns out that, for the same base size, IXAes gets better results, even if it was trained on a smaller corpus (OSCAR) which is publicly available since 2019 (Ortiz Suárez, Sagot, and Romary, 2019). OSCAR is based on Common Crawl, covers 166 languages, and uses a very light publicly available filtering software, while the BNE corpus was filtered in-house following previous work (Virtanen et al., 2019). The strongest performers (XLM-RoBERTa and mDeBERTa) also use a filtered version of Common Crawl, CC100, which in this case was publicly released by Facebook around 2020 (Conneau et al., 2020). There are evidences that high-quality filtering does not improve downstream performance and that size seems to be equally important (Artetxe et al., 2022). Perhaps an audit of a sample of the BNE corpus compared with the other corpora used to train the models would provide further light on this issue. On this line of research, two possible strategies would be to: (i) use the same architecture and training procedure but with different corpora (Artetxe et al., 2022); (ii) fix the corpus used for training varying the training method and specifications.

Other explanations may be related to how much training procedure and hyperparameters vary from one model to the other (see Section 3). Although an exhaustive analysis is not feasible, two key factors could be the *size of the vocabulary* (Zheng et al., 2021) and the number of *training examples seen in training*. In fact, the Spanish models have relatively small vocabularies compared to their XLM-RoBERTa and DeBERTa counterparts, and BETO and Electricidad have smaller vocabulary size than the better performing IXAes and MarIA. Thus, vocabulary size might be part of the explanation, but it does not explain the differences in results between the Spanish models with the same vocabulary, so we may need to consider other possible explanations.

If we look at the number of steps in training, MarIA uses a strategy which is substantially different to the rest of the models, in particular to XLM-RoBERTa and mDeBERTa. Both longer (Devlin et al.,

2019; Conneau et al., 2020) and shorter (Komatsuzaki, 2019) training have been recommended. In the light of the results, one would say that the strategy from XLM-RoBERTa and mDeBERTa is the best, so in this case it would look like as if some of the Spanish models have been undertrained. However, in order to have a more conclusive answer, it would be necessary to experiment with the number of steps fixing the other variables involved in the training process.

Summarizing, it seems that publicly available corpora suffice for optimal results, and that the larger the model and the vocabulary the better. Additionally, the number of steps could also play an important role. Unfortunately, the post-hoc analysis carried out in this paper cannot give a more precise picture, and carefully designed experiments along the lines of the ones suggested above would be necessary to shed some more light and perhaps to improve results.

Training a monolingual model, is it worth it? Common wisdom indicates that monolingual models improve over multilingual models (Martin et al., 2020; Agerri et al., 2020; Virtanen et al., 2019; Tanvir, Kittask, and Sirts, 2021; Armengol-Estapé et al., 2021), which led to a proliferation of models for many target languages. Most of the models have been shown to outperform their multilingual counterparts, but often have only considered multilingual BERT completely ignoring XLM-RoBERTa (Nozza, Bianchi, and Hovy, 2020).

Part of the mixed signals could be also caused by the size of the language: while large languages like Spanish and English are very well represented in multilingual models, low-resource languages tend to have a very small quota of training instances. Training a model using larger amounts of better quality corpora for low-resource languages could thus explain the good results of monolingual models with respect to multilingual ones (Agerri et al., 2020; Bhattacharjee et al., 2021; Nzeyimana and Rubungo, 2022), but this may not necessarily be the case for high-resource languages, as evidenced by the results reported in Table 2.

Our work shows that some monolingual base models such as IXAes or MarIA do slightly improve over the results of a comparable XLM-RoBERTa-base multilingual model. However, the two best performing models for Spanish are currently mDeBERTa (base) and XLM-RoBERTa-large. Considering these results and the literature mentioned above, it would seem that the amount and quality of publicly available Spanish corpora suffices, and that future improvements will need to come from larger models or architecture improvements, as shown by DeBERTa or T5 for English, or by careful experimentation as outlined above.

Better research reporting practices should be encouraged. The XLM-RoBERTa models were widely known and available when the Spanish models were built, but none of the publications on language models in Spanish compared their results to XLM-RoBERTa, implicitly sending the wrong message that ignoring XLM-RoBERTa was the best option when working with Spanish language models. As our results show, XLM-RoBERTa is currently the strongest option to build NLP applications in Spanish.

Comparison to the state-of-the-art. In relation to the previous point, research on language models seem to be inadvertently forgetting the primary objective of building language models in the first place, namely, improving the state-of-the-art of NLP technology. Thus, previous published work do not mention what the state-of-the-art is for each of the tasks used to benchmark the models. Without doing so, it is just not possible to know how much a given language model is actually advancing NLP technology. Therefore, we first reevaluate three tasks (PAWS-X and Capitel and UD POS) to report the most common accuracy metric usually used for those tasks (instead of the F1 score used in previous evaluations of language models in Spanish). Table 3 offers the overall results with PAWS-X, Capitel and UD PoS evaluated using accuracy. The new results were obtained by fine-tuning all 12 models following the methodology provided in Section 4. As it can be seen, they confirm the trends already observed and discussed above.

Based on Table 3 we can now compare the results of the models with respect to the state-of-the-art in each task. First, it should be noted that for five tasks (Capitel PoS, Ancora 2.0 NER, STS, SQAC and CoMeta) their results have been published for the first time during the evaluation of

	Spanish Base					*Multilingual Base*				*Large*		Prev
Dataset	Beto	Bertin	Elect.	MarIA	IXAes	IXAm	mBERT	XLM-R	mDeB3	MarIA **L**	XLM-R**L**	SOTA
PoS UD	99.10	99.11	98.37	99.14	99.17	98.98	99.01	99.16	**99.20**	99.12	99.19	99.05
PoS Capitel	98.57	98.63	98.40	98.67	98.75	98.55	98.60	98.68	98.76	98.73	**98.82**	-
NERC CoNLL	87.59	88.35	79.54	88.51	88.70	87.85	86.91	88.11	88.73	88.23	**89.02**	95.90
NERC Ancora	92.46	92.15	85.66	93.34	**93.57**	92.58	92.58	92.47	93.02	92.45	93.13	-
NERC Capitel	87.72	88.56	80.35	89.60	89.83	88.65	88.10	88.55	89.86	**90.51**	90.19	90.34
STS	81.59	79.45	80.63	**85.33**	83.82	83.09	81.64	83.47	83.61	84.11	84.04	-
MLDoc	**97.14**	96.68	95.65	96.64	96.78	96.70	96.17	96.30	96.62	97.02	97.05	96.80
PAWS-X	89.15	90.35	89.20	90.45	90.75	89.15	89.30	90.35	**92.50**	90.95	92.05	90.70
XNLI	81.30	78.90	78.78	80.16	82.40	79.40	78.76	81.14	84.85	82.63	**84.95**	85.50
SQAC	79.23	76.78	73.83	79.23	78.91	77.38	75.62	77.28	80.78	82.02	**84.10**	-
CoMeta	64.28	61.52	61.18	63.08	64.79	62.04	61.77	63.82	**67.46**	62.02	67.44	67.46
Average	87.10	86.41	83.78	87.65	87.95	86.76	86.22	87.21	88.67	87.98	**89.09**	
Average*	89.39	88.90	86.04	90.11	90.27	89.23	88.67	89.55	90.79	90.58	91.25	
Wins group	1.5			1.5	8	1			10	2	9	
Wins all	1			1	1				3	1	**4**	

Table 3: Same results as in Table 2, but using standard metrics (accuracy for PAWS-X, word accuracy for PoS UD and Capitel). We also report previous state-of-the-art results where available. See text for details.

language models in Spanish (including this one). Out of the six remaining tasks, the best results of the models on NERC CoNLL and XNLI remain far from the state-of-the-art reported by Wang et al. (2021) and Aghajanyan et al. (2021), with a 95.90 F1 score for NERC and 85.50 in accuracy in XNLI. For PoS UD, our best model scores 99.20 (mDeBERTa), comparable to (Straka, Straková, and Hajic, 2019), which scored 99.05. The same can be said regarding NERC Capitel, where the difference between the best score by MarIA large (90.51) and the previous best (90.34) is rather anecdotical (Agerri, 2020), and MLDoc, for which BETO slightly outscores 97.17 vs 96.80, the previous best result published (Lai et al., 2019). Finally, for PAWS-X only XLM-RoBERTa and mDEBERTa clearly outperform the state-of-the-art previously reported by Yang et al. (2019).

Summarizing, out of the 11 datasets, the Spanish monolingual language models obtain minimal better results for three tasks only: PoS UD, NERC Capitel and MLDoc, although the differences are too small to be significant. Furthermore, they underperform in the rest of the tasks with respect to previously published state-of-the-art results.

What should be the next steps for Spanish models? One could argue that given the better results of the multilingual models released by large companies, there is no need to devote resources to build better models for Spanish. Unfortunately, there is no guarantee that large companies will keep releasing updated models, which will make the models obsolete very quickly. As an example, all models are trained on texts before Covid-19, and thus have no notion of what the latest pandemic is about. It will also leave the leadership of NLP for Spanish at the hand of third parties. Given the foundational nature of language models it is necessary to ensure that new updated versions of the best performance are produced regularly.

Our analysis has shown that it is not trivial to produce high-performance language models, as it is still an open, resource-heavy, research problem. In addition, new and powerful models are being developed at a fast pace, including encoder-decoder models like T5 (Raffel et al., 2020), with its superior performance in many downstream tasks when compared to encoder-only models like BERT (Devlin et al., 2019), or decoder-only models like GPT-3 (Brown et al., 2020), which has facilitated good results in generation tasks, but also in zero- and few-shot approaches to regular NLP tasks (Brown et al., 2020).

In other countries other than Spain, policy-makers and research funding agencies have recognised the strategic importance of this field and its research-intensive and ambitious nature. For example, the European Language Equality (ELE) project[12] has defined an European strategy where three main requirements are identified: expert re-

[12] https://european-language-equality.eu

searchers, (public) data, and computational power (GPUs). However, expert researchers with experience in this field do not abound, and the GPUs needed are a substantial investment which should be carefully designed to meet the demands of training language models.

In our opinion, it is necessary to launch a multi-year research program devoted to language models in Spanish, which should match the ambition of this strategic field and which should marry the following: (i) The expertise of the best researchers in the field of language models. Unfortunately they are a scarce resource, as they are being actively recruited by large companies. We believe that only an attractive research landscape which includes the resources mentioned next will allow to attract them to this program. (ii) The necessary resources, either monetary or in the form of sustained access to powerful GPUs. In order to explore and understand the reasons for the results reported here, it is necessary to set an experimental program where variants of language models are trained on different experimental conditions.

7 Conclusion

In this paper we have presented a comprehensive head-to-head comparison of language models for Spanish. The results show that (i) multilingual models from large companies fare better than monolingual models; (ii) results across the monolingual Spanish models are not conclusive, with supposedly smaller and inferior models performing competitively. Based on these empirical results, we have argued for the need of further research to understand the factors underlying these results. Thus, the effect of corpus size, quality and pre-training techniques need to be further investigated to be able to obtain Spanish monolingual models significantly better than the multilingual ones released by large private companies, specially in the face of rapid ongoing progress in the field.

While the recent activity in the development of language technology for Spanish is to be welcomed, our results show that building language models remains an open, resource-heavy problem which requires to marry monetary and computational resources with the best research expertise and practice.

Other future work should include GPT-3 style improvements at scale for Spanish. Furthermore, most of the current few-shot and generative-related work for languages other than English is being done with multilingual models such as mBART and mT5. Thus, a lot of work remains to be done if Spanish as language is to be at the forefront of language technology.

Acknowledgments

We would like to thank the authors of MarIA models for their valuable help in using their evaluation scripts. This has allow us to follow the same evaluation methodology thereby facilitating comparability of results.

This work has been partially supported by the HiTZ center and the Basque Government (Research group funding IT-1805-22). We also acknowledge the funding from the following projects: (i) DeepKnowledge (PID2021-127777OB-C21) MCIN/AEI/10.13039/501100011033 and ERDF A way of making Europe; (ii) Disargue (TED2021-130810B-C21), MCIN/AEI/10.13039/501100011033 and European Union NextGenerationEU/PRTR (iii) Antidote (PCI2020-120717-2), MCIN/AEI/10.13039/501100011033 and by European Union NextGenerationEU/PRTR; (iv) DeepR3 (TED2021-130295B-C31) by MCIN/AEI/10.13039/501100011033 and EU NextGeneration programme EU/PRTR. Rodrigo Agerri currently holds the RYC-2017-23647 fellowship (MCIN/AEI/10.13039/501100011033 and by ESF Investing in your future).

References

Agerri, R. 2020. Projecting heterogeneous annotations for named entity recognition. In *IberLEF@SEPLN*.

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 *LREC 2020*, pages 4781–4788.

Aghajanyan, A., A. Shrivastava, A. Gupta, N. Goyal, L. Zettlemoyer, and S. Gupta. 2021. Better fine-tuning by reducing representational collapse. In *ICLR*.

Agirre, E., C. Banea, C. Cardie, D. M. Cer, M. T. Diab, A. Gonzalez-Agirre, W. Guo, I. Lopez-Gazpio, M. Maritxalar, R. Mihalcea, G. Rigau, L. Uria, and

J. Wiebe. 2015. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *SemEval@NAACL-HLT*.

Agirre, E., C. Banea, C. Cardie, D. M. Cer, M. T. Diab, A. Gonzalez-Agirre, W. Guo, R. Mihalcea, G. Rigau, and J. Wiebe. 2014. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In *SemEval*.

Armengol-Estapé, J., C. P. Carrino, C. Rodriguez-Penagos, O. de Gibert Bonet, C. Armentano-Oller, A. Gonzalez-Agirre, M. Melero, and M. Villegas. 2021. Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Artetxe, M., I. Aldabe, R. Agerri, O. P. de Viñaspre, and A. S. Etxabe. 2022. Does corpus quality really matter for low-resource languages? In *EMNLP*.

Bhattacharjee, A., T. Hasan, K. Samin, M. S. Rahman, A. Iqbal, and R. Shahriyar. 2021. BanglaBERT: Combating Embedding Barrier for Low-Resource Language Understanding. In *ArXiv*, volume abs/2101.00204.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. 2020. Language models are few-shot learners. In *arXiv*, volume 2005.14165.

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*.

Clark, K., M.-T. Luong, Q. V. Le, and C. D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *ICLR*.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *ACL*.

Conneau, A., G. Lample, R. Rinott, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *EMNLP*.

De la Rosa, J., E. G. Ponferrada, M. Romero, P. Villegas, P. G. de Prado Salas, and M. Grandury. 2022. BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling. *Procesamiento del Lenguaje Natural*, 68:13–23.

de Vries, W., A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, and M. Nissim. 2019. BERTje: A Dutch BERT Model. In *ArXiv*, volume abs/1912.09582.

Devlin, J., M. Chang, K. Lee, and K. Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, and M. Villegas. 2022. MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, 68:39–60.

He, P., J. Gao, and W. Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In *ArXiv*, volume abs/2111.09543.

Komatsuzaki, A. 2019. One Epoch Is All You Need. In *ArXiv*.

Kreutzer, J., I. Caswell, L. Wang, A. Wahab, D. van Esch, N. Ulzii-Orshikh, A. Tapo, N. Subramani, A. Sokolov, C. Sikasote, M. Setyawan, S. Sarin, S. Samb, B. Sagot, C. Rivera, A. Rios, I. Papadimitriou, S. Osei, P. O. Suarez, I. Orife, K. Ogueji, A. N. Rubungo, T. Q. Nguyen, M. Müller, A. Müller, S. H. Muhammad,

N. Muhammad, A. Mnyakeni, J. Mirzakhalov, T. Matangira, C. Leong, N. Lawson, S. Kudugunta, Y. Jernite, M. Jenny, O. Firat, B. F. P. Dossou, S. Dlamini, N. de Silva, S. Çabuk Ballı, S. Biderman, A. Battisti, A. Baruwa, A. Bapna, P. Baljekar, I. A. Azime, A. Awokoya, D. Ataman, O. Ahia, O. Ahia, S. Agrawal, and M. Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Lai, G., B. Oğuz, Y. Yang, and V. Stoyanov. 2019. Bridging the domain gap in cross-lingual document classification. In *ArXiv*, volume abs/1909.07009.

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. In *ArXiv*, volume abs/1907.11692.

Martin, L., B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot. 2020. CamemBERT: a tasty French language model. In *ACL*.

Nozza, D., F. Bianchi, and D. Hovy. 2020. What the [MASK]? Making Sense of Language-Specific BERT Models. In *ArXiv*, volume abs/2003.02912.

Nzeyimana, A. and A. N. Rubungo. 2022. KinyaBERT: a Morphology-aware Kinyarwanda Language Model. In *ACL*.

Ortiz Suárez, P. J., B. Sagot, and L. Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In P. Bański, A. Barbaresi, H. Biber, E. Breiteneder, S. Clematide, M. Kupietz, H. Lüngen, and C. Iliadi, editors, *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, pages 9–16.

Otegi, A., A. Gonzalez-Agirre, J. A. Campos, A. S. Etxabe, and E. Agirre. 2020. Conversational Question Answering in Low Resource Scenarios: A Dataset and Case Study for Basque. In *LREC*.

Pires, T. J. P., E. Schlinger, and D. Garrette. 2019. How Multilingual is Multilingual BERT? In *ACL*.

Porta, J. and L. Espinosa-Anke. 2020. Overview of CAPITEL Shared Tasks at IberLEF 2020: Named Entity Recognition and Universal Dependencies Parsing. In *IberLEF@SEPLN*.

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.

Sanchez-Bayona, E. and R. Agerri. 2022. Leveraging a new spanish corpus for multilingual and crosslingual metaphor detection. In *CoNLL*.

Scao, T. L., A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, J. Tow, A. M. Rush, S. Biderman, A. Webson, P. S. Ammanamanchi, T. Wang, B. Sagot, N. Muennighoff, A. V. del Moral, O. Ruwase, R. Bawden, S. Bekman, A. McMillan-Major, I. Beltagy, H. Nguyen, L. Saulnier, S. Tan, P. O. Suarez, V. Sanh, H. Laurençon, Y. Jernite, J. Launay, M. Mitchell, C. Raffel, A. Gokaslan, A. Simhi, A. Soroa, A. F. Aji, A. Alfassy, A. Rogers, A. K. Nitzav, C. Xu, C. Mou, C. Emezue, C. Klamm, C. Leong, D. van Strien, D. I. Adelani, D. Radev, E. G. Ponferrada, E. Levkovizh, E. Kim, E. B. Natan, F. De Toni, G. Dupont, G. Kruszewski, G. Pistilli, H. Elsahar, H. Benyamina, H. Tran, I. Yu, I. Abdulmumin, I. Johnson, I. Gonzalez-Dios, J. de la Rosa, J. Chim, J. Dodge, J. Zhu, J. Chang, J. Frohberg, J. Tobing, J. Bhattacharjee, K. Almubarak, K. Chen, K. Lo, L. Von Werra, L. Weber, L. Phan, L. B. allal, L. Tanguy, M. Dey, M. R. Muñoz, M. Masoud, M. Grandury, M. Šaško, M. Huang, M. Coavoux, M. Singh, M. T.-J. Jiang, M. C. Vu, M. A. Jauhar, M. Ghaleb, N. Subramani, N. Kassner, N. Khamis, O. Nguyen, O. Espejel, O. de Gibert, P. Villegas, P. Henderson, P. Colombo, P. Amuok, Q. Lhoest, R. Harliman, R. Bommasani, R. L. López, R. Ribeiro, S. Osei, S. Pyysalo, S. Nagel, S. Bose, S. H. Muhammad, S. Sharma, S. Longpre, S. Nikpoor, S. Silberberg, S. Pai, S. Zink, T. T. Torrent, T. Schick, T. Thrush, V. Danchev, V. Nikoulina,

V. Laippala, V. Lepercq, V. Prabhu, Z. Alyafeai, Z. Talat, A. Raja, B. Heinzerling, C. Si, E. Salesky, S. J. Mielke, W. Y. Lee, A. Sharma, A. Santilli, A. Chaffin, A. Stiegler, D. Datta, E. Szczechla, G. Chhablani, H. Wang, H. Pandey, H. Strobelt, J. A. Fries, J. Rozen, L. Gao, L. Sutawika, M. S. Bari, M. S. Al-shaibani, M. Manica, N. Nayak, R. Teehan, S. Albanie, S. Shen, S. Ben-David, S. H. Bach, T. Kim, T. Bers, T. Fevry, T. Neeraj, U. Thakker, V. Raunak, X. Tang, Z.-X. Yong, Z. Sun, S. Brody, Y. Uri, H. Tojarieh, A. Roberts, H. W. Chung, J. Tae, J. Phang, O. Press, C. Li, D. Narayanan, H. Bourfoune, J. Casper, J. Rasley, M. Ryabinin, M. Mishra, M. Zhang, M. Shoeybi, M. Peyrounette, N. Patry, N. Tazi, O. Sanseviero, P. von Platen, P. Cornette, P. F. Lavallée, R. Lacroix, S. Rajbhandari, S. Gandhi, S. Smith, S. Requena, S. Patil, T. Dettmers, A. Baruwa, A. Singh, A. Cheveleva, A.-L. Ligozat, A. Subramonian, A. Névéol, C. Lovering, D. Garrette, D. Tunuguntla, E. Reiter, E. Taktasheva, E. Voloshina, E. Bogdanov, G. I. Winata, H. Schoelkopf, J.-C. Kalo, J. Novikova, J. Z. Forde, J. Clive, J. Kasai, K. Kawamura, L. Hazan, M. Carpuat, M. Clinciu, N. Kim, N. Cheng, O. Serikov, O. Antverg, O. van der Wal, R. Zhang, R. Zhang, S. Gehrmann, S. Pais, T. Shavrina, T. Scialom, T. Yun, T. Limisiewicz, V. Rieser, V. Protasov, V. Mikhailov, Y. Pruksachatkun, Y. Belinkov, Z. Bamberger, Z. Kasner, A. Rueda, A. Pestana, A. Feizpour, A. Khan, A. Faranak, A. Santos, A. Hevia, A. Unldreaj, A. Aghagol, A. Abdollahi, A. Tammour, A. HajiHosseini, B. Behroozi, B. Ajibade, B. Saxena, C. M. Ferrandis, D. Contractor, D. Lansky, D. David, D. Kiela, D. A. Nguyen, E. Tan, E. Baylor, E. Ozoani, F. Mirza, F. Ononiwu, H. Rezanejad, H. Jones, I. Bhattacharya, I. Solaiman, I. Sedenko, I. Nejadgholi, J. Passmore, J. Seltzer, J. B. Sanz, K. Fort, L. Dutra, M. Samagaio, M. Elbadri, M. Mieskes, M. Gerchick, M. Akinlolu, M. McKenna, M. Qiu, M. Ghauri, M. Burynok, N. Abrar, N. Rajani, N. Elkott, N. Fahmy, O. Samuel, R. An, R. Kromann, R. Hao, S. Alizadeh, S. Shubber, S. Wang, S. Roy, S. Viguier, T. Le, T. Oyebade, T. Le, Y. Yang, Z. Nguyen, A. R. Kashyap, A. Palasciano, A. Callahan, A. Shukla, A. Miranda-Escalada, A. Singh, B. Beilharz, B. Wang, C. Brito, C. Zhou, C. Jain, C. Xu, C. Fourrier, D. L. Periñán, D. Molano, D. Yu, E. Manjavacas, F. Barth, F. Fuhrimann, G. Altay, G. Bayrak, G. Burns, H. U. Vrabec, I. Bello, I. Dash, J. Kang, J. Giorgi, J. Golde, J. D. Posada, K. R. Sivaraman, L. Bulchandani, L. Liu, L. Shinzato, M. H. de Bykhovetz, M. Takeuchi, M. Pàmies, M. A. Castillo, M. Nezhurina, M. Sänger, M. Samwald, M. Cullan, M. Weinberg, M. De Wolf, M. Mihaljcic, M. Liu, M. Freidank, M. Kang, N. Seelam, N. Dahlberg, N. M. Broad, N. Muellner, P. Fung, P. Haller, R. Chandrasekhar, R. Eisenberg, R. Martin, R. Canalli, R. Su, R. Su, S. Cahyawijaya, S. Garda, S. S. Deshmukh, S. Mishra, S. Kiblawi, S. Ott, S. Sang-aroonsiri, S. Kumar, S. Schweter, S. Bharati, T. Laud, T. Gigant, T. Kainuma, W. Kusa, Y. Labrak, Y. S. Bajaj, Y. Venkatraman, Y. Xu, Y. Xu, Y. Xu, Z. Tan, Z. Xie, Z. Ye, M. Bras, Y. Belkada, and T. Wolf. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. In *arXiv*.

Schwenk, H. and X. Li. 2018. A corpus for multilingual document classification in eight languages. In *LREC*.

Straka, M., J. Straková, and J. Hajic. 2019. Evaluating Contextualized Embeddings on 54 Languages in POS Tagging, Lemmatization and Dependency Parsing. In *ArXiv*, volume abs/1908.07448.

Tanvir, H., C. Kittask, and K. Sirts. 2021. EstBERT: A Pretrained Language-Specific BERT for Estonian. In *NODALIDA*.

Taulé, M., M. A. Martí, and M. Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *LREC*.

Tiedemann, J. and S. Thottingal. 2020. OPUS-MT – Building open translation services for the World. In *European Association for Machine Translation Conferences/Workshops*.

Tjong-Kim-Sang, E. 2002. Introduction to the CoNLL-2002 Shared Task: Language-

Independent Named Entity Recognition. In *CoNLL*.

Urbizu, G., I. San Vicente, X. Saralegi, R. Agerri, and A. Soroa. 2022. BasqueGLUE: A natural language understanding benchmark for Basque. In *LREC*.

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. In *ArXiv*, volume abs/1912.07076.

Wang, X., Y. Jiang, N. Bach, T. Wang, Z. Huang, F. Huang, and K. Tu. 2021. Automated concatenation of embeddings for structured prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP*.

Wu, S. and M. Dredze. 2020. Are All Languages Created Equal in Multilingual BERT? In *Workshop on Representation Learning for NLP*.

Xue, L., N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *NAACL*.

Yang, Y., Y. Zhang, C. Tar, and J. Baldridge. 2019. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *EMNLP*.

Zhang, S., S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. In *arXiv*.

Zheng, B., L. Dong, S. Huang, S. Singhal, W. Che, T. Liu, X. Song, and F. Wei. 2021. Allocating large vocabulary capacity for cross-lingual language model pre-training. In *EMNLP*.

Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 171-185 recibido 16-12-2022 revisado 31-01-2023 aceptado 09-02-2023

Named Entity Recognition: a Survey for the Portuguese Language

Reconocimiento de Entidades Nombradas: una investigación para el idioma Portugués

Hidelberg O. Albuquerque[1,2], Ellen Souza[1,3], Carlos Gomes[4], Matheus Henrique de C. Pinto[3], Ricardo P. S. Filho[5], Rosimeire Costa[5], Vinícius Teixeira de M. Lopes[6], Nádia F. F. da Silva[3,5], André C. P. L. F. de Carvalho[3], Adriano L. I. Oliveira[2]

[1]MiningBR Research Group, Federal Rural University of Pernambuco, Brazil
[2]Centre of Informatics, Federal University of Pernambuco, Brazil
[3]Institute of Mathematics and Computer Sciences, University of São Paulo, Brazil
[4]Institute of Mathematics and Technology, Federal University of Catalão
[5]Institute of Informatics, Federal University of Goiás, Brazil
[6]Federal University of Campina Grande
{hidelberg.albuquerque,ellen.ramos}@ufrpe.br, alio@cin.ufpe.br,
andre@icmc.usp.br, matheuscerqueira@usp.br, cadyoba@gmail.com,
{rpsfilho93,rosimeire_pereira}@discente.ufg.br, nadia.felix@ufg.br,
vinicius.teixeira.melo.lopes@ccc.ufcg.edu.br

Abstract: Named Entity Recognition (NER) is an important task in Natural Language Processing, as it is a key information extraction sub-task with numerous applications, such as information retrieval and machine learning. However, resources are still scarce for some languages, as it is the case of Portuguese. Thus, the objective of this research is to map NER techniques, methods and resources for the Portuguese language. Manual and automated searches were applied, retrieving 447 primary studies, of which 45 were included in our review. The growing number of studies reveal a greater interest of researchers in the area. 21 studies focused on the comparative analysis between techniques and tools. 24 new or updated NER corpora were mapped, in several domains. The most used text pre-processing techniques were tokenization, embeddings, and PoS Tagging, while the most used methods/algorithms were based on BiLSTM, CRF, and BERT models. The most relevant researchers, institutions and countries were also mapped, as well as the evolution of publications.
Keywords: Named Entities Recognition, Review, Portuguese.

Resumen: El Reconocimiento de Entidades con Nombre (en inglés, *NER*) es una tarea importante en el Procesamiento del Lenguaje Natural, ya que es una subtarea clave de extracción de información con numerosas aplicaciones, como la recuperación de información y el aprendizaje automático. Sin embargo, los recursos aún son escasos para algunos idiomas, como es el caso del portugués. Por lo tanto, el objetivo de esta investigación es mapear técnicas, métodos y recursos de NER para la lengua portuguesa. Se aplicaron búsquedas manuales y automatizadas, recuperando 447 estudios primarios, de los cuales 45 se incluyeron en nuestra revisión. El creciente número de estudios revela un mayor interés de los investigadores en el área. 21 estudios se centraron en el análisis comparativo entre técnicas y herramientas. Se mapearon 24 corpora NER nuevos o actualizados, en varios dominios. Las técnicas de preprocesamiento de texto más utilizadas fueron *tokenization*, *embeddings* y *PoS Tagging*, mientras que los métodos/algoritmos más utilizados fueron los basados en *BiLSTM*, *CRF* y de los modelos *BERT*. También se mapearon los investigadores, instituciones y países más relevantes, así como la evolución de las publicaciones.
Palabras clave: Reconocimiento de Entidades Nombradas, Revisión, Portugués.

ISSN 1135-5948 DOI 10.26342/2023-70-14

1 Introduction

Natural Language Processing (NLP) is one of the multidisciplinary areas involving the fields of Linguistics and Artificial Intelligence. NLP is a challenge for researchers and professionals because it corresponds to how natural language, with all its richness, complexities, and variances can be transformed and used by computational systems (Finatto, Lopes, and Silva, 2015). Named Entity Recognition (NER) is a NLP technique that aims to identify entities in the text and classify them into sets of universal syntactic or semantic categories (Maynard, Bontcheva, and Augenstein, 2016), or the ones specific to a particular language or domain (De Araujo et al., 2020). The classified data and the extracted features are used in text mining systems (Nadeau and Sekine, 2007) or in Machine Learning models (Bonifacio et al., 2020), and other applications.

For the recovery, processing and textual analysis to be effective, it is important to determine which methods are the best for each domain or language, which can explain why much of the research focuses on a monolinguistic approach (Akbik et al., 2016). Studies about NER for Portuguese show evidence that the models used for this language have challenges not found for other languages, which can be explained by the low volume of corpora, tools and pre-trained models developed for Portuguese (Castro, 2018). Researches in this language need a greater effort, mainly in the development of resources, approaches and tools, as occurs to English language (Pirovani, 2019).

Based on the guidelines proposed by Kitchenham, Charters, and others (2007) and Petersen et al. (2008), the main objective of this work is to characterize the current researches that report the use of techniques for NER in Portuguese, seeking to answer the general research question: *What is the current status of NER tasks for the Portuguese language?*

In this way, the automated and manual search procedures retrieved 447 papers published between January/2010 and June/2022 from which 63 were pre-selected and 45 were included in this study. Data extracted from primary studies were systematically structured and analyzed to answer historical, descriptive, and classificatory research questions presented below:

- RQ1: What are the existing corpora for NER in Portuguese Language?
- RQ2: What algorithms, techniques, and tools were used to build and validate the Portuguese NER models?
- RQ3: How the Portuguese NER models have been used in NLP tasks?
- RQ4: What has been the evolution of the number of publications until the year 2022?
- RQ5: What individuals, organizations and countries are the main contributors in this research area?

The remainder of this paper is structured as follows: Section 2 presents the related work. Section 3 details the review method. In Section 4, a comprehensive set of results is presented. Section 5 discusses the results, and contains conclusions and directions for future works. Finally, the Appendix A presents the list of primary studies selected, with their respective access links.

2 Related Work

The mapping of techniques, methods and resources are an indisputable key to the progress of any research, and an invaluable source for any researcher. This section highlights some initiatives with relevant contributions.

Nadeau and Sekine (2007) performed a survey on Named Entity Recognition and Classification (NERC), in a hundred studies published in English in ten different events, between 1991 and 2006. The review reports studies performed in over 20 languages, a wide range of named entity types, their semantic challenges, and hierarchical subcategories. Most studies have focused on limited domains and textual genres. The work also provides an overview of the studies selected from the challenges or techniques used: diversity of languages, types of text used, textual genres, application domains, corpora, disambiguation rules, machine learning algorithms, features, as well as evaluation methods. Finally, the work highlights the great importance of NER for NLP-based systems.

Sun et al. (2018) conducted a research using 162 publications from NLP conferences, between the years 1996 to 2017. The authors discuss about two aspects of research in NER: the first one, based on target languages (covering papers from more than 200 languages,

with mono, bi, and multilingual approaches), and the second one, a more technical approach with statistical analysis used in NER tasks. Some results brought by the authors were the mapping of the number of publications, the most used languages, the proportion between publications with different approaches, and the different methods.

Yadav and Bethard (2019) explore the advances of recent architectures with better deep learning results for state-of-the-art. Studies that combined learning models based on minimal resources were selected, which were compared with models of feature-based learning and with different representations of words. An automatic search was used in three search engines, with a search string. The papers were initially classified by total of citations, being pre-selected those that used an unpublished NER neural architecture, or a representation of a high-performance model for NER datasets, independent of domain or language. When published architecture was found, citation tracing back to the architecture's original source was performed. 154 papers were reviewed and 83 were selected. This results were subdivided into NER datasets, evaluation metrics and systems based on different techniques and architectures. The authors compared the results found in four languages (Spanish, Dutch, English and German), highlighting the need for future progress using insights from previous work applied to current neural network models.

Li et al. (2020) also focuses on studies of deep learning models for NER. After a brief review of traditional NER techniques, the authors make an intense review of studies, applications and deep learning techniques for NER, using universal entities in English. It was proposed a new taxonomy, which systematically organizes the approaches along three axes: distributed representations for input, context encoder (to capture contextual dependencies), and tag decoder (to predict word labels). In addition, the paper presents relevant secondary results, such as, corpora annotated in English, NER tools, summary of recent works on neural NER, besides presenting challenges and future directions.

3 Review Method

Secondary studies review all the primary studies relating to a specific research question with the aim of integrating/synthesizing evidence related to a subject (Keele and others, 2007). In this study, the search for primary studies was done in six steps, as shown in Fig. 1, detailed in the following subsections.

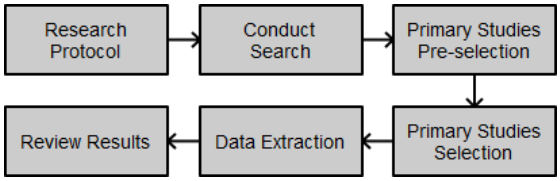


Figure 1: Review method (Keele and others, 2007).

3.1 Research Protocol

The research protocol outcome is the review scope, which includes, among other things, the research questions presented in Section 1, the inclusion and exclusion criteria, and data sources selection. Primary studies that reported NER corpora, algorithms, methods, and techniques for Portuguese were searched in the literature, in the last 12 years. Studies that met at least one of the following exclusion criteria were removed from this review: (i) written in a language other than English or Portuguese; (ii) not available on online scientific libraries; (iii) keynote speeches, workshop reports, books, theses and dissertations.

3.2 Conduct Search

Two different types of searches were performed: automated and manual. First, a manual search was performed, followed by the automatic one, performing the removal of duplicate studies. In the former, a search string was used to retrieve papers from digital libraries, using the following terms and their synonyms:

- Named Entity Recognition: NER, Recognition of Entities, *Reconhecimento de Entidades Nomeadas*, *REN*, *Reconhecimento de Entidades Mencionadas*, *REM*, *Reconhecimento de Entidades com Nome*, Entity Extraction Task, Name Entity, NE.
- Portuguese: Portuguese, *Língua Portuguesa*, *Português*.

ACL Anthology[1] and Google Scholar[2] digital libraries were selected to conduct

[1]https://aclanthology.org
[2]https://scholar.google.com

the automated searches, covering the period between January/2010 and June/2022. The ACL Anthology currently hosts 80,558 papers on the study of computational linguistics and natural language processing, indexing several events and journals in the NLP area.

For the manual search, we selected the main venues focusing on the Portuguese language (Souza et al., 2016; Souza et al., 2018): The International Conference on the Computational Processing of Portuguese (PROPOR); Portuguese Conference on Artificial Intelligence (EPIA); Brazilian Symposium on Information and Human Language Technology (STIL); Brazilian Conference on Intelligent Systems (BRACIS); Brazilian National Meeting of Artificial and Computational Intelligence (ENIAC) and Language Resources and Evaluation Conference (LREC), which is the major event on Language Resources and Evaluation for Language Technologies in several languages.

Manual Search	**Pre-sel.**	**Sel.**
BRACIS	7	6
ENIAC	3	3
EPIA	2	1
LREC	7	4
PROPOR	3	3
STIL	7	5
TOTAL	**29**	**22**
Automated Search	**Pre-sel.**	**Sel.**
EPIA	2	1
IberLEF	4	4
LREC	5	3
Events with one paper	17	9
Journals with one paper	6	6
TOTAL	**34**	**23**
Manual+Automated	**63**	**45**

Table 1: Review by data sources.

3.3 Primary Studies: Pre-selection and Selection

A pre-selection was accomplished in accordance with the inclusion and exclusion criteria established in the review protocol. Primary studies were analyzed using the same procedure for both automated and manual search strategies. Two researchers applied the inclusion and exclusion criteria, after reading the title, abstract, and keywords.

The selection process was done by six researchers. Each researcher was responsible for reading the *full* paper and presenting it at a weekly consensus meeting, where all researchers decided to include or exclude the primary studies. Table 1 shows the number of potentially relevant primary studies (Pre-selection step) and number of included studies (Selection step).

3.4 Data Extraction and Review Results

In this fifth step, data from included primary studies were extracted and synthetized to answer the research questions. The researchers worked independently to extract data from the included papers, using an extraction form. Finally, one researcher inspected the extracted data and *ad hoc* consensus online meetings were held. The details of the extraction form with its results were packed for later replication[3].

4 Results

In this section, the obtained results are presented, organized according to the five specific research questions.

4.1 RQ1: What are the existing corpora for NER in Portuguese Language?

Information about corpora, entities, annotation method, agreement level, domain, type of text used, and language variants were extracted from primary studies (PS).

In general, the studies presented a pattern in the use of manual annotation, absence of agreement level measure among annotators, use of formal texts in the construction of the corpora, and absence of comparison of entities for corpora with the same domain. The works that differ are: (i) PS05, PS07, PS21 and PS22, which used automatic annotation; PS06, PS17, PS25, PS34 and PS45 used hybrid annotation, and PS43 does not inform the used method; (ii) for the agreement measure, PS19 presented a measure of 95.8 % (without specifying which one was used) and PS01, in general, 91 % (using Cohen's Kappa); (iii) regarding the type of text, PS18 do not inform which type was used, PS09 used a mix between formal and informal usage. PS15, PS16, PS22, PS42, PS43, and PS45 used informal texts; (iv) regarding the difference of entities in the same domain, PS24 used only two entities against six of

[3]Available in https://bit.ly/extraction-form-survey

Corpora/PS	Entities	Annotation method	Domain	Text type	Language variant
Aposentadoria/PS25	Act, Act_Name, Class, Cod_Enrollment_Act, Company_Act, Legal_Fund, Position, Frame, Pattern, and Process.	Hybrid	Legal	Formal	PT-BR
DataSense NER Corpus/PS11	Bank Identification Number, Credit Card Number, Date, Driving License Number, E-mail address, Identification Number, Job, Local, Med, National Health Number, Organization, Passport Number, Person, Postal Code, Social Security Number, Tax Identification Number, Telephone Number, and Value.	Manual	Sensitive Data	Formal	PT-EU
Dicionário Histórico-Biográfico Brasileiro (DHBB)/PS17	Document, Event, Local, Organization, Person, Political Formulation, and Time.	Hybrid	History	Formal	PT-BR
DrugSeizures-Br/PS07	Drug, Location, Organization, Other, Person, and Time.	Automatic	Legal	Formal	PT-BR
EHR-Names/PS40	Person	Manual	Medical	Formal	PT-BR
Financial Market Corpus/PS34	Organization, Person and Place	Hybrid	Financial	Formal	PT-BR
GeoCorpus/PS03 and GeoCorpus-2/PS10	Aeon, Era, Period, Epoc, Age, Siliciclastic Sedimentary Rock, Carbonate Sedimentary Rock, Chemical Sedimentary Rock, Organic-rich Sedimentary Rock, Brazilian Sedimentary Basin, Basin Geological Context, Lithostratigraphic Unit, and Miscellaneous.	Manual	Geology	Formal	PT-BR
LeNER-Br/PS20	Legal cases, Legislation, Location, Organization, Person, and Time.	Manual	Legal	Formal	PT-BR
PS06	CPF_CNPJ, Marital status, Name, Nationality, OAB, and RG.	Hybrid	Legal	Formal	PT-BR
PS16	Date, Location, Organization and Person	Manual	General	Informal	PT-EU
PS19	Characterization, Test, Evolution, Genetics, Anatomical Site, Negation, Additional Observations, Condition, Results, DateTime, Therapeutics, Value, and Route of Administration.	Manual	Neurology	Formal	PT-EU
PS22	Location, Organization, and Person	Automatic	Journalistic	Informal	PT-EU
PS23	Location, Organization, and Person.	Manual	Police	Formal	PT-BR
PS24	Legal cases and Legislation	Manual	Legal	Formal	PT-BR
PS35	Person	Manual	Legal	Formal	PT-BR
PS36	Place and Person	Manual	Literature	Formal	PT-BR
PS42	Brand, Camera quality, Color, Display size, Internal memory, Model, Operating system, Processor, SIM card capacity, and WIT (What Is This)	Manual	E-commerce	Informal	PT-BR
PS43	Organization, Person, and Location	*Not Informed*	Jornalistic	Informal	PT-BR
PS45	Location and Event	Hybrid	Traffic	Informal	PT-BR
Second HAREM/PS15	Abstraction, Event, Location, Organization, Other, Person, Thing, Time, Title, and Value	Manual	General	Informal	PT-BR & PT-EU
SESAME/PS21	Location, Organization, and Person	Automatic	General	Formal	PT-BR
Summ-it++/PS05	Abstraction, Event, Organization, Other, Person, Place, Thing, Time, Value, and Work.	Automatic	General	Formal	PT-BR
UlyssesNER-Br/PS01	Date, Event, Law Fundation, Law product, Location, Organization, and Person	Manual	Legislative	Formal	PT-BR

Table 2: Portuguese NER corpora.

PS20; PS37 compares universal entities, such as Person, Place, Organization, Value, Time, Abstraction, Work, Event, and Thing using HAREM and SPA Conll-2002 models; PS01 and PS11 adopt more specificity when compared to the entities of PS15 and PS20. In turn, PS15 updates the golden collection of HAREM (Santos and Cardoso, 2006), presenting the main improvements: removal of repeated texts, cleaning of uncertain sequences, accounting of partially correct entities and systematization in the treatment of entities. HAREM is a huge Portuguese language NER corpus widely used by the Portuguese NLP community.

Among the primary studies that were included, 27 (60 %) did not indicate whether or how interference occurs based on text ty-

Domain	Corpora/PS	Public link
E-commerce	PS42	—
Financial	Financial Market Corpus/PS34	http://bit.ly/finmktcorpus
General	brWaC/PS07, PS38, and PS39	https://bit.ly/BrWaC-corpus
	Floresta Sintática/PS11	https://bit.ly/floresta-corpus
	Freeling/PS04 and PS12	https://bit.ly/freeling-corpus
	HAREM I, HAREM II and MiniHAREM/PS15	http://bit.ly/haremcorpus
	Paramopama/PS09 and PS21	https://bit.ly/paramopama
	PS16	https://bit.ly/ps16-ptools
	SESAME/PS21	https://bit.ly/sesamecorpus
	Summ-it++/PS05	https://bit.ly/summ-it
	WikiNER/PS08, PS09, PS21, PS23, and PS30	https://bit.ly/wikiner
Geology	GeoCorpus/PS03	https://bit.ly/GeoCorpus
	GeoCorpus-2/PS10	https://bit.ly/geocorpus2
History	Dicionário Histórico-Biográfico Brasileiro (DHBB)/PS17	https://bit.ly/DHBB-corpus
Jornalistic	aTribuna/PS08 and PS30	https://bit.ly/atribuna-corpus
	CETEMPúblico/PS43	https://bit.ly/CETEMPublico
	CETENFolha/PS43	https://bit.ly/cetenfolha
	PS22	—
	PS43	—
	SIGARRA News/PS16	https://bit.ly/sigarranews
Legal	Acordaos-TCU/PS07	https://bit.ly/acordaos-tcu
	Aposentadoria/PS25	https://bit.ly/aposentadoria-corpus
	Data-lawyer/PS09	—
	DrugSeizures-Br/PS07	—
	LeNER-Br/PS20	https://bit.ly/lener-br
	PS06	—
	PS24	—
	PS35	—
Legislative	UlyssesNER-Br/PS01	https://bit.ly/ulyssesner-br
Literature	PS36	—
Medical/ Clinical	EHR-Names/PS40	—
	PS19	—
	PS44	https://bit.ly/ps44-BioBERTpt
	SemClinBr/PS44	https://bit.ly/SemClinBr
Police	PS23	—
Sensitive Data	DataSense NER Corpus/PS11	—
Traffic	PS45	—

Table 3: Corpora per domain.

pe or domain in the NER task. Analyzing the studies that provide this information, it is possible to correlate domain specificity to decreased efficiency in the results.

PS21 indicates that the general domain facilitates information extraction and enables cross-referencing of information in order to increase complexity. In the legal domain, PS20 presents unique entities to represent laws and legal cases, PS35 reports that capitalization of proper names increases the generation of false positives, and PS01 is the only study that present entities for the legislative subdomain. In PS10 and PS17, for the Geology and History domains, respectively, the universal entities (such as Person, Organization, and Place) were insufficient to represent the complexity of the research.

In the medical/clinical domain, there is a significant worsening of F-measure compared to the general domain in PS40, caused by the concatenation of the corpora used; in PS41, this worsening is justified by the use of terms and abbreviations unique to the domain, while in PS44, the existence of multiple labels is pointed out as the main factor. PS08, PS09, PS30, and PS32 show that, in comparison, the NER task performs poorly for the clinical domain and reasonably to the police domain when compared to the general domain. In the E-commerce domain, PS42, the lack of syntactic structure makes attribute extraction difficult. Finally, PS43 states that in the journalistic domain, when approaching the gene-

Techniques	Total	Primary Studies
Tokenization	19	PS02, PS06, PS09, PS11, PS19, PS20, PS21, PS25, PS27, PS28, PS31, PS32, PS33, PS35, PS36, PS37, PS39, PS42, PS44
Embeddings	13	PS01, PS06, PS08, PS09, PS10, PS19, PS20, PS24, PS33, PS38, PS39, PS40, PS41
PoS Tagging	10	PS03, PS04, PS05, PS11, PS12, PS14, PS19, PS22, PS28, PS36
Lower case	6	PS20, PS33, PS35, PS37, PS39, PS42
Spell-check	5	PS02, PS16, PS29, PS32, PS34
Format Conversion	5	PS10, PS13, PS14, PS24, PS28
Special character removal	5	PS34, PS35, PS39, PS42, PS45
Features	4	PS02, PS12, PS13, PS32
Non-text removal	4	PS03, PS21, PS39, PS45
Removal of repeated sentences (outliers)	3	PS01, PS10, PS42
Removal of pre-textual or post-textual elements	3	PS03, PS21, PS36
Stop Words	2	PS27, PS42
Removal of HTML Tags	2	PS32, PS39

Table 4: Pre-processing techniques.

ral domain, the task is facilitated, probably due to language simplification.

Few articles presented applications for different language variants. PS12, PS13, PS25, and PS33 do not report whether language variation interferes with the NER tasks. The strategy of PS21 (using the DBPedia ontology[4]), avoids the inclusion of Portuguese language papers. PS42 indicates that, when tuning the BERTimbau model[5] with HAREM, there was a subtle worsening in performance. PS27, when comparing Portuguese variants, registered a difference in performance, justified by linguistic and cultural differences.

From the selected works, 21 ($\sim 47\,\%$) studies did not create, modify or update corpora: PS02, PS04, PS08, PS09, PS12 to PS14, PS18, PS26 to PS33, PS37 to PS39, and PS41. These works carried out comparative analyzes between tools, methods or the application of NER tasks in multidomains, in different textual genres or in textual semantic relations, using some variation of the HAREM corpus. Table 2 summarizes the 24 studies in which a new corpus was created or updated from an existing corpus. Some of them did not have a clearly identifiable name for the created or modified corpus, and were instead referred to by the PS number. Finally, Table 3 presents all corpora used in the studies, organized by domain. Some of these corpora are publicly available. Due to limited space, the entities labels are detailed in the document specified in the third footnote (Section 3.4).

4.2 RQ2: What algorithms, techniques, and tools were used to build and validate the Portuguese NER models?

37 PS ($\sim$82 %) mention the use of some type of preprocessing techniques as shown in Table 4. The most used techniques were tokenization, embeddings, and PoS Tagging. Regarding pre-processing, it was possible to observe that some works focused on the use and analysis of the influence of embeddings and features, among others.

Table 5 presents the most used methods, algorithms, and tools: learning models based on BiLSTM were the most used, followed by more traditional methods that use only CRF, as well as systems developed specifically for NER tasks, such as the NERP-CRF and PALAVRAS parser. More recent works used BERT deep learning model. The metrics used in most studies were Accuracy, Precision, Recall, and F-score. However, as not all papers presented all this information, the table shows the F-score range of performance obtained. Other tools, algorithms or techniques that had only one mention were not listed.

[4]https://www.dbpedia.org/resources/ontology

[5]https://github.com/neuralmind-ai/portuguese-bert

Algorithms/Methods/Tools	Total	F-score(%)	Primary Studies
BiLSTM+CRF	10	67 ∼ 97	PS06, PS09, PS10, PS20, PS24, PS33, PS38, PS39, PS40, PS41
CRF	7	48 ∼ 97	PS01, PS04, PS13, PS19, PS28, PS36, PS45
CRF+LG	6	53 ∼ 76	PS02, PS09, PS29, PS30, PS31, PS32
BiLSTM	3	53 ∼ 93	PS01, PS11, PS43
NERP+CRF	3	53 ∼ 93	PS05, PS12, PS14
PALAVRAS	3	57 ∼ 62	PS04, PS12, PS17
BERT+CRF	2	75 ∼ 95	PS34, PS44
FreeLing	2	54 ∼ 56	PS04, PS12

Table 5: Algorithms/methods/tools.

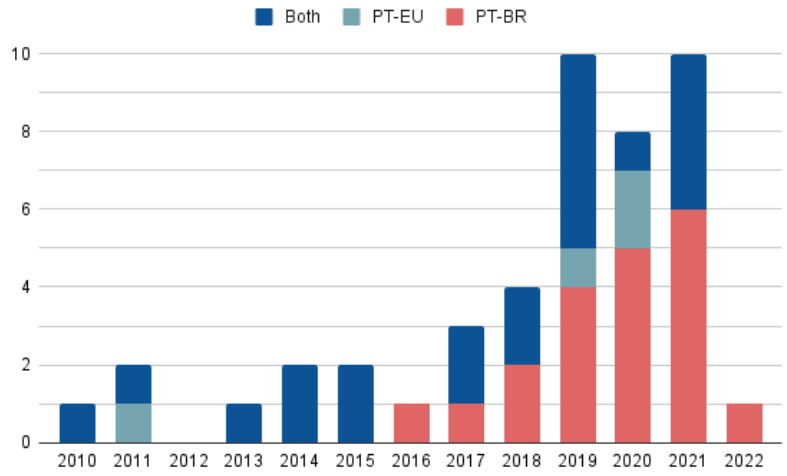


Figure 2: Temporal distribution of PS, between January/2010 to June/2022[6].

4.3 RQ3: How the Portuguese NER models have been used in NLP tasks?

As previously mentioned, a significant part of the selected studies focused their research on comparative analyzes between NER tasks *per se*. Some of these works showed great potential for applicability and/or improvement of information extraction systems, such as studies PS12, PS13, PS20, PS23, and PS29.

Among the selected studies, 20 applied NER models directly or indirectly in other NLP tasks, which are shown in Table 6. Of these, Information Retrieval appears as the main applied task (55 %), followed by Relation Extraction (30 %), Morphosyntactic Annotation and Semantic Similarity tasks (both with 15 %). Two studies explored the impact of language models with Machine Learning, using non-common models in the state-of-the-art, proposing improvements based on the results found.

4.4 RQ4: What has been the evolution of the number of publications until the year 2022?

As shown in Figure 2, the first primary study selected was published in 2010 (PS15). The number of studies ranged in the first years (1∼2 articles per year), increasing from 2017, with the exception of 2022 (the current year of this research[6]). Primary studies were classified according to the Portuguese language variant: studies that apply only European Portuguese represent 8.89 %, while Brazilian Portuguese comprises 44.44 %. Finally, 46.67 % of studies used text written in both languages, most of them with a HAREM corpus variation.

[6]The searches in automated sources were performed in June/2022, which explains the low number of publications for this year.

Task	Total	Subtask	Total	Primary Studies
Information Retrieval	12	-	-	PS01, PS07, PS10, PS11, PS15, PS17, PS18, PS19, PS22, PS42, PS44, PS45
Relation extraction	6	-	-	PS05, PS09, PS15, PS17, PS26, PS34
Morphosyntactic Annotation	3	-	-	PS05, PS17, PS22
Semantic similarity	3	-	-	PS30, PS38, PS44
Classification	2	Document Classification	1	PS18
		Text Classification	1	PS42
Co-reference resolution	2	-	-	PS05, PS15
Data Privacy	2	De-Identification	1	PS40
		Sensitive Data	1	PS11
Machine Learning for NLP	2	Deep Active Learning	1	PS25
		Transfer Learning	1	PS07
Tracking	1	-	-	PS45
Word sense disambiguation	1	-	-	PS44

Table 6: Related NLP tasks and subtasks.

4.5 RQ5: What individuals, organizations and countries are the main contributors in this research area?

A total of 145 researchers from 45 organizations were mapped. The data showed that Brazil has a greater number of researchers (∼79 %) and research institutes (∼67 %) in the area. Tables 7 and 8 list the main researchers and institutions, with emphasis on the Pontifical Catholic University of Rio Grande do Sul (PUCRS) and researchers Renata Vieira, Daniela O. F. do Amaral and Joaquim Santos, from the same institution. The vast majority of institutions in the papers are Colleges, Universities or Institutes of Higher Education (∼71 %). We also believe it is worth mentioning some private institutions that provided support for research in NER, such as Petrobras Research and Development Center (PS10), IBM Research (PS17, PS37, and PS39), Americanas S.A. Digital Lab (PS42), Viatecla SA (PS22), and some institutions linked to public or political administration, like Public Ministry of the State of Mato Grosso do Sul (PS07), Brazilian Federal Police (PS18 and PS23), and Brazilian Chamber of Deputies (PS01).

5 Discussion and Conclusion

Analyzing the data found in the included primary studies, it was possible to observe a growing interest in the development of research in NER for the most diverse fields of the Portuguese language, partly due to its potential applications, e.g., opinion mining, information retrieval systems, development of new general-purpose or domain-specific corpora, and optimization of machine and deep learning models. However, even with the good results achieved, the amount of research is still small when compared to other languages such as English. The amount of private or unpublished corpora and pre-trained learning models could be greater, which would improve the research area.

Looking at the research questions, it is possible to point out some limitations in the included selected studies, among which we highlight: the vast majority of studies did not deepen the discussion about the interference of the Portuguese language variant in the NER tasks; only two works showed the explicit use of a measure of agreement between manual annotators, which could influence the quality of the corpora; among the studies that used hybrid annotation, there is no comparison of the results between the types of annotation; no comparison was found between types of texts, and we think that the use of informal texts could give high complexity and richness to the textual analysis, by expressing with greater precision the colloquial form of the language; no comparison methods were found between annotation processes for entities from the same domain, it was not pointed out if there are semantic differences between entities from different domains.

Regarding the used techniques and algorithms, the most used pre-trained models

Quant.	Author	Institution	Quant.	Author	Institution
14	Renata Vieira	PUCRS-BR	5	Juliana Pirovani	UFES-BR
6	Daniela O.F. do Amaral	PUCRS-BR	4	Sandra Collovini	PUCRS-BR
6	Joaquim Santos	PUCRS-BR	3	Evandro Brasil da fonseca	PUCRS-BR
5	Bernardo Consoli	PUCRS-BR	3	Juliano Terra	PUCRS-BR
5	Elias S. Oliveira	UFES-BR	3	Nádia F. F. da Silva	UFG-BR

Table 7: Number of articles published by main researchers.

Quant.	Institution	Country	Quant.	Institution	Country
14	PUCRS	Brazil	3	IBM Research	Brazil
5	UFES	Brazil	3	PUC-Rio	Brazil
4	UFRPE	Brazil	3	UnB	Brazil
4	University of Évora	Portugal	3	UFG	Brazil

Table 8: Number of researchers per organization.

of machine learning were the classic models from the state-of-the-art, even after the advances of the recent models; more detailed information about used techniques and statistical measures was not found in the vast majority of the works.

There are some threats to the validity of our research that are worth highlighting: (i) it is possible that some relevant studies were not included throughout the search process. We attempted to mitigate this weakness by conducting extensive research as well always observing the research protocol used, carefully comparing the results and removing duplicate studies, and; (ii) as the studies were classified based on personal judgment, it is possible that some studies were classified incorrectly. In order to mitigate this threat, the classification step was performed for more than one researcher.

Finally, we emphasize that the strong point of this work is to promote the growth of research on Named Entities Recognition in the Portuguese Language, through the discrimination of their studies, resources, techniques, researchers, and institutions. We believe that the information described in this work can help other researchers/practitioners in the area to discover what has been researched and achieved, in addition to listing some gaps. The lack of some relevant data and published corpora and tools makes it difficult to carry on analysis in the research area. We plan to apply other mapping techniques to increase coverage, such as snowballing the included primary studies. Besides, an extension of this research is being produced, focusing on NER in the legislative field in several languages.

Acknowledgements

This research is carried out in the context of the Ulysses Project, of the Brazilian Chamber of Deputies. Ellen Souza and Nádia Félix are supported by FAPESP, agreement between University of São Paulo (USP) and the Brazilian Chamber of Deputies. Ricardo P. S. Filho are supported by FUNAPE. André C. P. L. F. de Carvalho and Adriano L. I. Oliveira are supported by CNPq. We express our gratitude to the Brazilian Chamber of Deputies and research funding agencies for their support this research.

References

Akbik, A., L. Chiticariu, M. Danilevsky, Y. Kbrom, Y. Li, and H. Zhu. 2016. Multilingual information extraction with polyglotie. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 268–272.

Bonifacio, L. H., P. A. Vilela, G. R. Lobato, and E. R. Fernandes. 2020. A study on the impact of intradomain finetuning of deep language models for legal named entity recognition in portuguese. In *Brazilian Conference on Intelligent Systems*, pages 648–662. Springer.

Castro, P. 2018. *Deep learning for named entity recognition in legal domain*. Ph.D. thesis, Master's thesis, Universidade Federal de Goiás.

De Araujo, P. H. L., T. E. de Campos, F. A. Braz, and N. C. da Silva. 2020. Victor:

a dataset for brazilian legal documents classification. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1449–1458.

Finatto, M. J. B., L. Lopes, and A. C. Silva. 2015. Processamento de linguagem natural, linguística de corpus e estudos linguísticos: uma parceria bem-sucedida. *Domínios de lingu@gem. Uberlândia, MG. Vol. 9, n. 5 (dez. 2015), p.[41]-59.*

Keele, S. et al. 2007. Guidelines for performing systematic literature reviews in software engineering. Technical report, Technical report, ver. 2.3 ebse technical report. ebse.

Kitchenham, B., S. Charters, et al. 2007. Guidelines for performing systematic literature reviews in software engineering version 2.3. *Engineering*, 45(4ve):1051.

Li, J., A. Sun, J. Han, and C. Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Maynard, D., K. Bontcheva, and I. Augenstein. 2016. Natural language processing for the semantic web. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 6(2):1–194.

Nadeau, D. and S. Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Petersen, K., R. Feldt, S. Mujtaba, and M. Mattsson. 2008. Systematic mapping studies in software engineering. In *12th International Conference on Evaluation and Assessment in Software Engineering (EASE) 12*, pages 1–10.

Pirovani, J. P. C. 2019. *CRF+ LG: uma abordagem híbrida para o reconhecimento de entidades nomeadas em português*. Ph.D. thesis, Universidade Federal do Espírito Santo, Vitória (Brésil).

Santos, D. and N. Cardoso. 2006. A golden resource for named entity recognition in portuguese. In *International Workshop on Computational Processing of the Portuguese Language*, pages 69–79. Springer.

Souza, E., D. Costa, D. W. Castro, D. Vitório, I. Teles, R. Almeida, T. Alves, A. L. I. Oliveira, and C. Gusmão. 2018. Characterising text mining: a systematic mapping review of the portuguese language. *IET Software*, 12(2):49–75.

Souza, E., D. Vitório, D. Castro, A. L. I. Oliveira, and C. Gusmão. 2016. Characterizing opinion mining: A systematic mapping study of the portuguese language. In J. Silva, R. Ribeiro, P. Quaresma, A. Adami, and A. Branco, editors, *Computational Processing of the Portuguese Language*, pages 122–127, Cham. Springer International Publishing.

Sun, P., X. Yang, X. Zhao, and Z. Wang. 2018. An overview of named entity recognition. In *2018 International Conference on Asian Language Processing (IALP)*, pages 273–278.

Yadav, V. and S. Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*.

A Appendix: Primary Studies (PS)

01. Albuquerque, H.O., R. Costa, G. Silvestre, E. Souza, N.F.F. Silva, D. Vitório, G. Moriyama, L. Martins, L. Soezima, A. Nunes, F. Siqueira, J.P. Tarrega, J.V. Beinotti, M. Dias, M. Silva, M. Gardini, V. Silva, A.C.P.L.F. Carvalho, and A.L.I. Oliveira. 2022. UlyssesNER-Br: A Corpus of Brazilian Legislative Documents for Named Entity Recognition. In *Proceedings of 15th edition of the International Conference on the Computational Processing of Portuguese (PROPOR 2022)*. DOI: 10.1007/978-3-030-98305-5_1

02. Alves, D., B. Bekavac, and M. Tadić. 2021. The Optimization of Portuguese Named-Entity Recognition and Classification by Combining Local Grammars and Conditional Random Fields Trained with a Parsed Corpus. In *Proceedings of NooJ 2020 International Conference*. DOI: 10.1007/978-3-030-70629-6_17

03. Amaral, D., S. Collovini, A. Figueira, R. Vieira, and Marco Gonzalez. 2017. Processo de construção de um corpus anotado com Entidades Geológicas visando REN. In *Proceedings of XI Brazilian Symposium in Information and Human Language Technology and Collocated Events (STIL 2017)*. Available in <https://sol.

sbc.org.br/index.php/stil/article/view/4032>.

04. Amaral, D.O.F., E.B. Fonseca, L. Lopes, and R. Vieira. 2014. Comparative Analysis of Portuguese Named Entities Recognition Tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* Available in <http://www.lrec-conf.org/proceedings/lrec2014/index.html>.

05. Antonitsch, A., A. Figueira, D. Amaral, E. Fonseca, R. Vieira, and S. Collovini. 2016. Summ-it++: an enriched version of the Summ-it corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC' 16).* Available in <https://aclanthology.org/L16-1324>.

06. Batista, H.H.N., A.C.A. Nascimento, R. F. Melo, P.B.C. Miranda, I.W.S. Maldonado, and J.L.M. Coelho Filho. 2021. A comparative analysis of text embedding approach to extract named entities in Portuguese legal documents. In *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2021).* DOI: 10.5753/eniac.2021.18255.

07. Bonifacio, L.H., P.A. Vilela, G.R. Lobato, and E.R. Fernandes. 2020. A Study on the Impact of Intradomain Finetuning of Deep Language Models for Legal Named Entity Recognition in Portuguese. In *Proceedings of 9th Brazilian Conference on Intelligent Systems (BRACIS 2020).* DOI: 10.1007/978-3-030-61377-8_46.

08. Castro, P.V.Q., N.F.F. Silva, and A.S. Soares. 2019. Contextual Representations and Semi-Supervised Named Entity Recognition for Portuguese Language. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019).* Available in <https://ceur-ws.org/Vol-2421>.

09. Collovini, S., J. Santos, B. Consoli, J. Terra, R. Vieira, P. Quaresma, M. Souza, D.B. Claro, and R. Glauber. 2019. IberLEF 2019 Portuguese Named Entity Recognition and Relation Extraction Tasks. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019).* Avaliable in <https://ceur-ws.org/Vol-2421>.

10. Consoli, B., J. Santos, D. Gomes, F. Cordeiro. R. Vieira, and V. Moreira. 2020. Embeddings for Named Entity Recognition in Geoscience Portuguese Literature. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020).* Available in <https: aclanthology.org/2020.lrec-1.568>.

11. Dias, M., J. Boné, J.C. Ferreira, R. Ribeiro, and R. Maia. 2020. Named Entity Recognition for Sensitive Data Discovery in Portuguese. Applied Sciences, 10(7):2303. DOI: 10.3390/app10072303

12. Do Amaral, D. O., E. Fonseca, L. Lopes, and R. Vieira. 2014. Comparing NERP-CRF with publicly available Portuguese named entities recognition tools. In *Proceedings of International Conference on Computational Processing of the Portuguese Language (PROPOR 2014).* DOI: 10.1007/978-3-319-09761-9_27.

13. Do Amaral, D.O.F., and R. Vieira. 2013. O Reconhecimento de Entidades Nomeadas por meio de Conditional Random Fields para a Língua Portuguesa. In *Proceedings of IX Brazilian Symposium in Information and Human Language Technology (STIL 2013).* Available in <https: //sites.google.com/usp.br/stil>.

14. Do Amaral, D.O.F., M. Buffet, and R. Vieira. 2015. Comparative Analysis between Notations to Classify Named Entities using Conditional Random Fields. In *Proccedings of X Brazilian Symposium in Information and Human Language Technology and Collocated Events (STIL 2015).* Available in <https://aclanthology.org/W15-5603>.

15. Freitas, C., C. Mota, D. Santos, H.G. Oliveira, and P. Carvalho. 2010. Second HAREM: Advancing the State of the Art of Named Entity Recognition in Portuguese. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).* Avaliable in <https://aclanthology.org/L10-1284>.

16. Gonçalves, M., L. Coheur, J. Baptista, and A. Mineiro. 2020. Avaliação de de Recursos Computacionais para o Português. *Linguamática 2020, 12, 51-68.* DOI: 10.21814/lm.12.2.331

17. Higuchi, S., C. Freitas, B. Cuconato, and A. Rademaker. 2018. Text Mining for History: first steps on building a large dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Available in <https://aclanthology.org/L18-1593>.

18. Junior, O.D., and D.B. Claro. 2011. Uma Análise do Reconhecimento Textual de Nomes de Pessoas e Organizações na Computação Forense. In *Proceddings of Sixth Internacional Conference on Forensic Computer Science (ICoFCS 2011)*. DOI: 10.5769/C2011001

19. Lopes, F., C. Teixeira, H.G. Oliveira. 2019. Named Entity Recognition in Portuguese Neurology Text Using CRF. In *Proceedings of 19th EPIA Conference on Artificial Intelligence (EPIA 2019)*. DOI: 10.1007/978-3-030-30241-2_29

20. Luz de Araujo, P.H., T.E. de Campos, R.R.R. de Oliveira, M. Stauffer, S. Couto, and P. Bermejo. 2018. LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text. In *Proceedings of Computational Processing of the Portuguese Language (PROPOR 2018)*. DOI: 10.1007/978-3-319-99722-3_32

21. Menezes, D., R. Milidiú, and P. Savarese. 2019. Building a Massive Corpus for Named Entity Recognition Using Free Open Data Sources. In *Proceedings of 8th Brazilian Conference on Intelligent Systems (BRACIS 2019)*. DOI: 10.1109/BRACIS.2019.00011.

22. Miranda, N., R. Raminhos, P. Seabra, J. Sequeira, T. Gonçalves, and P. Quaresma. 2011. Named Entity Recognition using Machine Learning techniques. In *Proceedings of EPIA-11, 15th Portuguese Conference on Artificial Intelligence*. Available in <https://portulanclarin.net/static/docs/uevora-tagger/miranda2011epia.pdf>.

23. Moreira, F., and R. Vieira. 2019. Aplicação de Reconhecimento de Entidades Nomeadas em investigação de Crimes Financeiros. In *Proceedings of XII Symposium in Information and Human Language Technology and Collocates Events (STIL 2019)*. Available in <http://comissoes.sbc.org.br/ce-pln/stil2019/proceedings.html>.

24. Mota, C., A. Nascimento, P. Miranda, R. Mello, I. Maldonado, and J.C. Filho. 2021. Reconhecimento de entidades nomeadas em documentos jurídicos em português utilizando redes neurais. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2021)*. DOI: 10.5753/eniac.2021.18247

25. Neto, J.R.C.S.A.V.S., and T.d.P. Faleiros. 2021. Deep Active-Self Learning Applied to Named Entity Recognition. In *Proceedings of 10th Brazilian Conference on Intelligent Systems (BRACIS 2021)*. DOI: 10.1007/978-3-030-91699-2_28

26. Oliveira, E., G. Dias, J. Lima, and J.P.C. Pirovani. 2021. Using Named Entities for Recognizing Family Relationships. In *Anais do IX Symposium on Knowledge Discovery, Mining and Learning*. DOI: 10.5753/kdmile.2021.17457.

27. Pinheiro, B., et al.. 2021. A Comparative Analysis of Machine Learning Named Entity Recognition Tools for the Brazilian and European Portuguese Language Variants. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2021)*. DOI: 10.5753/eniac.2021.18257

28. Pires, A., J. Devezas, and S. Nunes. 2017. Benchmarking Named Entity Recognition Tools for Portuguese. *In Proceedings of Ninth INForum: Simpósio de Informática.* Available in <https://api.semanticscholar.org/CorpusID:51991813>.

29. Pirovani, J., and E. Oliveira. 2018. Portuguese Named Entity Recognition using Conditional Random Fields and Local Grammars. In *Proceedings of Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Availabe in <https://aclanthology.org/L18-1705>.

30. Pirovani, J.P.C., and E. Oliveira. 2021. Studying the adaptation of Portuguese NER for different textual genres. *The Journal of Supercomputing, v. 77, n. 11, p. 13532-13548*. DOI: 10.1007/s11227-021-03801-9.

31. Pirovani, J.P.C., E. Oliveira. 2018. CRF+LG: A Hybrid Approach for the Portugue-

se Named Entity Recognition. In *Proceedings of 17th International Conference on Intelligent Systems Design and Applications (ISDA 2017).* DOI: 10.1007/978-3-319-76348-4_11.

32. Pirovani, J.P.C., J. Alves, M.A. Spalenza, W. Silva, C.S. Colombo, and E. Oliveira. 2019. Adapting NER (CRF+LG) for Many Textual Genres. *In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019).* Available in <https://ceur-ws.org/Vol-2421>.
33. Quinta de Castro, P.V., N.F.F. Silva, and A.S. Soares. 2018. Portuguese Named Entity Recognition Using LSTM-CRF. In *Computational Processing of the Portuguese Language (PROPOR 2018).* DOI: 10.1007/978-3-319-99722-3_9.
34. Reyes, D.D.L., D. Trajano, I.H. Manssour, R. Vieira, and R.H. Bordini. 2021. Entity Relation Extraction from News Articles in Portuguese for Competitive Intelligence Based on BERT. *In Proceedings of 10th Brazilian Conference on Intelligent Systems (BRACIS 2021).* DOI: 10.1007/978-3-030-91699-2_31
35. Rodríguez, M.M.M.S., and B.L.D. Bezerra. 2020. Processamento de Linguagem Natural para Reconhecimento de Entidades Nomeadas em Textos Jurídicos de Atos Administrativos (Portarias). *Revista de Engenharia e Pesquisa Aplicada. 5, 1, 67-77.* DOI: 10.25286/repa.v5i1.1204.
36. Sampaio, V.A., M.J.C. França, P.B.L. Silva, G.A.L. Campos, and L.D. Hissa. 2019. A Brief Survey of Deep Learning based methods against OpenNLP NameFinder for Named Entity Recognition on Portuguese Literary Texts. In *Proceedings of XII Symposium in Information and Human Language Technology and Collocates Events (STIL 2019).* Available in <http://comissoes.sbc.org.br/ce-pln/stil2019/proceedings.html>.
37. Santos, C.N., V. Guimaraes. 2015. Boosting Named Entity Recognition with Neural Character Embeddings. *arXiv preprint.* DOI: 10.48550/arXiv.1505.05008.
38. Santos, J., B. Consoli, and R. Vieira. 2020.Word Embedding Evaluation in Downstream Tasks and Semantic Analogies. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020).* Available in <https://aclanthology.org/2020.lrec-1.594>.
39. Santos, J., B. Consoli, C. dos Santos, J. Terra, S. Collonini, and R. Vieira. 2019. Assessing the Impact of Contextual Embeddings for Portuguese Named Entity Recognition. In *Proceedings of 8th Brazilian Conference on Intelligent Systems (BRACIS 2019).* DOI: 10.1109/BRACIS.2019.00083.
40. Santos, J., H.D.P. dos Santos, F. Tabalipa, and R. Vieira. 2021. De-Identification of Clinical Notes Using Contextualized Language Models and a Token Classifier. In *Proceedings of 10th Brazilian Conference on Intelligent Systems (BRACIS 2021).* DOI: 10.1007/978-3-030-91699-2_3.
41. Santos, J., J. Terra, B. Consoli, and R. Vieira. 2019. Multidomain Contextual Embeddings for Named Entity Recognition. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019).* Available in <https://ceur-ws.org/Vol-2421>.
42. Silva, D.F., A.M. Silva, B.M. Lopes, K.M. Johansson, F.M. Assi, J.T.C. Jesus, R.N. Mazo, D. Lucrédio, H.M. Caseli, and L. Real. 2021. Named Entity Recognition for Brazilian Portuguese Product Titles. In *Proceedings of 10th Brazilian Conference on Intelligent Systems (BRACIS 2021).* DOI: 10.1007/978-3-030-91699-2_36.
43. Silva, R.A., L. Silva, M.L. Dutra, and G.M. Araujo. 2020. A New Entity Extraction Model Based on Journalistic Brazilian Portuguese Language to Enhance Named Entity Recognition. In *Proceedings of International Conference on Data and Information in Online.* DOI: 10.1007/978-3-030-50072-6_5.
44. Souza, J.V.A., E.T.R. Schneider, J.O. Cezar, L.E.S. Oliveira, Y.B. Gumiel, E.C. Paraiso, D. Teodoro, and C.M.C.M. Barra. 2020. A Multilabel Approach to Portuguese Clinical Named Entity Recognition. *Journal of Health Informatics, Número Especial SBIS - Dezembro: 366-72.* Available in <https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/840>.
45. Teteo, L., P. Moura, E. Soares, and Carlos Campos. 2019. Um Framework de Ex-

tração e Etiquetamento de Informações de Trânsito. In *Anais do XVIII Workshop em Desempenho de Sistemas Computacionais e de Comunicação*. DOI: 10.5753/wperformance.2019.6472.

Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 187-198 recibido 16-12-2022 revisado 01-02-2023 aceptado 12-02-2023

Violencia Identificada en el Lenguaje (VIL). Creación de recurso para mensajes violentos

Violence Identified in Language (VIL). Creation of a resource for the detection of violent messages

Beatriz Botella, Robiert Sepúlveda-Torres, Patricio Martínez Barco, Estela Saquete
Department of Software and Computing Systems, University of Alicante, Spain
{beatriz.botella, rsepulveda, patricio, stela}@dlsi.ua.es

Resumen: La sociedad avanza cargada de conocimientos nuevos y muy accesibles, que se publican en el mundo virtual. Es una realidad que las Tecnologías de la Información y la Comunicación (TIC) han traído muchos beneficios a nuestras vidas pero también vemos como año tras año aumenta el uso de violencia en plataformas digitales. Nuestro trabajo se enfoca en la creación de recursos que permitan la detección de mensajes violentos en la red social Twitter. Se parte de la creación de una guía de anotación de grano fino para anotar un corpus de mensajes violentos (VIL) con el fin de utilizar herramientas de aprendizaje automático que nos ayuden a detectar automáticamente el problema. Con este corpus se entrenan dos modelos de lenguaje (BETO y RoBERTa_base) con los que se alcanza un valor en la métrica F_1m de 97.03 % y 96.51 % clasificando si un tuit es o no violento.
Palabras clave: Procesamiento Lenguaje Natural, Guía Anotación, Anotación Corpus, Detección Mensajes Violentos.

Abstract: Society is moving forward full of new and very accessible knowledge, which is published in the virtual world. It is a reality that ICTs have brought many benefits to our lives but we also see how year after year the use of violence on digital platforms increases. Our work focuses on the detection of violent messages in the social network Twitter. Starting from the creation of a fine-grained annotation guide to obtain a corpus of violent messages (VIL) in order to use Machine Learning tools that help us to automatically detect the problem Two language models are trained with this corpus (BETO and RoBERTa_base) with which a value of 97.03 % and 96.51 % is reached in the F_1m metric, classifying whether or not a tweet is violent.
Keywords: Natural Language Processing, Annotation Guideline, Dataset Annotation, Detection of Violent Messages.

1 Introducción

Internet se ha convertido en parte imprescindible de nuestras vidas, siendo utilizado prácticamente en todas las actividades cotidianas de la sociedad. Actualmente es posible tener contacto con cualquier persona del mundo a través de un dispositivo electrónico de manera inmediata. La sociedad avanza cargada de conocimientos nuevos y muy accesibles que se publican en el mundo virtual. Las relaciones personales también se han visto afectadas, no solo en el ámbito privado, sino también en el laboral.

Según WeAreSocial y Hootsuite (2022), casi 44 millones de personas en España son usuarias de Internet pasando más de 6 horas al día en la Red y alrededor de 41 millones de españoles son usuarios de redes sociales. Es una realidad que las TIC han traído muchos beneficios a nuestras vidas, pero también, gracias a la posibilidad de ser un usuario anónimo y la ausencia de observar cara a cara el daño que pueden generar nuestras palabras, se crean problemas aún por solucionar (Flores y Casal, 2008). En especial, muchos los investigadores denominan a este tipo de acción violenta como discurso del odio, una conducta ofensiva a través del lenguaje hacia personas o colectivos y cuya detección está siendo un problema para los investigadores, ya que, cabe la posibilidad de que la violencia no esté empleada de una forma explicita en un discurso, si no, ser una única pala-

ISSN 1135-5948 DOI 10.26342/2023-70-15

bra o incluso mediante una forma implícita con el uso de emoticonos (Alonso y Vázquez, 2017), o usando el humor, la ironía, el sarcasmo (Frenda, Patti, y Rosso, 2022; Frenda et al., 2022) o esteriotipos (Sánchez-Junquera et al., 2021).

Dada la cantidad de usuarios presentes en las redes sociales se hace imposible un control manual de los comentarios que se registran y su intención, creando una impunidad a las personas que utilizan estas redes con el fin de hacer daño. La identificación de mensajes violentos y controlar el discurso del odio en Internet se ha abordado desde diferentes puntos de vista, siendo imprescindible la utilización de Procesamiento del Lenguaje Natural (PLN) para desarrollar sistemas computacionales que ayuden a interpretar y procesar el lenguaje humano de forma rápida y efectiva.

Una barrera que encontramos nada más empezar el estudio es la recopilación de mensajes en las redes sociales, ya que como apunta Bruns (2019), la restricción al acceso de datos de las redes sociales dificulta el análisis de cuestiones de gran importancia como el lenguaje abusivo, el acoso, el discurso de odio o las campañas de desinformación.

Es por ello por lo que en la presente investigación se usará la red social Twitter donde cómo define Ott (2017): "El discurso de Twitter es irrespetuoso porque su registro es informal, y porque despersonaliza las interacciones sociales". Esta investigación persigue el objetivo de aportar soluciones a los problemas existentes en la detección de mensajes violentos en redes sociales de una forma rápida, automática y eficaz. La principal contribución del trabajo es un esquema de anotación de grano fino que vaya más allá de marcar un mensaje como violento o no, sino que permite una anotación semántica mucho más compleja del mismo, permitiendo un nivel de detalle mucho más exhaustivo que la simple detección binaria.

El artículo está estructurado de la siguiente manera: Sección 2, se muestran los principales trabajos realizados en la materia y las formas de detección; en la sección 3 describimos cual han sido los pasos de anotación para etiquetar mensajes violentos. La sección 4 explica el proceso de compilación, anotación, así como una prueba piloto para verificar la anotación de nuestra guía. La validación de nuestro corpus y experimentos se encuentran en la sección 5; la sección 6, muestra los resultados de la experimentación realizada y por último en la sección 7, conclusiones y trabajo futuro.

2 Estado de la cuestión

Son muchos los estudios que se han llevado a cabo sobre el análisis de mensajes violentos en redes sociales y medios de comunicación. En concreto, se puede encontrar mucha investigación centrada en descubrir las características del comportamiento humano que promueven la emisión de dichos mensajes, así como los que se centran en descubrir las características de los propios mensajes a través de técnicas de PLN. Si bien nuestro estudio está enfocado a este último grupo, revisaremos algunos de los trabajos más importantes para ambos casos.

2.1 Estudio del lenguaje y comportamiento

Hay una gran cantidad de estudios acerca del comportamiento humano ante los mensajes violentos y el lenguaje empleado. Como dijo McMenamin (2017), "el discurso del odio se estudia según cómo se define, cómo se interpreta, y cuáles son las mejores prácticas para enfrentarlo". Es por ello que encontramos trabajos como Salado (2022), que basaron su investigación en un análisis sintáctico del lenguaje, y descubrieron que hay distintos elementos lingüísticos a tener en cuenta que están presentes en las formas del habla violentas como, la categoría lingüística, el léxico empleado o cómo están colocadas las palabras. Del Arco et al. (2022) realiza un estudio de los fenómenos lingüísticos implícitos y explícitos del lenguaje ofensivo. Otros como Gitari (2015) se centraron en algo tan específico como la creación de un listado de verbos que pueden ser indicadores de mensajes violentos. Por otra parte existen trabajos que se centran en los roles presentes en estos actos como por ejemplo, Nielsen (2002) que a través de unas entrevistas y estudio de los participantes, observó las consecuencias para la víctima, su daño y la posibilidad de delito en los mensajes.

2.2 PLN aplicado a la detección de mensajes violentos

La aplicación del PLN es fundamental en este tipo de investigaciones dado el gran volumen de datos existentes, lo que facilita un gran avance en la investigación de la detección de

este tipo de mensajes, gracias a las siguientes técnicas:

- **Clasificadores basados en palabras claves**

  Una parte de las investigaciones en este campo se han centrado en la elaboración de lista de insultos que ayuden a una detección automática. En este sentido, se han desarrollado lexicones y diccionarios con el fin de observar si la presencia de estos términos determina la violencia en el mensaje (Sood, Churchill, y Antin, 2012).

  Aunque este tipo de listas han ayudado a la detección, se ha quedado escaso a la hora de ser la única herramienta para determinar la violencia. El lenguaje violento evoluciona constantemente,varía según el lugar donde ocurra y es posible que existan términos que en algunas zonas geográficas sean insultos y en otras no (Nobata et al., 2016).

- **Aprendizaje automático**

  La mayoría de los trabajos relacionados con la detección de mensajes violentos abordan esta problemática con la utilización de algoritmos clásicos de aprendizaje automático (ML). Trabajos como Xu et al. (2012) y Dadvar et al. (2013) han utilizado máquinas de soporte vectorial (SVM) en sus investigaciones obteniendo resultados satisfactorios, demostrando ser muy eficaz con muestras de entrenamiento de grandes dimensiones. SVM no es el único algoritmo clásico utilizado en las investigaciones de este campo, trabajos como Arcila-Calderon et al. (2021), utilizaron otros algoritmos, mostrando en sus resultados que el que ofrecía mejor rendimiento es la regresión logística, seguida de Naive Bayes y las SVM.

  La mayoría de los clasificadores basados en ML utilizan representaciones de textos tradicionales como bolsa de palabras (BOW), n-grams, frecuencia de términos (TF), entre otras. En Burnap y Williams (2014) se utilizan todas las técnicas citadas anteriormente. Esta investigación compara los resultados obtenidos de forma individual por los clasificadores con la utilización de un conjunto de clasificadores (ensemble) que los integra a todos, demostrando mayor precisión en este el último. El análisis de sentimientos es otra de las herramientas más utilizadas en este campo. Con ella podemos extraer la polaridad del mensaje y utilizar este indicador junto a otras tareas para determinar con mayor exactitud si estamos ante un mensaje violento o no (Martins et al., 2018).

  El desarrollo de corpus, tienen un papel importante en las investigaciones del lenguaje ofensivo cuando se aplican técnicas de ML. En los últimos años hemos observado un gran volumen de trabajo por parte de investigadores en PLN para generar estos recursos (Wiegand et al., 2018; Qian et al., 2019; Olteanu et al., 2018; Fortuna y Nunes, 2018; Poletto et al., 2021; Rosenthal et al., 2020). Estos autores crearon recursos en inglés, siendo SOLID (Rosenthal et al., 2020) el recurso que contiene más de nueve millones de tuits en inglés etiquetados de forma semisupervisada.

  Por otra parte, HurtLex (Bassignana, Basile, y Patti, 2018) es un léxico multilingüe de palabras de odio que abarcan varios idiomas y Hatebase3[1] es un repositorio colaborativo de discurso de odio también multilingüe. El principal inconveniente de estos recursos es su escasez de términos en español, y los que están presentes se han recopilado utilizando una traducción semiautomática de otro idioma, dejando de lado la importancia de los factores culturales y lingüísticos de cada país. Sin embargo, a pesar de que el español es una de las lenguas más habladas del mundo, encontramos escasez de recursos en este idioma para llevar a cabo la tarea de detección.

  Existen recursos en español de palabras ofensivas como Plaza-Del-Arco et al. (2020) para términos misóginos y xenófobo; y Share (Plaza-del Arco et al., 2022) que los etiquetan como ofensivo y no. Tras el estudio realizado sobre la literatura se considera necesaria la elaboración de otro corpus donde recoger más características presentes en los mensajes violentos, que puedan ayudar en la explicabilidad y el detalle de la detección.

- **Aprendizaje profundo**

[1] `https://hatebase.org/`

Dentro de la Inteligencia Artificial existen otras técnicas más complejas que también se han utilizado en esta tarea. Nos referimos al aprendizaje profundo (DL), como es el caso de la investigación de Arcila-Calderón et al. (2021) que tras utilizar las herramientas de aprendizaje automático y redes neuronales, estas últimas mejoraron las métricas de evaluación frente a los modelos generados con algoritmos de ML tradicionales. Con el mismo fin Badjatiya et al. (2017) usa modelos de DL para entrenar diferentes incrustaciones de palabras validando que, utilizar estas representaciones, obtiene mejor resultados que representaciones tradicionales como frecuencia de término – frecuencia inversa de documento (TF-IDF) o BoW.

Los modelos basados en arquitectura *transformer* como es el caso de BERT, RoBERTa y ALBERT, ostentan los mejores resultados del estado del arte en la detección de mensajes violentos en tareas reconocidas como OffensEval o HatEval (Sarkar et al., 2021). En Sarkar et al. (2021) se realiza un ajuste fino (*fine-tuning*) a BERT utilizando SOLID, el mayor corpus de identificación de lenguaje ofensivo en inglés, mejorando los resultados obtenidos con BERT en las tareas mencionadas anteriormente.

En Song et al. (2021) se utiliza un conjunto de clasificadores (*ensemble*) basados en RoBERTa y BERT que obtiene los mejores resultados en la tarea compartida "SemEval-2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense"[2] que incluye una subtarea de detección de mensajes ofensivos. Este trabajo consiste en hacer un ajuste fino de estos modelos para crear un clasificador y agruparlos en un conjunto de clasificadores basados en *stacking*.

Una vez estudiada la literatura al respecto de la tarea y la importancia de la aplicación de PLN y técnicas de ML y DL, y debido a que estas técnicas se nutren de datos de entrenamiento, se concluye la necesidad de crear un recurso en español que pueda ser utilizado en la detección efectiva de mensajes violentos, con un nivel de detalle que va más allá de la simple detección binaria, marcando características, que detallamos en nuestra guía de anotación, como el grado de violencia, el rol o el tipo de violencia, puesto que consideramos que si detección en los mensajes mensajes que puedan ayudar a la futura explicabilidad en las decisiones tomadas.

3 Guía de anotación

Persiguiendo el objetivo de crear un recurso que ayude a la detección de los mensajes violentos, decidimos generar una guía de anotación de grano fino para los mensajes, con un cierto grado de complejidad semántica, donde no solamente se marca si un mensaje es violento o no, si no también determinados elementos importantes al respecto del contenido del mensaje.

3.1 Violento vs NoViolento

Para empezar nuestra anotación elegimos si el mensaje es violento o no, definiendo que entendemos por mensaje violento:

- **Violento**: Cuando el contenido del mensaje contiene la acción de hacer daño o emplea violencia en sus palabras, ejemplos: "Deberían estar en la cárcel", "eres idiota","no soporto la mierda de este puto país".
- **NoViolento**: El mensaje no contiene acción de hacer daño, aunque puede existir "palabras violentas", como por ejemplo: "Que idiota soy, por casi me lo creo", "Hoy es un gran día", "me he levantado insoportable hoy".

  Si el contenido del mensaje es *NoViolento*, solo anotaremos si contiene insultos o palabras negativas. Las demás categorías no se anotarán.

3.2 Contenido del mensaje

En este apartado nos centraremos en analizar el contenido del mensaje más en detalle para determinar tres elementos fundamentales en el mismo: i) el grado de violencia, ii) el rol del mensaje y iii) el tipo de violencia. Dentro del contenido del mensaje observamos por ejemplo si los mensajes contienen insultos o expresiones negativas por si puede ser un indicador de violencia, ejemplo: "matón de patio", "me cago en tus muertos", así como la identificación de la estructura de un

[2] http://bit.ly/3J8uHOX

mensaje violento, que no contenga un insulto o palabra ofensiva directa, ejemplo: "Te vas a enterar", "Espero encontrarte a solas en la calle". Para todo ello se seleccionan las palabras que contienen violencia.

3.2.1 Grado de Violencia

Según el contenido del mensaje podemos catalogarlo en 2 niveles de violencia:

- **Moderado**: Se recogen aquí todos los mensajes que lleven contenido violento pero no atenten contra la vida o integridad física de las personas. Groserías, desaprobación con personas o cosas, ridiculizar, insultos por ideología o política. Ejemplos: "Ese es Gilipollas", "Que asco de hotel", "Maria la tetona", "me cago en tu puta madre".
- **Grave**: En este nivel los mensajes atentan contra la vida o integridad de las personas. Amenazas, agresión física, desear el mal. Encontramos verbos como ojalá o deseo con acciones negativas y acusar a personas de delitos graves, como asesino, violador, proxeneta. Ejemplos: "Ojalá te murieses", "Ten cuidado a ver si te pasa algo", "te voy a dar una ostia cuando te vea", "eres una asesina".

3.2.2 Rol del mensaje

Definir de que forma actúan los usuarios en mensajes violentos.

- **Rol 1 - Incitador**: Su mensaje incita a los demás lectores a que escriban mensajes violentos o propicia el odio en la red. Ejemplos: "Deberíamos decirle los españoles lo idiota que es esta tía","Tendríamos que ir todos a tu casa a darte tu merecido".
- **Rol 2 - Ejecutor**: Mensaje de un usuario individual con acción directa de violencia. Ejemplos: "Deberías morirte","Eres retrasado".
- **Rol 3 - Pasivo**: Emplea violencia sin estar dirigida a nadie en concreto. Ejemplos: "La política de este país es una mierda", "Siempre nos cuentan lo mismo, se creen que somos idiotas".
- **Rol 4 - Informativo**: Es mero transmisor de la violencia pero no participa en ella. Ejemplo: "No soporto la violencia".

3.2.3 Tipo de violencia

Es importante definir que tipo de violencia se está empleando en el mensaje. Según el Ministerio del Interior del Gobierno de España, en su informe emitido en 2021 [3] los delitos de odio cometido en internet y redes sociales son: racismo/xenofobia en primer lugar, seguido de orientación sexual e identificación de género, ideología, discriminación por razón de sexo, creencias o prácticas religiosas, discriminación generacional, delitos de odio contra personas con discapacidad, descriminación por razón de enfermedad, antigitanismo, antisemitismo y aporofobia. Dado el gran volumen existente de datos, se decidió etiquetar los mensajes en 5 tipos de violencias. Añadiendo la violencia machista, no presente en el informe anterior, por su alerta en la sociedad y con el fin de estudiar este problema en futuras investigaciones.

- **Machista**: el contenido del mensaje conlleva una actitud despectiva contra las mujeres. Ejemplos: "Lo has conseguido por ser una guarra y ponerte de rodillas", "Las mujeres no sabéis hacer otra cosa", "Estás con él por tu dinero".
- **Homófoba**: mensajes violentos hacia la homosexualidad o las personas homosexuales. Ejemplos: "Los gays están enfermos y tienen que curarse", "Bolleras de mierda".
- **Ideología_religiosa**: ataques contra las ideologías religiosas. Ejemplos: "Me río de tu dios Alá", "Los católicos son asquerosos".
- **Política**: violencia hacia cualquier ideología política o persona/s política que los representa. Ridiculizar nombres políticos. Ejemplos: "Peperos de mierda", "Los de podemos son asquerosos".
- **Xenofobia/Racismo**: rechazo a cualquier persona por el mero hecho de no compartir la misma nacionalidad o actitud o ideología donde una raza o grupo étnico se considera superior a otra. Ejemplos: "Moro de mierda", "Panchitos", "No podemos ser iguales que los negros, ellos son una escala inferior".
- **Otro**: otro tipo de violencia que no corresponda a los anteriores. Ejemplos:

[3] https://bit.ly/3FXZyxt

"Madrilistas de mierda"(Violencia Deportiva), "Ojala maten al perro"(Violencia animal), "Asco de los toreros"(Profesión). Este ultimo apartado se les pedía a los que en el apartado de Observaciones escribieran qué tipo de violencia era la que habían catalogado como Otro.

4 Corpus VIL

Una vez definida la guía de anotación, se procede a la construcción de un corpus basándonos en esa guía. Las fases seguidas en la construcción del recursos serán presentados en la siguientes subsecciones.

4.1 Proceso de compilación

Para poder tener un corpus de mensajes violentos etiquetados, primero se pensó en que red social era la más apropiada para descargar mensajes de los usuarios. Basado en la justificación mostrada en la sección 1, se escogió Twitter debido a la manera informal de expresarse que tienen los usuarios en esta red social. Además, Twitter permite descargar tuits con gran facilidad. A continuación, se pensaron 3 escenarios en los que se presencian opiniones que afectan a la sociedad actual, con lo que obtendríamos tuits reales con alta probabilidad de violencia. Los tuits seleccionados están relacionados con 3 acontecimientos ocurridos en España:

- Entrevista de La Sexta realizada a la política Cayetana Álvarez de Toledo.
- La campaña de Irene Montero, ministra de igualdad sobre la campaña "Sola y borracha quiero llegar a casa".
- Isabel Díaz Ayuso, en la manifestación que ocurrió el 13 de noviembre 2022 por la Sanidad en Madrid.

Esta elección se hizo en base a la actualidad en la sociedad española y el odio existente a los políticos de nuestro país. Los tuits se descargaron mediante la herramienta "Social Analytics" (Fernández et al., 2017), en total unos 12500 tuits. Con los tuits descargados se realizó un proceso de limpieza para eliminar tuits repetidos y retuits, generándose un total de 90 paquetes de tuits, donde cada paquete contiene 100 tuits.

4.2 Prueba piloto anotación

Para asegurarnos de que nuestra guía de anotación era correcta, contamos con la ayuda de 6 anotadores entrenados por un experto en la guía de anotación, cada uno de ellos anotó la misma cantidad de tuits (40 tuits). En el proceso de anotación de esta prueba piloto se observó que la guía era demasiado compleja, derivando en confusión para los anotadores, con anotaciones incorrectas. Con esta prueba se decidió hacer una serie de modificaciones a la guía de anotación, simplificando los pasos a seguir por los anotadores. En la guía inicial contábamos con 3 niveles de violencia, (leve, moderado y grave), pero la línea de decisión entre las opciones leve y moderado era muy difícil de definir por los anotadores, debido a la subjetividad de la violencia dependiendo de la persona que etiquetaba. Por ese motivo se modificó dejando solos dos niveles de violencia (presentes en la guía actual) y se añadieron más ejemplos que permitiesen reducir todas las dudas que suscitaban. También se añadieron ejemplos en el resto de opciones para asegurar un etiquetado correcto.

4.3 Anotación del corpus VIL

Como resultado del proceso de anotación después de las dos primeras fases explicadas anteriormente se procedió a la anotación masiva de los tuits recopilados. Con esta anotación se obtiene el corpus Violencia Identificada en el Lenguaje (VIL), el cual contiene un total de 2874 tuits anotados con 1491 *Violento* y 1383 *NoViolento*. La cantidad de tuits *Violento* y *NoViolento* que contiene el corpus es similar, evitando así que los modelos que lo utilicen se vean afectados por un posible desbalance entre las clases que contiene.

Con el fin de evaluar el rendimiento de futuros modelos entrenados utilizando este corpus, se realiza un particionamiento del mismo en (entrenamiento, validación y prueba). La partición de prueba fue extraída aleatoriamente utilizando el 20 % de los tuits anotados, de los tuits restantes el 20 % se reserva para evaluar experimentos (partición de validación) y el resto para la partición de entrenamiento. La tabla 1 muestra la distribución de etiquetas por cada partición. El conjunto de datos VIL esta disponible para su descarga y utilización en `http://bit.ly/3ZVwUnL`.

Más concretamente, dato este conjunto de datos seleccionados sobre los tres eventos mencionados actualmente, la distribución por tipo de violencia es la siguiente: 13 mensajes machistas, 5 mensajes homófobos, mensajes de 0 ideología religiosa, 174 mensajes políti-

	Violento	NoViolento	Total
Entrenamiento	957	882	1839
Validación	236	224	460
Prueba	298	277	575
Total	1491	1383	2874

Tabla 1: Distribución de etiquetas en las particiones de entrenamiento, validación y prueba.

cos, 2 de Xenofobia y racismo y 1381 mensajes que no corresponden a ninguno de los anteriores. Este grupo sería el más amplio dada la complejidad de la clasificación, siendo para estos eventos recopilados concretamente los mensajes racistas los más escasos y ninguno de ideología religiosa. El tipo de violencia va muy ligado al tipo de situación o evento que se esté analizando, y debido a este balanceo en trabajos futuros será necesario la ampliación de los tipos de violencia más escasos.

Para la anotación de este corpus se ha utilizado la herramienta de anotación Brat (Stenetorp et al., 2012). Esta permite la anotación de mensajes de una forma intuitiva mostrando una ventana para seleccionar la anotación deseada. Previamente se configura los campos específicos de la anotación así como la jerarquía en las anotaciones. Los insultos etiquetados mediante Brat se encuentran disponibles en el siguiente repositorio GitHub: `https://bit.ly/3ZVwUnL`. La figura 1 muestra algunos ejemplos de tuits anotados en el conjunto de datos VIL.

5 Validación del esquema de anotación y del corpus VIL

Esta sección presenta una validación realizada al esquema de anotación para corroborar que la guía de anotación es clara y precisa, derivando en una anotación del corpus acorde con la definición de la misma. Para cumplir este objetivo se realiza una validación de acuerdo entre anotadores. Por otra parte, se valida la pertinencia del corpus VIL para crear un sistema de detección de mensajes violentos en Twitter.

5.1 Validación entre anotadores

Con el objetivo de medir la calidad de la tarea de anotación se realizó un acuerdo entre dos anotadores. Los dos anotadores elegidos son criminólogos y está selección se hizo por su conocimiento en el comportamiento violento entre humanos. Estos anotaron independientemente 200 tuits entre *Violento* y *NoViolento*, calculando un índice de acuerdo en la anotación. Se utilizó el índice *kappa* (Cohen, 1960) para calcular el acuerdo en las anotaciones (índice común en procesos de validación de anotaciones entre dos anotadores). Se obtuvo un *kappa* de 0,868, lo que representa un valor alto de acuerdo entre dos anotadores, validando así el proceso de anotación.

Adicionalmente se calculó el acuerdo teniendo en cuenta el contenido del mensaje marcado, primeramente evaluando si el tuit contiene insultos, alcanzándose un *kappa* de 0.896. Por último, teniendo en cuenta la anotación del grado de violencia, el índice de acuerdo es de 0.753, sustancialmente menor que el resto de anotaciones, lo que evidencia que esta etiqueta es la más compleja de anotar.

En cualquier caso, se considera que los valores *kappa* obtenidos son suficientes para garantizar la calidad del corpus.

5.2 Experimentos

Un sistema capaz de detectar tuits violentos es de gran relevancia en el contexto actual de ataques constantes a través de redes sociales. Para probar la validez del conjunto de datos VIL, se realizaron dos experimentos que lo utilizan como base para entrenar modelos de lenguaje y evaluar el rendimiento de los mismos para predecir si un tuit es violento o no.

Para llevar a cabo estos experimentos se utilizan dos modelos de lenguaje en español (BETO y RoBERTa_base), basados en arquitectura *transformers* descritos en Canete et al. (2020) y Gutiérrez-Fandiño et al. (2021) respectivamente.

BETO está basado en el modelo de lenguaje BERT, diseñado para representar relaciones bidireccionales profundas a partir de texto sin etiquetar, utilizando mecanismos de atención (Devlin et al., 2018). Para la creación de BETO se realizaron una serie de optimizaciones similares a las llevadas a cabo para obtener el modelo RoBERTa (Liu et al., 2019). En este caso se entrena utilizando textos en español de la enciclopedia libre (Wiki-

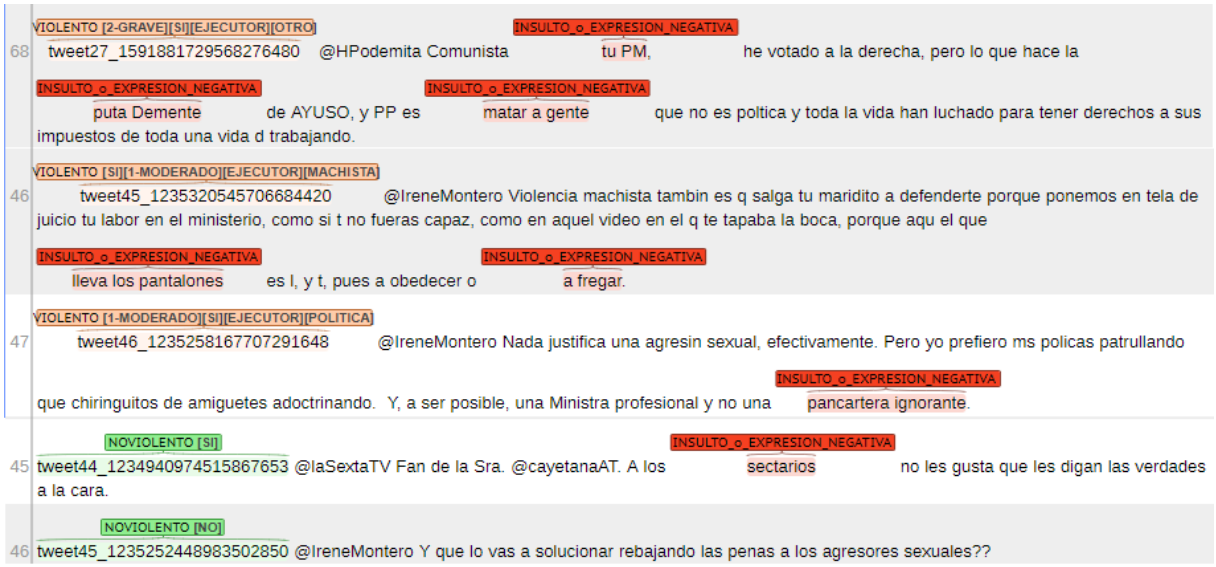


Figura 1: Ejemplos de tuits anotados en el corpus VIL.

pedia) y todas las fuentes del Proyecto OPUS (Tiedemann, 2012).

En el caso de RoBERTa_base en español se basa en el modelo de lenguaje RoBERTa. Para la obtención de este modelo se utiliza un total de 570 GB de texto limpio y recopilado por la Biblioteca Nacional de España de 2009 a 2019 (Gutiérrez-Fandiño et al., 2021).

En el contexto de estos experimentos los modelos BETO pre-entrenado[4] y RoBERTa_base[5] se utiliza en modo ajuste fino para ajustarlo a la tarea de detección de tuits violentos. En el primer experimento se entrenan los modelos de lenguajes utilizando como secuencia de entrada únicamente el texto del tuit. Por su parte, el segundo experimento concatena el texto del tuit con las frases anotadas como insultos para entrenar el modelo. Finalmente para llevar a cabo los experimentos se utilizó la biblioteca Simple Transformers[6] con la siguiente configuración de hiperparámetros en todos los experimentos: tasa de aprendizaje de 2e-5, tamaño de lote de 2 y número de iteraciones para entrenar 3.

6 *Resultados*

Los experimentos llevados a cabo en este artículo se pueden replicar descargando el código del siguiente repositorio GitHub: `http://bit.ly/3YGfVVs`.

[4] `http://bit.ly/3Feu5q6`
[5] `http://bit.ly/3FcBvu6`
[6] `https://simpletransformers.ai/`

Con la configuración inicial de hiperparámetros se realizó un ajuste fino preliminar para evaluar el comportamiento del modelo para predecir la partición de validación. La figura 2 muestra el comportamiento de las curvas de perdida, así como la métrica F_1m en cada iteración de entrenamiento. Esta figura corresponde con el ajuste fino del modelo BETO, sin embargo con el modelo RoBERTa_base el comportamiento es similar.

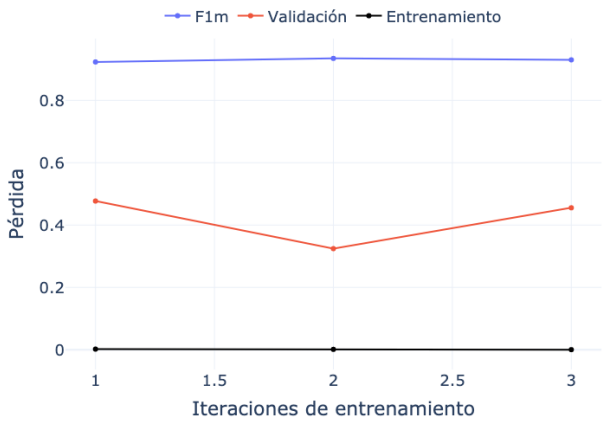


Figura 2: Curva de pérdida utilizando el conjunto de entrenamiento y validación durante el entrenamiento.

Como se puede apreciar en la figura 2, la pérdida en la curva de validación (línea roja) desciende de 0.47 a 0.32 de la primera iteración a la segunda, por el contrario en la tercera iteración esta aumenta hasta 0.45. Por su

parte la curva de la pérdida de entrenamiento (línea negra) disminuye en cada iteración. La disminución en la pérdida en el entrenamiento en todas las iteración así como el aumento para pérdida en validación (de la iteración 2 a la 3) es una evidencia contundente de que el modelo se empezó a sobreajustar. Por último la curva que representa la evolución de la métrica macro-promedio F_1 (F_1m) alcanza su máximo valor (93.47 %) en la segunda iteración, lo que se corresponde con el comportamiento de la curva de validación. Después de este análisis se decidió utilizar el modelo entrenado durante dos iteraciones para predecir la partición de prueba.

La tabla 2 muestra los resultados de las métricas puntuación F_1 por clase y el F_1m prediciendo la partición de prueba. Los valores se expresan en modo de porcentaje.

Los experimentos BETO (texto del tuit) y RoBERTa-base model (texto del tuit) obtienen un buen rendimiento en la predicción de si un tuit es o no violento. Estos solo utilizan como entrada al modelo de entrenamiento y predicción el texto del tuit. Los resultados alcanzados son similares a los obtenidos por Mathew et al. (2021) sobre el corpus HateXplain en idioma inglés para la detección de mensajes violentos, en este trabajo también se entrenan clasificadores basados en BERT. Para el caso del castellano, lo más cercano al trabajo aquí presentado sería la evaluación de SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter (Basile et al., 2019), en la subtarea B donde se clasifica la agresividad de los mensajes, obteniéndose valores para el castellano de alrededor del 70.5 %.

Por su parte, los experimentos que concatenan el texto del tuit con los insultos anotados en el tuit —BETO (texto del tuit y los insultos) y RoBERTa_base model (texto del tuit y los insultos)— mejoran en todas las métricas a los sistemas bases que solo utilizan los textos del tuit. Este hallazgo indica la pertinencia de preanotar insultos en los textos y luego clasificarlos.

Teniendo en cuenta los modelos de lenguajes utilizados, con el modelo BETO se obtienen mejores resultados en el experimento que utiliza los insultos anotados como entrada al modelo. Sin embargo, para el experimento que solo utiliza el texto como entrada al modelo RoBERTa_base, se mejora la métrica F_1m en un punto porcentual con respecto al experimento utilizando el otro modelo. En los escenarios mencionados anteriormente las diferencias son bajas, sin embargo se considera que los resultados difieren de los esperados debido a que se pensaba que el modelo RoBERTa_base obtendría los mejores resultados en ambos experimentos debido a que fue entrenado sobre un conjunto de datos mucho más extenso y utilizando optimizando el proceso de entrenamiento.

7 Conclusiones y trabajo futuro

Este trabajo se ha realizado con el fin de encontrar mejoras en la detección de los mensajes violentos en redes sociales, utilizando un esquema de anotación de grano fino para obtener un corpus de mensajes violentos con indicadores de nivel de violencia, rol, presencia de insultos y tipo de violencia. El corpus VIL ha sido utilizado para entrenar clasificadores basados en modelos de lenguaje *transformers*. Estos clasificadores obtienen resultados significativos cuando se utiliza el texto del tuit concatenado con las frases con insultos anotadas. En próximos trabajos esperamos utilizar mecanismo de Human in the Loop y Active Learning para obtener un dataset a gran escala con mayor cantidad de mensajes violentos de tipo *Grave*, debido que en esta primera versión solo se cuenta con 60 mensajes de este tipo, además, vamos a trabajar con un mayor numero de dominios para aumentar los distintos tipos de violencia.

Agradecimientos

Esta investigación ha sido financiada por MCIN/AEI/ 10.13039/501100011033 y la Unión Europea NextGenerationEU/PRTR a través de los proyectos “TRIVIAL” (PID2021-122263OB-C22) and “SocialTrust” (PDC2022-133146-C22). También cuenta con el apoyo de la Generalitat Valenciana a través del proyecto “NL4DISMIS” (CIPROM/2021/21).

Bibliografía

Alonso, L. y V. J. Vázquez. 2017. *Sobre la libertad de expresión y el discurso del odio: Textos críticos*. Athenaica ediciones universitarias.

Arcila-Calderón, C., J. J. Amores, P. Sánchez-Holgado, y D. Blanco-Herrero. 2021. Using shallow and deep learning to automatically detect hate motivated by

	F_1		
	Violento	**NoViolento**	F_1m
BETO (texto del tuit)	87.32	86.92	87.12
BETO (texto del tuit y los insultos)	**97.18**	**96.89**	**97.03**
RoBERTa-base model (texto del tuit)	88.96	87.26	88.11
RoBERTa-base model (texto del tuit y los insultos)	**96.39**	**96.64**	**96.51**

Tabla 2: Resultados de los experimentos utilizando el corpus VIL.

gender and sexual orientation on twitter in spanish. *Multimodal technologies and interaction*, 5(10):63.

Badjatiya, P., S. Gupta, M. Gupta, y V. Varma. 2017. Deep learning for hate speech detection in tweets. En *Proceedings of the 26th international conference on World Wide Web companion*, páginas 759–760.

Basile, V., C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, y M. Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. En *Proceedings of the 13th international workshop on semantic evaluation*, páginas 54–63.

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

Bruns, A. 2019. After the 'apicalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, 22(11):1544–1566.

Burnap, P. y M. L. Williams. 2014. Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making.

Canete, J., G. Chaperon, R. Fuentes, y J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. *PML4DC at ICLR*, 2020.

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

Dadvar, M., D. Trieschnigg, R. Ordelman, y F. d. Jong. 2013. Improving cyberbullying detection with user context. En *European Conference on Information Retrieval*, páginas 693–696. Springer.

del Arco, F. M. P., M. D. Molina-González, L. A. Ureña-López, y M.-T. Martín-Valdivia. 2022. Integrating implicit and explicit linguistic phenomena via multi-task learning for offensive language detection. *Knowledge-Based Systems*, 258:109965.

Devlin, J., M.-W. Chang, K. Lee, y K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fernández, J., F. Llopis, P. Martínez-Barco, Y. Gutiérrez, y Á. Díez. 2017. Analizando opiniones en las redes sociales. *Procesamiento del Lenguaje Natural*, 58:141–148.

Flores, J. y M. Casal. 2008. *Ciberbullying. Guía rápida para la prevención del acoso por medio de las nuevas tecnologías*.

Fortuna, P. y S. Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Frenda, S., A. T. Cignarella, V. Basile, C. Bosco, V. Patti, y P. Rosso. 2022. The unbearable hurtfulness of sarcasm. *Expert Systems with Applications*, 193:116398.

Frenda, S., V. Patti, y P. Rosso. 2022. Killing me softly: Creative and cognitive aspects of implicitness in abusive language online. *Natural Language Engineering*, páginas 1–22.

Gitari, N. D., Z. Zuping, H. Damien, y J. Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, A. Gonzalez-Agirre, C. Armentano-Oller, C. Rodriguez-Penagos, y M. Villegas.

2021. Spanish language models. *arXiv preprint arXiv:2107.07253*.

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

Martins, R., M. Gomes, J. J. Almeida, P. Novais, y P. Henriques. 2018. Hate speech classification in social media using emotional analysis. *Proceedings - 2018 Brazilian Conference on Intelligent Systems, BRACIS 2018*, páginas 61–66, 12.

Mathew, B., P. Saha, S. M. Yimam, C. Biemann, P. Goyal, y A. Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. En *Proceedings of the AAAI Conference on Artificial Intelligence*, volumen 35, páginas 14867–14875.

McMenamin, G. R. 2017. *Introducción a la lingüística forense: un libro de curso*. Press at California State University, Fresno.

Nielsen, L. B. 2002. Subtle, pervasive, harmful: Racist and sexist remarks in public as hate speech. *Journal of Social Issues*, 58:265–280, 1.

Nobata, C., J. Tetreault, A. Thomas, Y. Mehdad, y Y. Chang. 2016. Abusive language detection in online user content. En *Proceedings of the 25th international conference on world wide web*, páginas 145–153.

Olteanu, A., C. Castillo, J. Boy, y K. Varshney. 2018. The effect of extremist violence on hateful speech online. En *Proceedings of the international AAAI conference on web and social media*, volumen 12.

Ott, B. L. 2017. The age of twitter: Donald j. trump and the politics of debasement. *Critical studies in media communication*, 34(1):59–68.

Plaza-Del-Arco, F.-M., M. D. Molina-González, L. A. Ureña-López, y M. T. Martín-Valdivia. 2020. Detecting misogyny and xenophobia in spanish tweets using language technologies. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–19.

Plaza-del Arco, F. M., A. B. P. Portillo, P. L. Úbeda, B. Gil, y M.-T. Martín-Valdivia. 2022. Share: A lexicon of harmful expressions by spanish speakers. En *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, páginas 1307–1316.

Poletto, F., V. Basile, M. Sanguinetti, C. Bosco, y V. Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Qian, J., M. ElSherief, E. Belding, y W. Y. Wang. 2019. Learning to decipher hate symbols. *arXiv preprint arXiv:1904.02418*.

Rosenthal, S., P. Atanasova, G. Karadzhov, M. Zampieri, y P. Nakov. 2020. A large-scale semi-supervised dataset for offensive language identification. *arXiv preprint arXiv:2004.14454*.

Salado, M. R. 2022. Análisis lingüístico del discurso de odio en redes sociales. *VISUAL REVIEW. International Visual Culture Review/Revista Internacional de Cultura Visual*, 9(Monográfico):1–11.

Sánchez-Junquera, J., P. Rosso, M. Montes, B. Chulvi, y others. 2021. Masking and bert-based models for stereotype identication. *Procesamiento del Lenguaje Natural*, 67:83–94.

Sarkar, D., M. Zampieri, T. Ranasinghe, y A. Ororbia. 2021. Fbert: A neural transformer for identifying offensive content. *arXiv preprint arXiv:2109.05074*.

Song, B., C. Pan, S. Wang, y Z. Luo. 2021. Deepblueai at semeval-2021 task 7: Detecting and rating humor and offense with stacking diverse language model-based methods. En *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, páginas 1130–1134.

Sood, S. O., E. F. Churchill, y J. Antin. 2012. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology*, 63:270–285, 2.

Stenetorp, P., S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, y J. Tsujii. 2012. Brat:

a web-based tool for nlp-assisted text annotation. En *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 102–107.

Tiedemann, J. 2012. Parallel data, tools and interfaces in OPUS. En *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, páginas 2214–2218, Istanbul, Turkey, may. European Language Resources Association (ELRA).

WeAreSocial y Hootsuite. 2022. Digital report espaÑa 2022: Nueve de cada diez espaÑoles usan las redes sociales y pasan casi dos horas al dÍa en ellas.

Wiegand, M., J. Ruppenhofer, A. Schmidt, y C. Greenberg. 2018. Inducing a lexicon of abusive words–a feature-based approach. 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 1046–1056.

Xu, J.-M., K.-S. Jun, X. Zhu, y A. Bellmore. 2012. Learning from bullying traces in social media. En *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, páginas 656–666.

Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 199-212 recibido 16-12-2022 revisado 01-02-2023 aceptado 11-02-2023

Exploring politeness control in NMT: fine-tuned vs. multi-register models in Castilian Spanish

Estudio de la cortesía en traducción automática neuronal: modelos ajustados y modelos multirregistro para el castellano

Celia Soler Uguet[1] **Nora Aranberri**[2]
[1]University of the Basque Country (UPV/EHU)
[2]HiTZ Center - Ixa, University of the Basque Country (UPV/EHU)
csoler003@ikasle.ehu.eus
nora.aranberri@ehu.eus

Abstract: Nowadays neural machine translation can generate high quality translations with regard to grammatical accuracy and fluency. Therefore, it is time to broaden research efforts to consider aspects of language that go beyond the mentioned attributes to keep pushing the limits of the technology. In this work, we focus on politeness. Specifically, we adapt and explore, for Castilian Spanish, two different domain-adaptation approaches: fine-tuning and multilingual models. Results from automatic and manual evaluations seem to indicate that the latter might be a better solution to strike a quality balance between all registers (formal, informal, and neutral). Fine-tuning a baseline system for each specific register seems to suffer from a degree of catastrophic forgetting, which leads to a worse overall performance of the engines.
Keywords: neural machine translation, politeness, fine-tuning models, multi-register models.

Resumen: En la actualidad, la traducción automática neuronal es capaz de generar traducciones de alta calidad en lo que respecta a la precisión gramatical y la fluidez. Así, es hora de ampliar los objetivos de investigación y considerar aspectos de la lengua que van más allá de los atributos mencionados para seguir superando los límites de la tecnología. En este trabajo, nos centramos en la cortesía. En concreto, adaptamos y exploramos, para el castellano, dos enfoques diferentes de adaptación al dominio: modelos ajustados y modelos multilingües. Los resultados de las evaluaciones automáticas y manuales parecen indicar que el segundo podría ser mejor para lograr un equilibrio de calidad entre todos los registros (formal, informal y neutro). El ajuste de modelos parece sufrir de olvido catastrófico, lo que conduce a un peor rendimiento general de los motores.
Palabras clave: traducción automática neuronal, cortesía, modelos ajustados, modelos multirregistro.

1 Introduction

As Vanmassenhove, Shterionov, and Gwilliam (2021) suggest, now that Neural Machine translation (NMT) systems have reportedly reached a quality that is close to that of human translations, it is time to start paying attention to aspects of language that go beyond grammatical accuracy such as discourse phenomena. In this line, one such phenomenon is the level of politeness. Deviations from what is expected in its use can give rise to misunderstandings in communication, and although this might seem like a petty problem, it can become extremely critical for certain cultures and communicative situations (Haugh, 2005).

Now, what is politeness? Let us start by defining register. Register was described by Matthiessen and Halliday (1997) as the context of a situation in a speech act, which consists of three dimensions: field, mode and tenor. The field refers to the area in which

ISSN 1135-5948 DOI 10.26342/2023-70-16

the linguistic activity is operating (specialized vs. non-specialized discourse); the mode has to do with the means in which communication is taking place (written vs. oral); and the tenor denotes the relationship between the speakers (relatives vs. workmates) (Halliday, McIntosh, and Stevens, 1964). In this scenario, politeness presents itself as one of the aspects that comprise the tenor, described by Brown (2015, 1) as "a matter of taking into account the feelings of others as to how they should be interactionally treated, including behaving in a manner that demonstrates appropriate concern for the interactors' social status and their social relationship". Therefore, we can argue that politeness is one of the many aspects that NMT needs to address in order to adequately respond to a specific register.

In this work, we explore ways to control the level of politeness in NMT for an English to Spanish system. Specifically, we focus on Castilian, the Spanish variety spoken in Spain, where, as a general rule, different personal pronouns are used to address an interlocutor depending on the intended level of politeness: *tú* tends to be the form used in situations where interlocutors are (relatively) close, while *usted* tends to be the form used to show respect and distance. We explore two domain-adaption techniques, namely, a fine-tuning approach following research by Chu and Wang (2018) and a multi-register approach following Sennrich, Haddow, and Birch (2016a), by adapting their setups to address the new language.

Results seem to indicate that a multi-register system trained in three directions (formal, informal and neutral) using a mix of in-domain and out-of-domain data achieves better average scores when taking into account the three directions according to both automatic and human evaluations. This approach seems to slightly outperform the fine-tuning approach, which seems to suffer from catastrophic forgetting.

The remainder of this paper is organised as follows: Section 2 presents the related work on addressing politeness in NMT; Section 3 describes the experimental setup of our study; Section 4 reports the results obtained; and finally, Section 5 draws the main conclusions and presents some avenues for further analysis on the topic.

2 Related work

Domain-adaptation is a fairly researched area in MT, as general purpose systems usually perform poorly and systems geared towards specific domains are in high demand (Koehn and Knowles, 2017). One of the main approaches used in this area is the fine-tuning of a baseline system (Kell, 2018). In NMT, it involves leveraging out-of-domain corpora to improve in-domain translations (Kirkpatrick et al., 2017), and it has been implemented successfully in various works (Luong and Manning, 2015; Etchegoyhen et al., 2018).

Alternatively, some recent research has proposed strategies to guide and control NMT output, for example, translation memory guided neural fuzzy repair (Bulte and Tezcan, 2019), domain control using side constraints such as tag-tokens and word features (Kobus, Crego, and Senellart, 2017), terminology constraints (Dinu et al., 2019), or constrained decoding (Post and Vilar, 2018).

However, to the best of our knowledge, register-related work, in general, and politeness, in particular, have received very little attention so far. In fact, we found that, to date, experiments have only been carried out with one main approach for the linguistic phenomenon at hand, namely, the application of a multilingual model, proposed by Sennrich, Haddow, and Birch (2016a) and later recreated by Feely, Hasler, and de Gispert (2019) to address politeness in German and Japanese, respectively.

In the following lines, we describe the two approaches, fine-tuning and multilingual models, in more detail.

2.1 Fine-tuning approach

Fine-tuning is considered model centric, or more precisely, training-objective centric according to the classification by Chu and Wang (2018). Here, an NMT system is trained on a resource-rich out-of-domain corpus until convergence, and then its parameters are fine-tuned on a resource-poor, in-domain corpus. A good number of positive results have been reported in the literature. For example, Luong and Manning (2015) adapted a baseline system to spoken language by further training an existing model based on formal texts (provided at WMT 2015) for 12 epochs using a smaller set of spoken text (provided at IWSLT 2015) in which after the first epoch, learning rates (initially set

to 1.0) are halved every two epochs. They reported an improvement in BLEU of almost four points.

If we were to adapt this approach to tackle politeness, the baseline system could be trained with generic data, while data for specific politeness levels could be used to develop as many fine-tuned models as necessary. Yet, it is worth mentioning that, apart from the high maintenance requirements (Bapna, Arivazhagan, and Firat, 2019), one of the main drawbacks of these systems is what is called catastrophic forgetting. This is a phenomenon whereby a model that has been trained on task A and then retrained on task B forgets much of what it originally learned on task A (Kell, 2018). Yet, as Kell (2018) outlines, different approaches have been proposed for tackling this problem, such as combining multi-domain and fine-tuning methods or using regularization techniques such as elastic weight consolidation.

2.2 Multi-register approach

The *multi-register* approach was first introduced by Sennrich, Haddow, and Birch (2016a) and has then been used for other tasks such as multilingual NMT (Aharoni, Johnson, and Firat, 2019). This method uses the placement of tags in the training data to help the decoder at translation time. Instead of applying changes to a model architecture from a standard NMT system, it introduces an artificial token at the beginning of the input sentence to specify the required target language. Sennrich, Haddow, and Birch (2016a) performed English>German experiments on OpenSubtitles (Tiedemann, 2012), a parallel corpus of movie subtitles. They trained an attentional encoder-decoder NMT system using Groundhog[1] (Van Merriënboer et al., 2015) and used a joint BPE to represent the texts with a fixed vocabulary of subword units with size 90,000.

In their research the authors proved that it is possible to control the honorifics produced at test time by marking up the source side of the training data with a feature that encodes the use of honorifics on the target side. To automatically annotate politeness on a sentence level, they made use of rules based on the morphosyntactic annotation by ParZu (Sennrich, Volk, and Schneider, 2013), marking each instance as either being informal, formal or neutral (if none of the other two applied). Interestingly, to ensure that the engine learned to not overproduce honorifics when no side constraint was provided, they only marked a subset of the training instances with a politeness feature and set the probability that an instance was marked to 0.5.

They tested translations without side constraints (neutral) and with constrains (polite and informal), achieving 20.7, 17.9 and 20.2 BLEU points respectively. In another oracle experiment, they used the politeness label of the reference to determine the side constraint, which simulates a setting in which a user controls the desired politeness. In that case, BLEU was strongly affected by the choice in politeness: results showed an improvement of 3.2 BLEU points over the baseline.

3 Experimental setup

In this section we describe the steps taken to train our politeness-aware systems for the English-Castilian language combination. Firstly, we present the procedure followed to select, process and divide the data set according to the different levels of politeness. Secondly, we introduce the features of the fine-tuned and multilingual NMT models used for the experiment.

3.1 Data set

There is no bilingual data annotated according to its level of politeness for the English>Spanish language pair that can be used to train an NMT system. Therefore, our first task involved creating a set with those characteristics. We opted for the OpenSubtitles corpus (Tiedemann, 2012) and followed an automatic classification approach to divide it into the required subsets. OpenSubtitles consists of a parallel collection of user contributed subtitles of films and TV programs in various languages. The English-Spanish subset accounts for 46 million parallel segments. It must be noted that the alignment is not always correct but, most importantly for our experiment, the texts are not identified by diatopic varieties. This means that the bilingual corpus might contain instances from several dialects of the Spanish language, which use honorifics differently. In particular, in contrast to Castilian, in a number of Latin American countries the form

[1] github.com/sebastien-j/LV_groundhog

usted is used for familiar situations. Yet, we believed that the advantages of this corpus (mainly the orality, which results in the frequent use of second person pronouns) outweighed such disadvantage and turned it into an interesting case study.

Let us remember that one way of marking politeness in Spanish is by using different honorifics.[2] In Castilian, the personal pronoun *tú* tends to be the form used in situations where interlocutors are (relatively) close, while *usted* tends to be the form used to show respect and distance. Given that this is similar to how German works, we adapted the classification approach by Sennrich, Haddow, and Birch (2016b) to allocate segments into three register subsets: informal, formal and neutral (cases with no second person pronouns or verbs).

This involved a two-step exercise. Firstly, we searched for occurrences of lexical forms that belong to the paradigms of *tú* and *usted* using regex.[3] However, Spanish is a predominantly pro-drop language, that is, pronouns can be omitted if their information can be inferred pragmatically or grammatically. If we only use sentences with overt pronouns to train or fine-tune the formal and informal engines, the language produced could sound quite unnatural, since such engines might over-generate pronouns. To counterbalance this behaviour, we also identified grammatical forms, in particular, verbs, in segments with no overt lexical forms.

Remember that, in Castilian Spanish, the informal pronoun *tú* requires the verb to be conjugated with the mark for the second person singular, while the formal pronoun *usted* requires the verb to be conjugated with the mark for the third person singular. As a result, using Spacy[4], if the Spanish sentence contained a verb conjugated in the second person, we classified it as *informal*; if it contained a verb conjugated in the third person, we classified it as *formal*, and if there was no verb or there was a verb conjugated using a different person, we classified it as *neutral*. The challenge here lies in that the third person forms are ambiguous: they can belong to either *usted* or to the regular third person pronouns *él*, *ella*, *ellos* or *ellas*. The same happened with other lexical forms such as possessives *su*, *suyo*, *suya*, etc. Because Spacy could not disambiguate these cases efficiently, to classify these correctly, we searched for *you*, *your* or *yours* in the parallel source segment to identify second-person cases (Sennrich, Haddow, and Birch, 2016b).

We checked the accuracy of our approach by analysing a random set of 100 instances from each of the subsets (50 extracted using the regex approach, and 50 extracted by parsing). The majority of the segments was correctly classified, with an accuracy of 99%, 76% and 93% for the informal, formal and neutral subsets, respectively.

During a qualitative analysis of the results, we observed that to a large extent, the incorrect instances were due to errors in the disambiguation of third person verbs, the misalignment of the English *you*, originally misaligned source and target segments and segments of dubious quality. Solving the first two cases would require implementing a more complex disambiguation process and were not modified. After all, we expected that the amount of false positives in the formal corpus would not hurt the performance of our engines to a great extent, and if so, it could also shed some light on our study when comparing the different engines. However, for the problem with segments of dubious quality, we filtered our data using Marian's scorer[5] (Junczys-Dowmunt et al., 2018) following the advice of Bane and Zaretskaya (2021). The scorer calculates negative log likelihood of a segment with respect to a model. We used the Helsinki−NLP EN>ES model[6] Tiedemann and Thottingal (2020) and filtered our data with a threshold of -6.5, which reduced the data sets around 20% (see Table 1 for the distribution of register classes of the corpus).[7]

Not surprisingly, the number of segments

[2] In his study on registers, Briz (2010) gives a definition of what he denotes as the prototype of colloquial and formal registers. Among their characteristics, he mentions the use of an informal or a formal tone, and refers to politeness as one of the several features that conform register.

[3] Informal lexical forms: *tú, tu, tus, contigo, tuyo, tuyos, tuya, tuyas, ti, te, vosotros, vosotras, vuestro, vuestros, vuestras vuestros*; formal lexical forms: *usted, ustedes, le, les, su, sus, se, suyo, suyos, suya, suyas.*

[4] https://spacy.io

[5] https://marian-nmt.github.io

[6] https://github.com/Helsinki-NLP/Opus-MT

[7] The politeness-specific corpus is open-source and can be freely downloaded from github.com/c-soler-u/exploring_politeness_control

formal subset	1,821,381
informal subset	4,453,708
neutral subset	3,670,602

Table 1: Distribution of the corpus segments across register subsets after full processing.

allocated to each subset is different. Note, however, that a randomly selected even part of each subset was used for training, thus eliminating such unbalances.

3.2 NMT systems

We explored two domain-adaptation approaches to manage politeness in NMT: a fine-tuning approach (FTA) and a multilingual –or multi-register– approach (MRA). We used the Fairseq toolkit[8] (Ott et al., 2019) to train the NMT systems for both approaches. For tokenization and byte-per-encoding (BPE) segmentation, we used Moses[9] and Subword-NMT[10] (Sennrich, Haddow, and Birch, 2016b).

Fine-tuning approach

For the FTA, we first trained a baseline model using 3 million segments containing a balanced mix of formal, informal and neutral subsets (e.g. 1 million segments of each distribution). We trained a joint BPE vocabulary of size 32,000 and applied it to the training data. We used separate vocabularies created with Fairseq and trained a system based on the Transformer architecture (Vaswani et al., 2017) using Adam as an optimizer, a learning rate of 5e-4, dropout of 0.3, label-smoothing of 0.1 and 50 epochs. Our engine was trained with an early-stopping of 5 validation runs.

We then used 700,000 segments from the formal subset and 700,000 segments from the informal subset to fine-tune the baseline system towards these two directions using the last training epoch (see Table 2 for final segment configuration).

For the fine-tuned systems, we reused the BPE code from the baseline engine, but following Subword-NMT best practices (Sennrich, Haddow, and Birch, 2016b), we extracted the vocabulary for each register and passed it along when applying the BPE with a vocabulary threshold of 50 so that the script would only produce symbols which also appeared in the vocabulary. According to the authors, learning BPE on the concatenation of the involved languages increases the consistency of segmentation, and reduces the problem of inserting/deleting characters when copying/transliterating names. Moreover, applying a vocabulary to this would prevent words from being segmented in a way that was seen only in the other language (or register in our case). We used the parameters of the baseline system for the fine-tuned systems, which are trained for 10 epochs with early stopping of 2 validation runs reusing the separate vocabularies that were created for the baseline.

Multi-register approach

For the MRA approach we trained two engines. The first followed the work by Sennrich, Haddow, and Birch (2016a) where a portion of segments from the other registers was added to each subset to avoid excessive bias towards the trained register (MRAnoise). The second was treated as a multilingual system where the three different registers replaced the usual languages (MRAnonoise), which allowed us to check if the bias was effectively meaningful for our task. To signal the politeness on the target language, the authors prepend a token to each segment. However, for our research, we made use of Fairseq's implementation to train a multilingual system, which dealt with this process automatically.

We trained the MRAnonoise engine using 1.5 million segments from each register subset amounting to a total of 4.5 million segments (see Table 3). We trained a joint BPE code using the three directions and applied it as for the FTA systems, using separate vocabularies. The English vocabulary was trained using the English source data from all three subsets, while the vocabulary for each respective direction was extracted from their particular training-data. We used the Transformer architecture for multilingual translation from Fairseq and applied the same parameters as the previous model but with shared encoder-embeddings: Adam optimizer, learning rate of 5e-4, label-smoothing of 0.1 and dropout of 0.3. We trained the model for 50 epochs with early stopping of 5.

Starting from the data sets that were used to train the MRAnonoise engine (each set containing 1.5 million parallel segments as shown in Table 3), we trained the MRA-

[8]https://github.com/pytorch/fairseq
[9]https://github.com/moses-smt/mosesdecoder
[10]https://github.com/rsennrich/subword-nmt

Baseline system	Fine-tuned systems		
Training set	**Training set**	**Development set**	**Test set**
3,000,000	696,000	2,000	2,000

Table 2: Number of bilingual segments used for the FTA systems.

Politeness level	Training set	Development set	Test set
informal	1,498,600	700	700
formal	1,498,600	700	700
neutral	1,498,600	700	700
Total	4,495,800	2,100	2,100

Table 3: Number of bilingual segments used for the MRAnonoise system.

	Informal direction	Formal direction	Neutral direction
informal segments	750,000	0	750,000
formal segments	0	1,000,000	750,000
neutral segments	325,000	75,000	750,000
Total segments	1,075,000	1,075,000	2,250,000

Table 4: Number of bilingual segments used for the MRAnoise system.

noise by redistributing portions of the sentences following Sennrich, Haddow, and Birch (2016a) where, in order to reduce bias, the probability of an instance pertaining to either the formal or informal subset is marked to 0.5 (note that we did not re-marked it for each epoch of training) (see Table 4 for data size).

As it can be observed in Table 4, around half of the informal and formal training sets were used for their respective registers, while the other half were added to the neutral register. We also set aside 0.70 million segments from the neutral training set and divided them between the informal and formal sets (0.35 million each). However, to compensate for the higher level of noise in the formal set (see Section 3.1), we reduced this amount in its training data. The BPE code and vocabularies, and the training was carried out as following the same steps used in the MRAnonoise engine.

4 Results

In this section we report the results from the evaluation of each approach. We start by providing the score for a number of automatic metrics to test the overall quality of the systems (Section 4.1). Then, we describe the process and insights gathered from human assessments (Section 4.2).

When generating the translations that are used for testing, we use the last checkpoint from each engine with a beam search of 5 and batch size of 128.

4.1 Automatic evaluation

We carried out a two-fold automatic analysis, that is, we used a specific test set for each register of each engine (e.g. 9 specific test sets in total), as well as a common test set to all the engines. The first intends to test each system on a subset of the specific data distribution collected for their development (set aside prior to training), while the second aims at testing the relative performance of the engines. In order to compile the common set and find a balance between the varying data distributions of the engines, we extracted 200 segments from each of the following specific test sets: 600 segments from the FTA test set (200 from each the informal, the formal and the baseline test sets), and 600 from the MRA test set (200 from each the informal, the formal and the neutral test sets). Therefore, the final common test set contains 1,200 segments.

We obtained the automatic metric scores using MT-Telescope (Rei et al., 2021a) and report results for COMETINHO (Rei et al., 2021b), sacreBLEU (Post, 2018) and chr-F (Popović, 2015). For the common test set, we also perform significance testing using t-tests with bootstrap re-sampling (Koehn, 2004) with default parameters (re-samples of 0.5 and 300 iterations).

For the engine-specific test sets, results show solid +30 BLEU points for all direc-

System	sacreBLEU	COMETINHO	chr-F
FTA baseline	35.3	38.9	56.8
FTA informal	39.7	47.5	58.7
FTA formal	35.0	37.3	56.9
MRAnonoise neutral	36.8	38.6	58.0
MRAnonoise informal	40.3	46.6	59.5
MRAnonoise formal	38.4	42.2	59.3
MRAnoise neutral	30.3	25.5	53.0
MRAnoise informal	32.8	30.0	54.1
MRAnoise formal	31.8	27.8	55.1

Table 5: Automatic metric scores for all systems on the specific test sets.

System	sacreBLEU	COMETINHO	chr-F
FTA baseline	35.4**	36.3**	56.7**
FTA informal	30.5	28.3	52.7
FTA formal	30.7	27.3	53.1
FTA average	32.2	30.6	54.2
MRAnonoise neutral	30.1	23.6	52.8
MRAnonoise informal	32.3	30.8**	55.0**
MRAnonoise formal	33.7**	30.8**	55.5**
MRAnonoise average	32.0	28.4	54.4
MRAnoise neutral	**36.5***	**38.1***	**57.5***
MRAnoise informal	34.1*†	35.1*†	55.8*†
MRAnoise formal	32.8†	30.1†	55.2†
MRAnoise average	**34.8**	**34.4**	**56.3**

Table 6: Automatic metric scores for all systems on the 1,200 segment common test set. Best results are highlighted in bold. Statistically significant results are also marked: * for comparisons between the MRAnonoise and MRAnoise engines per direction, † for MRAnoise and FTA, and ** for MRAnonoise and FTA.

tions (see Table 5). In general, the informal directions achieve the overall highest scores for each approach, while the formal directions tend to achieve better scores than their respective baseline/neutral directions (except for the FTA engine, where the baseline outperforms the formal direction). Even when this seems to emerge as a trend, note that further analysis is required for precise conclusions, as these particular test sets are not directly comparable.

Comparisons across systems based on the common test set (see Table 6) show that the baseline/neutral engines achieve some of the best scores even when they obtained the worst scores in their specific test sets. This strengthens the idea that the informal engines might be in general over-fitted to their training data, while the baseline/neutral models might be better suited to respond to other data.

If we turn to the MRA engines, we see that MRAnoise achieves significantly better results than its MRAnonoise counterpart for the neutral and informal registers, and also significantly better results for the informal and the formal registers than the FTA engine. When comparing the MRAnonoise and the FTA engines, the informal and formal registers of the former significantly outperform the latter, yet, not the baseline. This might imply that, when fine-tuning a baseline to the different registers, there is a bigger drop in performance. This is not the case when training a multi-register model with noise added to each register.

In Table 6, we also present the average performance of each engine (averaging the scores from the baseline/neutral, formal and informal registers). As it can be seen, the directions from the MRAnoise engine achieve the best average scores for all metrics, with a difference of more than 2 points for each metric over the second best engine (FTA). The MRAnonoise engine presents the lowest scores.

4.2 Human evaluation

Automatic metrics are dependent on the reference segments and their original quality. Therefore, in order to have an assessment of the quality from a human perspective, we also performed a set of human evaluations.

For these assessments, we created a test suite *ad-hoc*, from now on LINGtest[11], which contains 50 segments divided into two categories: those with overt second person forms in the source (YOU_FORMS), intended to cover the different forms that *tú* and *usted* can take in Spanish (*you*, *your*, *yours*), and those with no overt forms or verbs (NO_FORMS). This will allow us to check how the systems perform when faced with overt and non-overt cases.

We translated the 50 segments from the LINGtest using each of the 9 directions of the three approaches trained, which amounted to 450 unique translations.

In order to create the sets for evaluators, we allocated 50 segments to each set in a way that all the sets included translations from all engines while no source segment was repeated, and we could collect responses for all 450 translations. Given the subjective nature of the evaluation, we collected three assessments per translation. A total of 30 volunteer evaluators (native or near-native speakers of Spanish with varying expertise in NLP) were asked to score the translations of the LINGtest according to accuracy and fluency on a 5-point scale. Additionally, they were given the opportunity to comment on any aspect they considered relevant. It is important to note that, to avoid bias, they were not aware of the focus of the assessment (politeness) nor that they were evaluating output from different engines.

To obtain the final human results, we averaged the scores for each translation given by each evaluator. For the general system-level score, we averaged the previous segment-scores again. The average inter-annotator agreement of our research was 0.25 (calculated using Fleiss' Kappa).

Results for quality assessment show that all engines achieve adequacy and fluency scores above 4 points, which in our measuring scale means all engines tend to preserve most of the meaning of the original sentence and have good fluency, although they are not flawless (see Table 7). Contrary to automatic metrics, human assessments seem to indicate that the FTA baseline achieves the best adequacy and overall scores, and is the second best for fluency.

Interestingly, for adequacy, we observe that, when compared to the formal and baseline/neutral registers within the same engine, all the informal directions achieve worse results except for MRAnoise. This might indicate that the MRAnoise informal direction indeed benefited from the addition of sentences belonging to the neutral and formal subsets. In fact, average scores for each approach show that MRAnoise achieves the best overall score.

To check whether the performance of the engines degrades with certain types of linguistic phenomena in particular, we next took a more detailed look into the scores given to the different types of segments (YOU_FORMS and NO_FORMS). In Table 8, we present the overall scores (calculated as the mean of adequacy and fluency) for each engine and register, as well as the difference in the performance between the YOU_FORMS and the NO_FORMS segments.

The results show a difference in behaviour according to the type of segment. The engines that were trained with more strictly filtered data show better performance in the YOU_FORMS segments, while their performance decreases with the NO_FORMS segments to some extent. However, the directions trained with data belonging to the different register subsets achieve worse performance in the YOU_FORMS segments but do not experience such a sharp decrease in quality with the NO_FORMS segments. This calls for further experiments to establish the optimal proportion of register segment types at training-time.

On the other hand, while MRAnonoise achieves some of the best results for the informal and formal registers in the YOU_FORMS segments, its neutral register lags behind, which suggests that this engine did not benefit from being trained with only segments extracted from the neutral subset.

We carried out a final analysis to focus on the specific handling of the honorifics by the engines. We reviewed all the translations in the LINGtest and annotated (1) whether the systems overgenerated honorifics for seg-

[11] Can we found in Appendix A.

System	Adequacy	Fluency	Overall
FTA baseline	**4.51**	4.45	**4.48**
FTA informal	4.05	4.32	4.18
FTA formal	4.18	4.14	4.16
FTA average	4.25	4.30	4.28
MRAnonoise neutral	4.21	4.16	4.18
MRAnonoise informal	4.13	4.43	4.28
MRAnonoise formal	4.47	4.35	4.41
MRAnonoise average	4.27	4.31	4.29
MRAnoise neutral	4.39	4.37	4.38
MRAnoise informal	4.36	4.42	4.39
MRAnoise formal	4.34	**4.47**†	4.35
MRAnoise average	**4.36**	**4.42**	**4.37**

Table 7: Average human assessment scores for adequacy and fluency on the LINGtest. Best scores are in bold. † marks statistically significant differences when comparing the MRAnoise and the FTA approach.

System	YOU_FORMS	NO_FORMS	DIFFERENCE
FTA baseline	4.38	**4.59**†	+0.21
FTA informal	4.41	3.79	-0.62
FTA formal	4.26	4.08	-0.18
MRAnonoise neutral	4.16	4.23	+0.7
MRAnonoise informal	**4.6**	3.87	-0.73
MRAnonoise formal	4.45	4.47	+0.02
MRAnoise neutral	4.51*	4.21	-0.3
MRAnoise informal	4.44	4.33*†	-0.11
MRAnoise formal	4.38	4.28	-0.1

Table 8: Average human assessment scores for adequacy and fluency on the LINGtest per segment type. Best scores are in bold. Statistically significant results are also marked: * for comparisons between the MRAnonoise and MRAnoise engines per register, † for MRAnoise and FTA, and ** for MRAnonoise and FTA.

System	POLITENESS ACCURACY	HALLUCINATIONS
FTA informal	96.7%	20%
FTA formal	90%	15%
MRAnonoise informal	100%	50%
MRAnonoise formal	96.7%	0%
MRAnoise informal	90.3%	5%
MRAnoise formal	90.3%	5%

Table 9: Politeness test of segments with second person forms in the source.

ments with no overt second person forms or no verbs in the source segments (NO_FORMS segments); and (2) whether they actually produced the correct formal and informal forms as intended (YOU_FORMS). *Politeness accuracy* is calculated as the number of times the informal and formal engines outputted the right register divided by the total number of YOU_FORMS instances (30), while *Halluciations* is calculated as the total number of segments with overgenerated honorifics divided by the total number of NO_FORMS instances (20). Scores for *Politeness accuracy* were not calculated for the neutral and baseline systems, since they were not intended to handle any particular register and, due to their training data, did not overgenerate honorifics in the NO_FORMS segments.

Results show that honorifics are very accurately handled in all engines and registers, with over 90% of the instances correctly generated (see Table 10). MRAnonoise is the best approach, at par with the FTA for the

informal register. However, we observe different tendencies with regards hallucinations: overall MRAnoise is the best performer, with the more consistent low level at 5%. MRAnonoise is able to avoid all overgeneration for the formal register, but reaches a 50% high for the informal register. Meanwhile, the proportions for the FTA engine remain between 15% and 20%.

5 Conclusions and future work

In this work, we studied ways to control politeness in NMT for Castilian Spanish. Our first contribution to this topic was the creation of politeness-specific sets for the new language pair –based on the approach used by Sennrich, Haddow, and Birch (2016a) for German–. By adapting their methodology, we classified Spanish segments from the OpenSubtitles corpus into three levels of politeness (formal, informal and neutral) with an average accuracy on a population sample of over 90%.

We then used the separate subsets to explore two main domain-adaptation techniques to address politeness in English>Castilian Spanish NMT: fine-tuning and multilingual models.

Automatic evaluation results seem to show that, overall for our case, the multi-register approach with noise might be better suited than the fine-tuning approach when a balance between accuracy at choosing the honorific and performance in the different registers is the key. However, whether these results are due to the domain adaptation technique used for training or to the incorporation of noise into the training data should be further studied.

We extended the evaluation of the results with multiple human evaluations, which help to understand the handling of the registers more in detail. According to the adequacy and fluency judgements, the ranking of the engines varies slightly. Adequacy and overall quality seem to be better achieved by the baseline system trained as part of the fine-tuning approach. The best overall fluency is achieved by the multi-register formal engine trained with noise.

It is interesting to note that several annotators reported concern about their assessment, stating that they were not too sure about how to evaluate politeness-related issues. We take these statements not as a weakness of the evaluation but rather as a clear sign that politeness is a relevant feature to establish the appropriateness -and quality- of a translation. Therefore, as politeness can be a factor that can direct the assessment, we suggest that evaluations, whether register-related or general- may benefit from including specific guidelines as to how to treat register-related issues.

Additionally, the specific politeness-related analyses showed that the engines did not always perform consistently for the different types of segments that display (or omit) register-related elements. In any case, we observed that the accuracy of honorifics was above 90% for all engines and registers. In terms of hallucinations, multi-register models performed better –except for the informal direction trained with no noise– while the FTA models seem to have suffered from some degree of catastrophic forgetting, which can lead to a worse overall performance of those models in segments with no second person forms and no verbs in the source when compared to the MRAnoise system.

In this line, in future work, we aim to explore the use of mixed fine-tuning (as proposed by Chu, Dabre, and Kurohashi (2017)) in the quest for palliating catastrophic forgetting in fine-tuned systems. Additionally, driven by the growing interest in the research and development of register-aware NLP technologies, we also intend to work with other features that configure politeness for Spanish beyond the use of honorifics and to provide an enlarged and refined version of the register-annotated corpus created for this work with the aim of contributing to the community with a high-quality resource to be used in other NLP applications beyond MT.

Acknowledgements

The research leading to this work was partially funded by the *SignOn* RIA ICT-57-2020-101017255 project, the *TANDO* ELKARTEK 20/49 project and the IT1570-22 project (Basque Government). We would also like to thank the volunteers who freely participated in the human evaluation process. Finally, we would like to express our gratitude to Fred Bane for laying the foundations of this research.

Bibliography

Aharoni, R., M. Johnson, and O. Firat. 2019. Massively multilingual neural machine translation. 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 3874–3884, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Bane, F. and A. Zaretskaya. 2021. Selecting the best data filtering method for NMT training. In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pages 89–97, Virtual, August. Association for Machine Translation in the Americas.

Bapna, A., N. Arivazhagan, and O. Firat. 2019. Simple, scalable adaptation for neural machine translation. *arXiv preprint arXiv:1909.08478*.

Briz, A. 2010. Lo coloquial y lo formal, el eje de la variedad lingüística. *De moneda nunca usada: Estudios dedicados a José Mª Enguita Utrilla*, 125:133.

Brown, P. 2015. Politeness and language. *The International Encyclopedia of the Social and Behavioural Sciences (IESBS),(2nd ed.)*, pages 326–330.

Bulte, B. and A. Tezcan. 2019. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1800–1809, Florence, Italy, July. Association for Computational Linguistics.

Chu, C., R. Dabre, and S. Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada, July. Association for Computational Linguistics.

Chu, C. and R. Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, page 1304–1319, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Dinu, G., P. Mathur, M. Federico, and Y. Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy, July. Association for Computational Linguistics.

Etchegoyhen, T., E. Martínez Garcia, A. Azpeitia, G. Labaka, I. Alegria, I. Cortes Etxabe, A. Jauregi Carrera, I. Ellakuria Santos, M. Martin, and E. Calonge. 2018. Neural machine translation of Basque. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 139–148, Alacant, Spain, May. European Association for Machine Translation.

Feely, W., E. Hasler, and A. de Gispert. 2019. Controlling japanese honorifics in english-to-japanese neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 45–53.

Halliday, M., A. McIntosh, and P. Stevens. 1964. *The language Science and Language Teaching*. London. Longman.

Haugh, M. 2005. The importance of "place" in japanese politeness: Implications for cross-cultural and intercultural analyses. *Japanese Politeness: Implications for Cross-Cultural and Intercultural Analyses*, 2(1):41–68.

Junczys-Dowmunt, M., R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. F. Aji, N. Bogoychev, A. F. T. Martins, and A. Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July. Association for Computational Linguistics.

Kell, G. 2018. *Overcoming catastrophic forgetting in neural machine translation*. Ph.D. thesis, MPhil dissertation, University of Cambridge.

Kirkpatrick, J., R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho,

A. Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Kobus, C., J. Crego, and J. Senellart. 2017. Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378, Varna, Bulgaria, September. INCOMA Ltd.

Koehn, P. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Koehn, P. and R. Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August. Association for Computational Linguistics.

Luong, M.-T. and C. Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam, December 3-4.

Matthiessen, C. and M. Halliday. 1997. *Systemic functional grammar*. Amsterdam and London: Benjamins & Whurr.

Ott, M., S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Popović, M. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Post, M. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Post, M. and D. Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. 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 1314–1324, New Orleans, Louisiana, June. Association for Computational Linguistics.

Rei, R., A. C. Farinha, C. Stewart, L. Coheur, and A. Lavie. 2021a. MT-Telescope: An interactive platform for contrastive evaluation of MT systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 73–80, Online, August. Association for Computational Linguistics.

Rei, R., A. C. Farinha, C. Zerva, D. van Stigt, C. Stewart, P. Ramos, T. Glushkova, A. F. T. Martins, and A. Lavie. 2021b. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online, November. Association for Computational Linguistics.

Sennrich, R., B. Haddow, and A. Birch. 2016a. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California, June. Association for Computational Linguistics.

Sennrich, R., B. Haddow, and A. Birch. 2016b. 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, August. Association for Computational Linguistics.

Sennrich, R., M. Volk, and G. Schneider. 2013. Exploiting synergies between open resources for German dependency parsing, POS-tagging, and morphological analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 601–609, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Tiedemann, J. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Tiedemann, J. and S. Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal, November. European Association for Machine Translation.

Van Merriënboer, B., D. Bahdanau, V. Dumoulin, D. Serdyuk, D. Warde-Farley, J. Chorowski, and Y. Bengio. 2015. Blocks and fuel: Frameworks for deep learning. *arXiv preprint arXiv:1506.00619*.

Vanmassenhove, E., D. Shterionov, and M. Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online, April. Association for Computational Linguistics.

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 *31st Conference on Neural Information Processing Systems (NIPS 2017)*, pages 1–11, Long Beach, CA, USA, December.

A Appendix: LINGtest

YOU_FORMS	NO_FORMS
- You should go to the doctor if you are feeling sick.	- Nonsense!
- What did you do yesterday?	- Why not?
- We are available via Whatsapp to solve any questions you may have during the purchase	- How cool!
- It was you who started the fight.	- Seriously?
- Who did it? Was it you?	- Postal code
- Yesterday, we went out for a couple of drinks downtown. What about you guys?	- Next item
- Is it you, Tom?	- Hey, there!
- You need to be the one that picks up the parcel.	- Welcome!
- Can you check your agenda and let me know when you are free?	- Where? There?
- How was your experience with us?	- Customized delivery services
- Did you break your arm?	- We are delighted to be here today.
- I believe that T-shirt was yours.	- I am really happy to be here today.
- Let's take my car, not yours.	- They were suppose to come today.
- Please enter your address.	- We enjoyed it so much!
- Where do you wish to receive your items?	- Personally, I think that is not true.
- Your purchase is almost done!	- He was such a nice person.
- How was your experience with us?	- She moved to Madrid to attend University.
- Come with us, please!	- Offering customized delivery services since 1996.
- Contact us at XXXXX.	- They asked me whether I wanted a refund.
- Call me when you get home.	- Let's go together.
- Click on the item you wish to purchase.	
- Look at this.	
- Can I come with you?	
- We have all these new items for you!	
- No, thank you	
- Please, do not hesistate to contact us and ask for a refund.	
- I made all this for you.	
- We would love to go to the cinema with you tonight.	
- Did she come with you?	
- I was waiting for you guys forever!	

Table 10: Test suite for human evaluation.

Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 213-223 recibido 16-12-2022 revisado 29-01-2023 aceptado 07-02-2023

Generación y pesado de skipgrams y su aplicación al análisis de sentimientos

Skipgrams Generation and Weighting and its Application to Sentiment Analysis

Javi Fernández, Yoan Gutiérrez, Patricio Martínez-Barco
Department of Software and Computing Systems
University of Alicante
{javifm,ygutierrez,patricio}@dlsi.ua.es

Resumen: El modelado de skipgrams es una técnica para la generación de términos multi-palabra que conserva parte de la secuencialidad y flexibilidad del lenguaje. Sin embargo, en algunos casos el número de skipgrams generados puede ser excesivo a medida que se aumenta la distancia entre palabras. Además, esta distancia no suele ser tenida en cuenta a la hora de valorar los términos que se generan. En este trabajo proponemos una técnica para la generación y filtrado eficientes de skipgrams y un esquema de pesado que tiene en cuenta la distancia entre los términos, dando más importancia a aquellos más cercanos. Aplicaremos y evaluaremos estas propuestas en la tarea de análisis de sentimientos.
Palabras clave: skipgrams, generación de términos, pesado de términos, análisis de sentimientos.

Abstract: Skipgram modelling is a technique for generating multi-word terms that preserves some of the sequentiality and flexibility of the language. However, in some cases the number of skipgrams generated may become excessive as the distance between words increases. Moreover, this distance is often not taken into account when evaluating the terms that are generated. In this paper we propose a technique for efficient skipgram generation and filtering, and a weighing scheme that takes into account the distance between terms, giving more importance to those closer. We will apply and evaluate these proposals in the task of sentiment analysis.
Keywords: skipgrams, term generation, term weighting, sentiment analysis.

1 Introducción

La técnica del *modelado de skipgrams* consiste en obtener términos multi-palabra[1] a partir de un texto, de forma similar a los n-gramas, pero permitiendo omitir algunas palabras intermedias. Más concretamente, en un *k-skip-n-gram*, n determina el número de palabras de los términos generados, y k el número de palabras que se omiten. También se puede trabajar con un número máximo de palabras por término y con un número máximo de omisiones. En este trabajo nos diferenciaremos estos casos denominándolos n_{max} y k_{max} respectivamente. Con esta técnica estamos generando términos adicionales que conservan parte de la secuencialidad de las palabras originales, pero de forma más flexible que los n-gramas. Cabe señalar que los n-gramas pueden definirse como skipgrams en los que $k = 0$ (sin omisiones o saltos).

Sin embargo, la principal desventaja de esta técnica radica en que el número de skipgrams generados puede ser muy grande. Para hacernos una idea del tamaño máximo que podrían alcanzar, obtener todos los términos posibles utilizando n-gramas (de cualquier tamaño) tiene una complejidad de $O(n^2)$ (en este caso n es el número de palabras del texto), pero utilizando skipgrams (cualquier tamaño y cualquier número de palabras omitidas), la complejidad sería del orden de $O(2^n)$. En la Figura 1, se puede ver un ejemplo práctico utilizando uno de los conjunto de datos del TASS 2020 (Vega et al., 2020), concretamente el conjunto `train` en castellano (`es`) de la

[1] En este artículo denominaremos *términos* a una secuencia de palabras cuyo orden importa.

ISSN 1135-5948 DOI 10.26342/2023-70-17

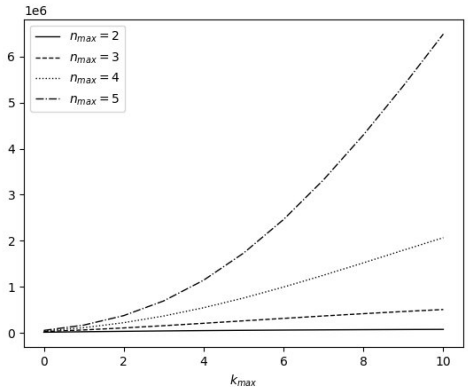


Figura 1: Número de términos generados utilizando skipgrams, según se va aumentando el valor de k_{max}, para diferentes valores de n_{max} máximo, en el conjunto de datos `train` de la tarea `1.1` del TASS 2020 en castellano.

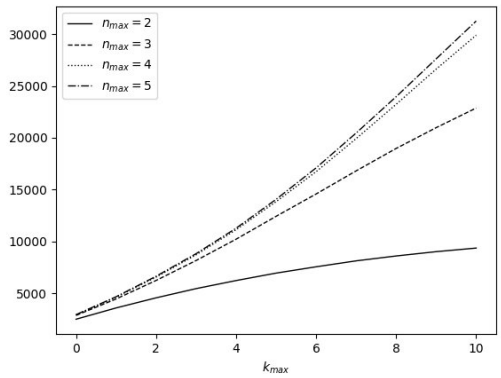


Figura 2: Número de términos generados utilizando skipgrams, según se va aumentando el valor de k_{max}, para diferentes valores de n_{max}, eliminando los que aparecen solo una vez en el conjunto de datos `train` de la tarea `1.1` del TASS 2020 en castellano.

tarea `1.1`, y generando términos con diferentes valores de n_{max} y k_{max}. Este es un conjunto de datos relativamente pequeño, con 1126 documentos y 4222 palabras diferentes, y en el caso de $n_{max} = 5$ y $k_{max} = 10$ máximos el número de términos generados casi alcanza los 6,5 millones.

Existen varios motivos para filtrar los términos generados, sobre todo en el contexto del aprendizaje automático. Por un lado, el número de skipgrams generados puede ser excesivo en algunos casos. Existen técnicas y modelos que nos sería imposible utilizar si disponemos de pocos recursos (procesamiento, memoria, espacio, tiempo) ya que no pueden manejar un número tan grande de términos o características. Por otro lado, reducir el número de términos también mejora el rendimiento de los sistemas que los utilizan, y puede disminuir el ruido si se seleccionan de la manera adecuada, obteniendo mejores resultados en ciertas tareas. Afortunadamente, una estrategia sencilla como eliminar los términos que aparecen solo una vez en el conjunto de datos puede reducir drásticamente la cantidad de términos, como se puede observar en la Figura 2, con la misma configuración que en el ejemplo anterior, donde el número de términos finales en el caso de $n_{max} = 5$ y $k_{max} = 10$ desciende hasta algo más de 30000 (una reducción de más del 99 % de términos).

No obstante, lo mencionado anteriormente no elimina el hecho de que para reducir el número de términos, primero debemos generarlos *todos* (sobretodo cuando los criterios del filtrado son estadísticos), algo que puede ser inviable en algunos conjuntos de datos o configuraciones. En este trabajo proponemos una técnica para realizar este filtrado de manera simultánea a la generación de los términos, con el objetivo de minimizar el número de generaciones realizadas y mejorar la eficiencia. Esta técnica se explicará en detalle en la Sección 3.1.

Por otra parte, la generación de términos utilizando el modelado de skipgrams ofrece una información valiosa que es frecuentemente ignorada: la distancia entre las diferentes ocurrencias del término. Saber si un término ha sido generado mayoritariamente por palabras adyacentes o cercanas, o al contrario por palabras más alejadas entre sí, es una información que podemos aprovechar tanto para la selección de términos como para la propia tarea en la que vamos a utilizar los términos. En este trabajo proponemos una forma de pesar los términos para aprovechar esta información, detallada en la Sección 3.2.

Este trabajo sigue la siguiente estructura. En la Sección 2 estudiaremos el uso de la técnica de modelado de skipgrams en la actualidad y la problemática asociada. La Sección 3 describirá las técnicas para la generación y pesado de términos mencionadas previamente. A continuación, en la Sección 4 aplicaremos las propuestas a la tarea de análisis de sentimientos en dos conjuntos de datos diferentes, para comprobar en qué medida es-

tas técnicas pueden influenciar los resultados. Finalmente, en la Sección 5 explicaremos las conclusiones sacadas de este trabajo y propondremos nuevas líneas para su continuación en el futuro.

2 Estado actual

La técnica del modelado de skipgrams tuvo auge hace años en el campo del PLN (Guthrie et al., 2006), y a día de hoy existen muchos enfoques que utilizan el modelado de skipgrams para relacionar y contextualizar palabras. Sin embargo, la mayoría de ellos solo contemplan las relaciones entre términos de par en par, utilizan esta técnica para crear un contexto, o siguen utilizando palabras o n-gramas (o un vector que representa una palabra) como unidades básica de información (Mikolov et al., 2013; Church, 2017; Vaswani et al., 2017; Zhao et al., 2017). En la actualidad, prácticamente todas las menciones a los skipgrams se refieren a su uso como contexto para generar *word embeddings* (Peng et al., 2020; Du et al., 2020; Santos et al., 2021).

Como hemos mencionado en la sección anterior, aumentar el número de términos u omisiones puede dar lugar a un número demasiado grande de combinaciones, por lo que a menudo no se aprovecha todo el potencial de esta técnica. El trabajo de Shazeer, Pelemans, y Chelba (2015) muestra como con valores grandes ($n = 5$, $k = 10$) se generan más de 60 mil millones de términos para algunos conjuntos de datos. Uno de los trabajos que más se ha centrado en generar skipgrams de manera eficiente es el de Gompel y van den Bosch (2016), donde en un primer paso se obtienen n-gramas con un filtrado progresivo, tras el cual los skipgrams son generados y filtrados a partir de esos n-gramas pero eliminando ciertas palabras intermedias. El método parece muy eficiente en términos espaciales y temporales. Sin embargo, el foco de este trabajo es la eficiencia en la generación y filtrado de n-gramas, no en los skipgrams, y no se estudia su repercusión en otras tareas. Otros trabajos, como por ejemplo los de Nguyen y Grishman (2016) o Hossny et al. (2020), explican brevemente que se generan skipgrams de manera eficiente pero no se dan detalles de la aproximación utilizada para hacerlo.

Respecto a aprovechar la información sobre la distancia para valorar los términos generados mediante el modelado de skipgrams, en la literatura podemos encontrar algunas aproximaciones, aunque no es algo común. Uno de los trabajos más enfocados en aprovechar esta información es el de Chang, Lee, y Lai (2017), en el que se propone una función gaussiana para valorar la relación de pares de palabras, con el objetivo de mejorar `word2vec` pero no para la generación de términos, aunque se muestra que los resultados mejoran al tener en cuenta esta información. Otros trabajos como Komninos y Manandhar (2016) o Mimno y Thompson (2017) mencionan que tienen en cuenta la distancia pero no se indica el método.

3 Propuesta

En este trabajo realizaremos dos propuestas. Por un lado proponemos una técnica para la generación y filtrado eficientes de skipgrams, con el objetivo de evitar la generación excesiva de términos. Por otro lado, diseñamos un esquema de pesado que tiene en cuenta la distancia entre las palabras utilizadas para crear los términos, que intenta dar mayor importancia a aquellos más cercanos.

3.1 Filtrado progresivo

Lo más común a la hora de generar términos a partir un conjunto de textos (crear una *bolsa de palabras*) es extraerlos todos luego elegir aquellos que dan más información utilizando diferentes técnicas o heurísticas. Sin embargo, como hemos visto previamente, el número de términos puede ser considerablemente grande en algunos casos, y procesar todos ellos puede requerir más recursos de los disponibles (tanto temporales y espaciales). Con el fin de obtener los términos más importantes de manera eficiente, intentando evitar tener que generar la totalidad de ellos, nuestro objetivo es filtrar los términos durante el propio proceso de generación.

Para ello aprovecharemos el hecho de que los términos con un cierto número de palabras tienen características en común con las palabras (u otros términos) que la forman. Por ejemplo, si tenemos un término multipalabra $t_1 = (w_1 w_2)$, en el que w_1 es la primera palabra y w_2 la segunda, podemos asegurar que t_1 aparecerá como máximo en tantos documentos como w_1 o w_2. Si nuestro criterio de filtrado es aparecer en un mínimo de documentos, si w_1 no lo cumple, entonces t_1 tampoco. Tampoco lo cumpliría ningún otro término derivado de t_1, como por ejem-

plo $t_2 = (w_1 w_2 w_3)$. Por lo tanto, sabiendo que w_1 no cumple este criterio, podemos evitar generar cualquier derivado, tanto t_1 como t_2, mediante un algoritmo de *ramificación y poda* (ver Figura 3), donde los términos generados se generan de izquierda a derecha a partir de otros generados previamente.

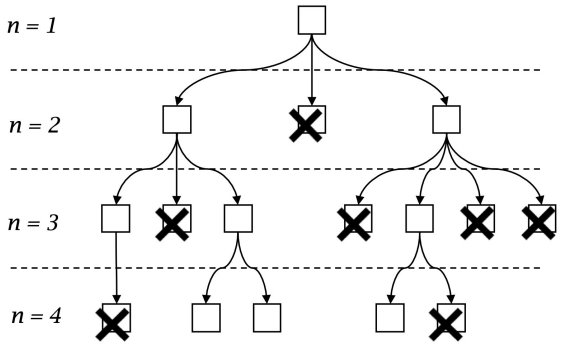


Figura 3: Ejemplo de ramificación y poda. Con una `X` se marcan aquellos términos que no seguían el criterio de filtrado, evitando la generación de todo un subárbol de términos derivados.

Sin embargo no todos los criterios de filtrado pueden hacerse durante la generación. Existen filtrados (la mayoría) que si se realizan durante la generación eliminarán más términos de los esperados, y nuestro objetivo es mejorar la eficiencia pero con el mismo resultado que filtrando al final. Por ejemplo, el porcentaje de ocurrencia en una categoría de un conjunto de datos etiquetado (normalmente denominado en estadística $P(t, c)$) no es candidato para realizarse de manera progresiva. Si un término $t_1 = (w_1 w_2)$ aparece un porcentaje de veces en la categoría c, no podemos saber si el término $t_2 = (w_1 w_2 w_3)$ aparecerá en más o en menos proporción hasta que no lo hayamos generado, por lo que este filtrado debería realizarse tras la generación. En el ejemplo de la Figura 4, si nuestro criterio de filtrado es que el 75 % de las ocurrencias de un término deben ser en documentos positivos, el primer término «qué» y el siguiente «qué máquina» no lo cumplirían, pero aún así no debemos dejar de generar el resto de términos ya que perderíamos el término «qué máquina eres» que sí que lo cumple y puede ser interesante.

Como ejemplos de criterios de filtrado progresivo podemos mencionar *a)* un mínimo de ocurrencias en el conjunto de datos, *b)* un mínimo de ocurrencias en una categoría concreta, *c)* combinaciones de palabras prohibidas,

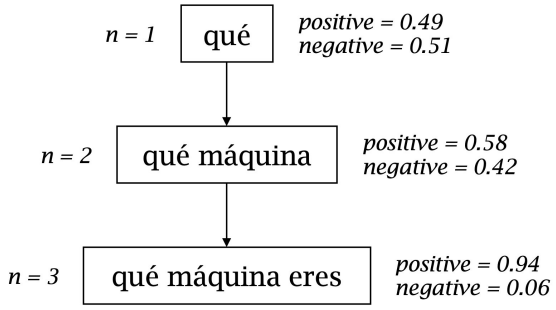


Figura 4: Ejemplo de generación de términos de hasta $n = 3$ palabras, donde *positive* el porcentaje de documentos positivos y *negative* el porcentaje de documentos negativos (ejemplo ficticio).

d) combinaciones de categorías gramaticales (PoS) prohibidas, etc. Como criterios de filtrado que no puede realizarse de manera progresiva podemos destacar *a)* una proporción mínima de ocurrencias en una categoría concreta, *b)* un umbral de puntuación obtenida a partir de un algoritmo de selección de características, etc. En el presente trabajo solo estudiaremos el mínimo de ocurrencias con el fin de mantener la aproximación simple y evitar propagar errores de herramientas externas en nuestro estudio.

En pseudocódigo para este procedimiento se puede observar en el Algoritmo 1, donde `Tokenize` se encarga de extraer las palabras de los textos, `Generate` realiza la generación de términos de izquierda a derecha a partir de otros términos calculados anteriormente (añadiendo palabras y realizando las omisiones oportunas), `Filter` realiza el filtrado progresivo seleccionado y `FilterFinal` realiza el filtrado final con los criterios que no se pueden comprobar progresivamente.

Datos: dataset
$X \longleftarrow$ `Tokenize`$(dataset)$;
$T_1 \longleftarrow$ `Generate`(X, k_{max});
$T_1 \longleftarrow$ `Filter`(T_1);
para $n \leftarrow 2$ **a** n_{max} **hacer**
  $T_n \longleftarrow$ `Generate`(X, k_{max}, T_{n-1});
  $T_n \longleftarrow$ `Filter`(T_n);
fin
$T \longleftarrow$ `FilterFinal`$(T_1, ..., T_{n_{max}})$;

Algoritmo 1: Algoritmo para la generación y filtrado progresivo de términos utilizando skipgrams.

En la Sección 4 realizaremos algunos ex-

perimentos para ver en qué medida se reduce la generación de términos utilizando esta técnica en diferentes conjuntos de datos.

3.2 Pesado por densidad

Las aproximaciones más comunes a la hora de pesar los términos dentro de un documento en una tarea de clasificación son la *binaria* (se pesa con 1 si el término aparece en el texto o 0 si no aparece) o el *conteo básico* (se pesa con el número de veces que aparece el término en el documento). A partir de ahí existen diferentes técnicas de normalización para que los valores se mantengan en el rango $[0, 1]$, como por ejemplo el *tf-idf*.

Estos pesados de términos están pensados para palabras o n-gramas, donde tenemos claro si realmente el término aparece o no en el texto. Sin embargo, en el modelado de skipgrams un término puede aparecer en un texto pero no de manera tan estricta. Por ejemplo, en el texto «La pantalla es muy brillante», podemos decir que el término *pantalla brillante* aparece si realizamos 2 omisiones o saltos. Sin embargo, en otro texto «Tiene una pantalla brillante» el término también aparece, pero sin realizar ninguna omisión. Con el fin de saber si esta distancia influye, proponemos una nueva forma de pesado que tiene en cuenta el número de omisiones, y que hemos denominado *pesado por densidad*. En la Ecuación 1 podemos ver la fórmula propuesta, donde el w_t es el peso del término t, y k es el número de omisiones realizadas para generarlo.

$$w_t = (1 + k)^{-1} \tag{1}$$

En el ejemplo anterior, el término «pantalla brillante» tendría un peso de $w_t = (1 + 2)^{-1} = 0{,}67$ para el primer texto y un peso de $w_t = (1 + 0)^{-1} = 1$ para el segundo texto, dando el peso máximo en el segundo porque no hay omisiones. En el caso de que haya varias ocurrencias de un mismo término en un documento, podemos tomar dos aproximaciones: la primera sería similar al pesado binario, indicando si el término aparece o no, pero en su lugar devolviendo el máximo peso por densidad en ese documento (ver Ecuación 2), y la otra similar al conteo de ocurrencias, devolviendo la suma de todos los pesos por densidad (ver Ecuación 3). En dichas ecuaciones, se calcula el peso $w_{t,d}$ de un término t en un documento d, donde $O_{t,d}$ es el conjunto de todas las ocurrencias del término t en el documento d y k_i es el número de omisiones utilizadas para generar la ocurrencia i.

$$w_{t,d} = max_{i \in O_{t,d}} (1 + k)^{-1} \tag{2}$$

$$w_{t,d} = \sum_{i \in O_{t,d}} (1 + k)^{-1} \tag{3}$$

En la Sección 4 realizaremos la experimentación para comprobar en qué situaciones esta información es valiosa, concretamente en el contexto del aprendizaje automático aplicado a la tarea de análisis de sentimientos.

4 Experimentación y resultados

Todos los experimentos realizados en esta sección se han realizado en conjuntos de datos de textos etiquetados para análisis de sentimientos, concretamente para la tarea clasificación de polaridad. Los conjuntos de datos elegidos están formados por textos cortos o frases, para no generar términos con palabras de frases diferentes y evitar el uso de herramientas de división de frases en textos. El preprocesamiento realizado en cada texto es básico: pasar los textos a minúsculas, eliminar acentos, eliminar repeticiones de caracteres (más de 3), y reemplazar menciones y hashtags por las cadenas `USER` y `HASHTAG` respectivamente (para conjuntos de datos de Twitter[2]). Para extraer las palabras también utilizaremos una aproximación simple, una expresión regular que divide por espacios y signos de puntuación, extrayendo solo palabras formadas por letras y/o números: `(?u)\b\w+\b`.

4.1 Reducción mediante filtrado progresivo

Para comprobar en qué medida el filtrado progresivo puede reducir el número de términos generados, realizaremos diferentes experimentaciones con diferentes conjuntos datos. La primera experimentación la hemos realizado en el dataset del TASS 2020 (Vega et al., 2020), concretamente el conjunto `train` en castellano (`es`) de la tarea `1.1`. Es un conjunto de datos que contiene tweets en castellano (textos cortos obtenidos de Twitter), formado por 1126 documentos y 5314 palabras diferentes, y con un tamaño medio por tweet de 14,82 palabras. En la Tabla 1 se pueden ver los diferentes números de términos

[2] `https://twitter.com`

para diferentes valores de n_{max} y k_{max}, donde SF es el número de términos sin filtrado, FP es el número de términos generados mediante filtrado progresivo, y FT es el número de términos tras el filtrado final. Por ejemplo, en el caso de $n = 3$ y $k = 4$ tenemos reducido los términos generados al 35,80 %, lo que significa que no hemos tenido que filtrar finalmente (FT) el 64,19 % de los términos ya que ni siquiera se han llegado a generar. Destacar que estos porcentajes no se refieren al número de términos filtrados finalmente sino al número de términos que se ha conseguido no tener que generar. En este trabajo nos hemos centrado en la generación, pero realmente el número de términos filtrados es mucho mayor, por ejemplo para el ejemplo anterior con $n = 3$ y $k = 4$ el número de términos final es de 11834 (una reducción del 94,75 %).

La segunda experimentación la hemos realizado en el dataset Movie Reviews (Pang y Lee, 2005), más específicamente el conjunto de frases con polaridad `sentence polarity dataset v1.0`, un conjunto de datos de críticas de películas en inglés, formado por 10695 frases y 18285 palabras diferentes. El tamaño medio por documento es de 18,06 palabras, más grande que el anterior, por lo que esperamos una cantidad mayor de términos generados. En la Tabla 2 se pueden ver los diferentes números de términos para diferentes valores de n y k. Podemos ver que a partir de esas 18285 palabras diferentes se llegan a generar casi 26 millones de términos en el caso de $n_{max} = 5$ y $k_{max} = 5$, reduciéndose a 359939 términos, de ahí la importancia del filtrado al utilizar skipgrams.

Comparando los datos de ambos conjuntos, podemos observar que en el segundo conjunto genera muchos más términos que el primero. Esto es algo de esperar ya que el número de palabras, el número de documentos y el tamaño medio por documento es mayor. Sin embargo, el porcentaje de reducción es menor en el segundo caso. Una posible explicación es que en el segundo conjunto los términos son más susceptibles de observarse más de una vez, algo también explicable por ser un conjunto mayor. En cualquier caso, observamos una reducción significativa de términos generados en valores altos de n_{max} y k_{max}, precisamente en los casos en los que consideramos que es más necesaria.

4.2 Pesado por densidad en análisis de sentimientos

En esta experimentación evaluaremos el comportamiento del modelado de skipgrams en la tarea de análisis de sentimientos. Para ello realizaremos experimentos utilizando *máquinas de soporte vectorial* (SVM) por sus buenos resultados para texto (Yadav et al., 2020). Utilizaremos la implementación de Scikit Learn (Pedregosa et al., 2011) `LinearSVC` con los parámetros por defecto.

Como características utilizaremos los términos extraídos de los conjuntos de datos anteriores (misma aproximación) con diferentes pesados: binario, *tf-idf* y nuestra propuesta de pesado por densidad (3.2, Ecuación 2). También utilizaremos una combinación de ambos pesados para para la experimentación (*tf-idf* y pesado por densidad), donde utilizaremos la fórmula del pesado *tf-idf* pero sustituyendo los número de ocurrencias por el pesado por densidad descrito en la Ecuación 3, dando como resultado la fórmula en la Ecuación 4, donde $w_{t,d}$ es el resultado del pesado por densidad.

$$tfidf_{t,d} = tf_{t,d} \cdot idf_t \rightarrow w_{t,d} \cdot idf_t \tag{4}$$

La evaluación la realizaremos mediante *validación cruzada estratificada* con 10 particiones, utilizando la métrica *F1* con promediado *macro* (misma importancia a todas las polaridades). Nuestro punto de partida o *baseline* serán las configuraciones sin pesado por densidad, y nuestro objetivo será mejorar los resultado de estas configuraciones.

La primera experimentación la realizaremos en el conjunto de datos del TASS 2020 (mencionado en la sección anterior) para diferentes valores de n_{max} y k_{max}. En la Tabla 3 podemos ver los resultados de esta primera experimentación, donde B indica que se ha realizado un pesado binario, D que se ha utilizado el pesado por densidad (fórmula en la Ecuación 2), T que se ha realizado un pesado por *tf-idf*, $D+T$ que se ha realizado un pesado por densidad (fórmula en la Ecuación 3) seguido de un pesado por *tf-idf*, y $DF = 1$, $DF = 2$ y $DF = 3$ que se ha utilizado un criterio de filtrado de términos con ocurrencia mayor o igual a 1, 2 y 3 respectivamente (cabe destacar que en los experimentos donde $DF = 1$ no se realiza ningún filtrado ya que todos los términos aparecen al menos una vez).

$k_{max} \rightarrow$	0	1	2	3	4	5
$n_{max} = 2$, *SF*	16301	27384	37101	45628	53104	59684
$n_{max} = 2$, *FP*	11021	17250	22511	26960	30809	34082
$n_{max} = 2$, *FT*	2661	3868	4971	5951	6840	7647
$n_{max} = 3$, *SF*	31021	69367	116106	168812	225472	284122
$n_{max} = 3$, *FP*	14837	28709	44750	62310	80741	99411
$n_{max} = 3$, *FT*	3094	4925	7052	9366	11834	14556
$n_{max} = 4$, *SF*	45244	122602	240101	399100	598443	834869
$n_{max} = 4$, *FP*	15623	31511	51769	76401	105138	138139
$n_{max} = 4$, *FT*	3164	5134	7456	10055	12999	16377
$n_{max} = 5$, *SF*	58426	183611	409139	762196	1264757	1931543
$n_{max} = 5$, *FP*	15734	31943	52873	78727	109558	145836
$n_{max} = 5$, *FT*	3184	5190	7560	10223	13217	16691

Tabla 1: Número de términos sin filtrar (*SF*), número de términos generados utilizando el filtrado progresivo (*FP*) y número de términos incluyendo el filtrado final (*FT*). El criterio de filtrado es que los términos deben aparecer en más de un documento ($df \geq 2$) en el conjunto de datos `train` de la tarea `1.1` del TASS 2020 en castellano.

$k_{max} \rightarrow$	0	1	2	3	4	5
$n_{max} = 2$, *SF*	124320	229385	323444	407768	483434	551530
$n_{max} = 2$, *FP*	109423	200351	281110	352897	416937	474173
$n_{max} = 2$, *FT*	30583	48491	64134	78428	91451	103555
$n_{max} = 3$, *SF*	284760	701386	1231744	1850792	2536975	3271609
$n_{max} = 3$, *FP*	186643	432725	733293	1075730	1448698	1843316
$n_{max} = 3$, *FT*	41149	75377	113432	156141	202733	252977
$n_{max} = 4$, *SF*	453546	1351736	2783988	4798134	7413739	10625918
$n_{max} = 4$, *FP*	213460	533488	975839	1545202	2242864	3068329
$n_{max} = 4$, *FT*	44397	84651	132123	189098	256428	334776
$n_{max} = 5$, *SF*	616087	2125177	4981590	9634849	16506362	25963045
$n_{max} = 5$, *FP*	220216	560629	1046063	1693926	2522137	3547672
$n_{max} = 5$, *FT*	45223	87235	137480	198747	272537	359939

Tabla 2: Número de términos sin filtrar (*SF*), número de términos generados utilizando el filtrado progresivo (*FP*) y número de términos incluyendo el filtrado final (*FT*). El criterio de filtrado es que los términos deben aparecer en más de un documento ($df \geq 2$) en el conjunto de datos `sentence polarity dataset v1.0` de Movie Reviews en inglés.

Observando los resultados obtenidos podemos ver que utilizar el pesado por densidad mejora el rendimiento de los skipgrams con cualquier configuración ($D > B$, $D+T > T$), menos en el caso de n-gramas, lo que es lógico ya que los n-gramas siempre tienen densidad 1, y únicamente en el caso de skipgrams con $n_{max} = 2, k_{max} = 1, df = 1$. La mejora media utilizando el pesado por densidad es de un 3,35 %, llegando al 7,67 % en el mejor caso. Por lo tanto podemos decir que el uso de la información sobre la distancia en los skipgrams es una información valiosa a la hora de pesar los skipgrams ya que ayuda a mejorar los resultados en esta tarea de análisis de sentimientos, en algunos casos de manera significativa. También podemos observar que sin el pesado por densidad (B y T), la utilización de skipgrams nunca mejora los resultados respecto a los n-gramas. Añadir la información sobre el número de omisiones no es solo recomendable sino también necesaria si queremos

$k_{max} \rightarrow$	0	1	2	3	4	5
$n_{max} = 2$, *DF=1, B*	0,501	0,504	0,495	0,496	0,504	0,497
$n_{max} = 2$, *DF=1, D*	0,501	0,509	0,506	0,506	0,509	0,509
$n_{max} = 2$, *DF=1, T*	0,499	0,503	0,489	0,489	0,481	0,483
$n_{max} = 2$, *DF=1, D+T*	0,499	0,499	0,504	0,502	0,503	0,501
$n_{max} = 2$, *DF=2, B*	0,506	0,497	0,494	0,483	0,478	0,478
$n_{max} = 2$, *DF=2, D*	0,506	0,506	0,508	0,513	0,504	0,506
$n_{max} = 2$, *DF=2, T*	0,515	0,503	0,493	0,482	0,493	0,489
$n_{max} = 2$, *DF=2, D+T*	0,515	**0.517**	0,509	0,505	0,510	0,509
$n_{max} = 2$, *DF=3, B*	0,491	0,488	0,478	0,475	0,481	0,475
$n_{max} = 2$, *DF=3, D*	0,491	0,493	0,496	0,489	0,489	0,491
$n_{max} = 2$, *DF=3, T*	0,496	0,502	0,481	0,484	0,492	0,485
$n_{max} = 2$, *DF=3, D+T*	0,496	0,504	0,498	0,499	0,494	0,491
$n_{max} = 3$, *DF=1, B*	0,509	0,493	0,488	0,477	0,470	0,474
$n_{max} = 3$, *DF=1, D*	0,509	0,505	0,503	0,504	0,506	**0.510**
$n_{max} = 3$, *DF=1, T*	0,502	0,488	0,467	0,463	0,457	0,458
$n_{max} = 3$, *DF=1, D+T*	0,502	0,497	0,495	0,488	0,486	0,489
$n_{max} = 3$, *DF=2, B*	0,495	0,488	0,494	0,487	0,471	0,469
$n_{max} = 3$, *DF=2, D*	0,495	0,500	0,503	0,499	0,497	0,503
$n_{max} = 3$, *DF=2, T*	0,507	0,494	0,478	0,472	0,472	0,474
$n_{max} = 3$, *DF=2, D+T*	0,507	0,505	0,501	0,493	0,501	0,497
$n_{max} = 3$, *DF=3, B*	0,486	0,482	0,487	0,481	0,469	0,466
$n_{max} = 3$, *DF=3, D*	0,486	0,492	0,488	0,487	0,485	0,491
$n_{max} = 3$, *DF=3, T*	0,499	0,491	0,479	0,477	0,476	0,468
$n_{max} = 3$, *DF=3, D+T*	0,499	0,493	0,489	0,491	0,488	0,486

Tabla 3: Evaluación realizada con la unión de los conjuntos de datos `dev` y `train` de la tarea `1.1` del TASS 2020 en castellano (`es`) para diferentes valores de n_{max} y k_{max}, utilizando la medida $F1$ con promediado *macro*, donde B indica que se ha realizado un pesado binario, D que se ha utilizado el pesado por densidad, T que se ha realizado un pesado por *tf-idf*, $D + T$ que se ha realizado un pesado por densidad seguido de un pesado por *tf-idf*, y $DF = 1$, $DF = 2$ y $DF = 3$ que se ha utilizado un criterio de filtrado de términos con ocurrencia mayor o igual a 1, 2 y 3 respectivamente.

utilizar skipgrams de manera efectiva. Destacar que el pesado por *tf-idf* por sí solo ha obtenido buenos resultados, pero que se han visto mejorados al añadir el pesado por densidad, que parece ser la mejor combinación para pesar todo tipo de términos.

Respecto al filtrado, en este conjunto de datos no podemos deducir si, además de mejorar la eficiencia reduciendo los recursos espaciales y temporales, se mejora también la efectividad de los modelos en análisis de sentimientos. Pero a pesar de ser reducciones tan agresivas (que pueden llegar al orden del 90 % como hemos visto en la Sección 3.1), los resultados son similares y en algunos casos pueden conseguirse mejoras significativas.

Con el fin de poner en contexto esta aproximación con el estado de la cuestión, comparamos la mejor de nuestras configuraciones (*DF=2, D+T*) con los resultados oficiales de la competición del TASS 2020, que se sitúan en el rango $[0{,}37, 0{,}67]$. Sin utilizar conocimiento externo, utilizando como entrenamiento los conjuntos `dev` y `train` provistos por la competición, y evaluando con el conjunto de `test`, para castellano (`es`), obtenemos una puntuación en el rango $[0{,}505, 0{,}522]$ (según diferentes valores de n_{max} y k_{max}), que nos situaría en una posición intermedia, superando a algunas aproximaciones que utilizan técnicas como *word embeddings* o *neural networks*. Si aumentamos el entrenamiento utilizando los conjuntos `dev` y `train` de todos los idiomas de la competición (variantes del español), para poder aumentar nuestro vocabulario, y evaluamos igualmente solo con el conjunto de `test` en castellano (`es`), la puntuación aumentaría hasta el ran-

go [0,531, 0,554], lo que nos dejaría cerca de algunas de las aproximaciones que utilizan *deep learning* o *transformers*. Es cierto que nuestros resultados se han obtenido una vez terminada la competición, pero se ha intentado simular el contexto para poder comprobar en qué situación quedaría nuestra propuesta respecto a técnicas más recientes.

La siguiente experimentación la realizaremos en el conjunto de datos de Movie Reviews (también mencionado en la sección anterior) igualmente para diferentes valores de n_{max} y k_{max}. En la Tabla 4 se pueden observar los resultados con la misma nomenclatura que en los experimentos anteriores.

Los resultados de esta experimentación nos ofrecen una visión similar sobre el pesado por densidad. Como en el conjunto de datos anterior, nuestra propuesta de pesado mejora el rendimiento de los skipgrams con cualquier configuración ($D > B$, $D + T > T$) menos en el caso de n-gramas. La mejora media utilizando el pesado por densidad es de un 1,65 %, llegando al 5,65 % en el mejor caso. A pesar de que el conjunto de datos tiene más documentos, más palabras diferentes, un tamaño medio mayor, y está en un idioma diferente, la información sobre la distancia en los skipgrams sigue siendo valiosa a la hora de pesar los skipgrams en análisis de sentimientos. Igualmente, sin el pesado por densidad los skipgrams no ofrecen mejoras respecto a n-gramas.

5 Conclusiones y trabajo futuro

En este trabajo hemos realizado dos propuestas para mejorar el uso del modelado de skipgrams. Por un lado, una técnica para la generación y filtrado progresivos, con el objetivo de reducir el número de términos que se generan a la hora de extraer términos automáticamente de un conjunto de datos. Hemos aplicado esta técnica en dos corpus diferentes y hemos visto una reducción significativa de términos generados, que depende del conjunto de datos utilizado, pero que es considerablemente mayor cuanto más términos y omisiones se realizan, precisamente los casos en los que es más necesaria por la gran cantidad que se genera. Esta reducción puede llegar a un 95 %, pero viendo la tendencia creemos que puede ser mucho mayor si realizamos la experimentación con mayores valores de n_{max} y k_{max}. En la experimentación posterior hemos observado que para tareas como el análisis de sentimientos, este filtrado no reduce el rendimiento sino que en algunos casos los puede incluso mejorar.

Por otro lado, hemos propuesto un esquema de pesado de términos que tiene en cuenta el número de omisiones realizadas al generar skipgrams. Se ha observado que el uso de skipgrams en análisis de sentimientos necesita de este tipo de pesado, ya que en caso contrario los resultados empeoran al utilizar skipgrams. Pero cuando se usa el pesado por densidad, los resultados pueden mejorar hasta un 7 % en los mejores casos. Además, mencionar que el esquema propuesto combinado con el clásico tf-idf es lo que mejores resultados ofrece en el contexto estudiado.

Este trabajo abre la puerta a nuevas investigaciones relacionadas, entre las que podemos destacar:

- Buscar y diseñar nuevos tipos de filtrado progresivo eficientes que sean capaces de eliminar los términos menos relevantes pero manteniendo la efectividad de los sistemas que los utilizan, empezando por técnicas de *selección de características* y *reducción de la dimensionalidad*.
- Estudiar esquemas de pesado existentes o diseñar nuevos, y probar diferentes combinaciones entre ellos para ver si pueden mejorar los resultados.
- Aplicar el modelado de skipgrams en diferentes conjuntos de datos, géneros textuales, idiomas e incluso tareas de PLN.
- Analizar las curvas de aprendizaje de diferentes modelos utilizando skipgrams, para comprobar en qué medida pueden ampliar la cobertura de los sistemas.
- Utilizar los términos generados para crear un lexicón o diccionario de sentimientos que pueda ser utilizado como recurso para otras herramientas.

Agradecimientos

Esta investigación ha sido financiada por la Universidad de Alicante, el Ministerio de Ciencia e Innovación de España, la Generalitat Valenciana y el Fondo Europeo de Desarrollo Regional (FEDER) a través de la siguiente financiación: a nivel nacional, se concedieron los proyectos *TRIVIAL* (PID2021-122263OB-C22), *SocialTrust* (PDC2022-133146-C22) y *CLEARTEXT* (TED2021-130707B-I00), financiados

$k_{max} \rightarrow$	0	1	2	3	4	5
$n_{max} = 2$, *DF=1, B*	0,774	0,774	0,769	0,762	0,759	0,757
$n_{max} = 2$, *DF=1, D*	0,774	0,774	0,775	0,774	0,773	0,777
$n_{max} = 2$, *DF=1, T*	0,784	0,783	0,778	0,774	0,773	0,771
$n_{max} = 2$, *DF=1, D+T*	0,784	**0.789**	0,787	0,787	0,787	0,787
$n_{max} = 2$, *DF=2, B*	0,760	0,762	0,758	0,756	0,754	0,751
$n_{max} = 2$, *DF=2, D*	0,760	0,765	0,766	0,767	0,765	0,764
$n_{max} = 2$, *DF=2, T*	0,777	0,778	0,776	0,772	0,772	0,770
$n_{max} = 2$, *DF=2, D+T*	0,777	0,781	0,784	0,782	0,785	0,786
$n_{max} = 2$, *DF=3, B*	0,750	0,751	0,751	0,748	0,745	0,744
$n_{max} = 2$, *DF=3, D*	0,750	0,755	0,757	0,758	0,756	0,757
$n_{max} = 2$, *DF=3, T*	0,771	0,771	0,771	0,768	0,77	0,768
$n_{max} = 2$, *DF=3, D+T*	0,771	0,775	0,775	0,778	0,78	0,781
$n_{max} = 3$, *DF=1, B*	0,773	0,764	0,747	0,737	0,730	0,726
$n_{max} = 3$, *DF=1, D*	0,773	0,771	0,772	0,769	0,768	0,767
$n_{max} = 3$, *DF=1, T*	0,777	0,768	0,761	0,753	0,751	0,745
$n_{max} = 3$, *DF=1, D+T*	0,777	0,777	0,777	0,775	0,776	0,776
$n_{max} = 3$, *DF=2, B*	0,761	0,759	0,753	0,754	0,748	0,750
$n_{max} = 3$, *DF=2, D*	0,761	0,764	0,764	0,765	0,765	0,766
$n_{max} = 3$, *DF=2, T*	0,780	0,776	0,776	0,772	0,772	0,768
$n_{max} = 3$, *DF=2, D+T*	0,780	0,783	**0.785**	0,783	0,784	**0.785**
$n_{max} = 3$, *DF=3, B*	0,751	0,747	0,748	0,745	0,743	0,744
$n_{max} = 3$, *DF=3, D*	0,751	0,753	0,755	0,757	0,757	0,756
$n_{max} = 3$, *DF=3, T*	0,770	0,771	0,767	0,765	0,768	0,766
$n_{max} = 3$, *DF=3, D+T*	0,770	0,774	0,774	0,776	0,780	0,779

Tabla 4: Evaluación realizada en el conjunto de datos `sentence polarity dataset v1.0` de Movie Reviews en inglés para diferentes valores de n_{max} y k_{max}, utilizando la medida $F1$ con promediado *macro*, donde B indica que se ha realizado un pesado binario, D que se ha utilizado el pesado por densidad, T que se ha realizado un pesado por *tf-idf*, $D + T$ que se ha realizado un pesado por densidad seguido de un pesado por *tf-idf*, y $DF = 1$, $DF = 2$ y $DF = 3$ que se ha utilizado un criterio de filtrado de términos con ocurrencia mayor o igual a 1, 2 y 3 respectivamente.

por MCIN/AEI/10.13039/501100011033 y European Union NextGenerationEU/PRTR; a nivel regional, la Generalitat Valenciana (Conselleria d'Educació, Investigació, Cultura i Esport), concedió financiación para NL4DISMIS (CIPROM/2021/21). Además, contó con el apoyo de dos acciones COST: CA19134 - "Distributed Knowledge Graphs" y CA19142 - "Leading Platform for European Citizens, Industries, Academia, and Policymakers in Media Accessibility".

Bibliografía

Chang, C.-Y., S.-J. Lee, y C.-C. Lai. 2017. Weighted word2vec based on the distance of words. En *2017 International Conference on Machine Learning and Cybernetics (ICMLC)*, volumen 2, páginas 563–568. IEEE.

Church, K. W. 2017. Word2vec. *Natural Language Engineering*, 23(1):155–162.

Du, X., J. Yan, R. Zhang, y H. Zha. 2020. Cross-network skip-gram embedding for joint network alignment and link prediction. *IEEE Transactions on Knowledge and Data Engineering*.

Gompel, M. v. y A. van den Bosch. 2016. Efficient n-gram, skipgram and flexgram modelling with colibri core. *Journal of Open Research Software*, 4:1–10.

Guthrie, D., B. Allison, W. Liu, L. Guthrie, y Y. Wilks. 2006. A closer look at skip-gram modelling. En *Proceedings of the fifth international conference on language resources and evaluation (LREC'06)*.

Hossny, A. H., L. Mitchell, N. Lothian, y

G. Osborne. 2020. Feature selection methods for event detection in twitter: A text mining approach. *Social Network Analysis and Mining*, 10(1):1–15.

Komninos, A. y S. Manandhar. 2016. Dependency based embeddings for sentence classification tasks. En *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, páginas 1490–1500.

Mikolov, T., K. Chen, G. Corrado, y J. Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mimno, D. y L. Thompson. 2017. The strange geometry of skip-gram with negative sampling. En *Empirical Methods in Natural Language Processing*.

Nguyen, T. H. y R. Grishman. 2016. Modeling skip-grams for event detection with convolutional neural networks. En *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, páginas 886–891.

Pang, B. y L. Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. En *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, páginas 115–124.

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, y E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peng, H., J. Li, H. Yan, Q. Gong, S. Wang, L. Liu, L. Wang, y X. Ren. 2020. Dynamic network embedding via incremental skip-gram with negative sampling. *Science China Information Sciences*, 63(10):1–19.

Santos, F. A. O., T. D. Bispo, H. T. Macedo, y C. Zanchettin. 2021. Morphological skip-gram: Replacing fasttext characters n-gram with morphological knowledge. *Inteligencia Artificial*, 24(67):1–17.

Shazeer, N., J. Pelemans, y C. Chelba. 2015. Sparse non-negative matrix language modeling for skip-grams. *Proceedings Interspeech 2015*, 2015:1428–1432.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, y I. Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vega, M. G., M. C. Díaz-Galiano, M. Á. G. Cumbreras, F. M. P. del Arco, A. Montejo-Ráez, S. M. J. Zafra, E. M. Cámara, C. A. Aguilar, M. A. S. Cabezudo, L. Chiruzzo, y others. 2020. Overview of tass 2020: Introducing emotion detection. En *IberLEF@ SEPLN*.

Yadav, B. P., S. Ghate, A. Harshavardhan, G. Jhansi, K. S. Kumar, y E. Sudarshan. 2020. Text categorization performance examination using machine learning algorithms. En *IOP Conference Series: Materials Science and Engineering*, volumen 981, página 022044. IOP Publishing.

Zhao, Z., T. Liu, S. Li, B. Li, y X. Du. 2017. Ngram2vec: Learning improved word representations from ngram co-occurrence statistics. En *Proceedings of the 2017 conference on empirical methods in natural language processing*, páginas 244–253.

Tesis

Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 227-230 recibido 14-11-2022 revisado 01-02-2023 aceptado 11-02-2023

Linguistic features integration for text classification tasks in Spanish

Integración de características lingüísticas para tareas de clasificación de texto en español

José Antonio García-Díaz
Facultad de Informática, Universidad de Murcia
joseantonio.garcia8@um.es

Abstract: This manuscript summarises the doctoral thesis of José Antonio García-Díaz at the University of Murcia, under the supervision of doctors Rafael Valencia-García and Pedro José Vivancos-Vicente. This doctoral thesis is published by compendium of publications under the industrial doctorate modality. The act of defence took place on Tuesday, July 5, 2022, before the court composed of doctors Salud María Jiménez-Zafra, from the University of Jaén; Miguel Ángel Rodríguez-García, from the Rey Juan Carlos University; and Mª del Pilar Salas-Zarate, from the National Technological Institute of Mexico. The qualification obtained was Outstanding Cum Laude unanimously. In addition, the mention of international doctorate was obtained.
Keywords: Linguistic features, feature integration, automatic document classification, natural language processing.

Resumen: Este documento resume la tesis doctoral por compendio de publicaciones de José Antonio García-Díaz en la Universidad de Murcia, bajo la supervisión de los doctores Rafael Valencia-García y Pedro José Vivancos-Vicente bajo la modalidad de doctorado industrial. El acto de defensa tuvo lugar el martes 5 de Julio de 2022 ante el tribunal compuesto por los doctores Salud María Jiménez-Zafra, de la Universidad de Jaén; Miguel Ángel Rodríguez-García de la Universidad Rey Juan Carlos; y Mª del Pilar Salas-Zarate, del Tecnológico Nacional de México. La calificación obtenida fue de Sobresaliente Cum Laude por unanimidad. Además, se obtuvo la mención de doctorado internacional.
Palabras clave: Características lingüísticas, integración de características, clasificación automática del texto, procesamiento del lenguaje natural.

1 Introduction

Natural Language Processing (NLP) is the branch of Artificial Intelligence (AI) and Linguistics that aims at easing the communication between computers and humans by means of human language.

The scope of this thesis is Automatic Document Classification (ADC), an NLP task which consists in assigning a set of predefined labels to a set of documents. ADC can be applied to Author Profiling (AP), Emotion Detection (ED), Sentiment Analysis (SA), or hate-speech detection among others. To do ADC, computers need practical ways to represent natural language. One of these ways is by means of Linguistic Features (LFs), which represent documents as a vector formed by the percentage of linguistically relevant traits, that indicate *what* a text says, and *how* it says it.

Two research hypotheses are raised in this thesis:

- **RH1**. The inclusion of a set of LFs that capture linguistic traits of the authors can improve the performance of ADC. We address this study in Spanish, including a wide variety of domains concerning infodemiology, hate-speech, humour, or irony among others.
- **RH2**. The inclusion of LFs improves the interpretability to the models with a fewer number of features that generalise better than systems built upon novel Language Models and Transformers.

To accomplish the hypotheses, we have

ISSN 1135-5948 DOI 10.26342/2023-70-18

obtained a taxonomy of LFs in Spanish, and we have developed two software tools: UMUTextStats and UMUCorpusClassifier. The validation of the research hypotheses has been conducted in several scenarios with the validation of the features and the compilation of several linguistic corpora in Spanish.

2 Structure and organisation

Chapter 1 details all the contributions derived from this work. Apart from the abstract, the introduction, its motivation and a state-of-the-art subsection with the methodologies and evaluation used, this chapter describes the system architecture of the two tools developed: UMUTextStats and UMUCorpusClassifier. Besides, it summarises the experimental results obtained during the validation of the tool, which have given rise to the publications that are presented by compendium and the participation in several international workshops.

Chapter 2 presents the research articles that are attached as the compendium of the doctoral thesis. These research articles are about: (1) an ontology-driven aspect-based sentiment analysis system, focused on infodemiolgy (García-Díaz, Cánovas-García, and Valencia-García, 2020); (2) the compilation and evaluation of the Spanish MisoCorpus 2020, focused on misogyny detection (García-Díaz et al., 2021); (3) the compilation process of the Spanish PoliCorpus 2020 and its evaluation with two author analysis tasks: an author profiling task to extract demographic and psychographic traits, and an authorship attribution task in order to obtain which the author of a set of anonymous documents (García-Díaz, Palacios, and Valencia-García, 2022); (4) and the compilation process of the Spanish SatiCorpus 2021, which includes satirical headlines and tweets from a wide variety of countries from Spain and Latin America newspapers (García-Díaz and Valencia-García, 2022).

Chapter 3 contains the conclusions, a summary of all the publications derived from this work, and a list of promising future research lines related to the LFs and ADC in Spanish.

3 Main contributions

The main contributions of this doctoral thesis are the UMUTextStats and UMUCorpusClassifier tools, and their validation in multiple scenarios. Accordingly, this section describes both tools and their validation.

3.1 Tools

3.1.1 UMUTextStats

UMUTextStats[1] (García-Díaz et al., 2022) is a tool for extracting LFs. This tool focuses for Spanish since it is one of the most used languages on the Internet. UMUTextStats is inspired in LIWC (Tausczik and Pennebaker, 2010). However, UMUTextStats solves some deficiencies identified in LIWC for Spanish (Garcí et al., 2007). For instance, LIWC does not capture inflection mechanisms that indicates the tense, mood, and the person to whom the verb refers in Spanish. In addition, LIWC is a commercial tool, and we aim to provide an open-source tool for the Spanish NLP community.

UMUTextStats captures a total of 365 LFs organised within the following taxonomy: (1) phonetics, (2) morphosyntax, (3) correction and style, (4) semantics, (5) pragmatics, (6) stylometry, (7) lexical, (8) psycho-linguistic processes, (9) register, and (10) social media.

3.1.2 UMUCorpusClassifier

The UMUCorpusClassifier tool[2] eases the compilation and annotation of linguistic corpora (García-Díaz et al., 2020), which is a very time-consuming task. Besides, the quality of manually annotated datasets is heavily influenced by disagreements between annotators. Therefore, the lack of supervision of the annotation process can lead to poor quality corpora.

The documents compiled from UMUCorpusClassifier can be classified using distant supervision or manual labelling. Besides, UMUCorpusClassifier allows to coordinate groups of annotators and measure their performance with several metrics concerning inter-annotator agreement.

3.2 Validation

3.2.1 Aspect-based Sentiment Analysis

We evaluate the LFs in an aspect-based SA study focused on infodemiology (García-Díaz, Cánovas-García, and Valencia-García, 2020). For this, a dataset from Twitter with short texts related to different infectious diseases was compiled. Once the dataset was

[1] https://umuteam.inf.um.es/umutextstats/

[2] https://umuteam.inf.um.es/corpusclassifier/

compiled, we extracted the LFs and used them to perform a multi-class SA, achieving an accuracy of 55.3% with the LFs. These results outperformed the rest of the features, which included non-contextual word embeddings trained with a convolutional or recurrent neural networks.

The aspects related to infodemiology were represented within a domain ontology, representing risks, symptoms, transmission methods or drugs related to infectious diseases. In this work, we assumed that one document contains only one sentiment. Accordingly, we ranked the relationship between the sentiment of the tweet with the ontology classes.

The interpretability of the resulting models was measured with the Information Gain of the LFs. We observed that numerals are correlated to negative documents and that the usage of colloquialism is more related to positive and neutral tweets than negative tweets.

Other validations in SA were our participation in TASS 2020 and EmoEvalEs shared tasks.

3.2.2 Hate-speech and misogyny detection

Our contributions regarding hate-speech started with the compilation and evaluation of the Spanish MisoCorpus 2020 (García-Díaz et al., 2021), which includes documents concerning violence against relevant women, messages written from Spain and Latin America, and general traits related to misogyny, such as discredit or dominance among others. The dataset is balanced, and it contains 3 841 misogynous documents. The best accuracy achieved was 85.175% with Support Vector Machine (SVM). Moreover, we observed that the combination of the LFs and the sentence embeddings outperformed the rest of the feature sets. This finding supports our first research hypothesis regarding the improvement of the results for ADC. As expected, we observed that LFs related to offensive language have a strong correlation for misogyny detection. A strong correlation between the grammatical gender and misogyny identification was also found. This is relevant because some words can be interpreted differently according to their gender. We also observed a strong correlation with correction and style features, such as the percentage of misspelled words.

In addition, we participated in the last two editions of EXIST (2021, 2022), and MeOffendES 2021.

3.2.3 Figurative language: Satire, Sarcasm, and Humor detection

We compile the Spanish SatiCorpus 2021 (García-Díaz and Valencia-García, 2022) to distinguish between satirical news and real news. The dataset is balanced and contains news headlines from Twitter. The accounts were selected from different Spanish spoken countries. Moreover, we decided to enlarge this dataset including tweets from Twitter accounts used for impersonating and satirise real relevant people. This dataset was automatically annotated, based on the idea that all tweets from satirical news media are satiric. We evaluated the LFs separately and combined with other types of features using different strategies. Our best result was achieved with a combination of the LFs and BERT with an accuracy of 97.405%. We observed that the number of orthographic errors is more common in non-satirical documents than in satirical documents. In contrast, the number of hashtags more commonly appears in non-satirical documents. Regarding morphological features, the use of pronouns and nouns is good for discerning between satirical and non-satirical documents, being the pronouns more frequently found in satirical documents whereas nouns are more common in non-satirical documents.

Besides, we participate in Hahackathon 2021 and HaHa 2021.

3.2.4 Author Analysis

We explored the reliability of the linguistic features in two experiments regarding author analysis: authorship attribution and profiling, and we annotated each user with their gender, year of birth, and political spectrum on two axes (binary and multiclass) (García-Díaz, Palacios, and Valencia-García, 2022).

Concerning the AP study, we evaluated the LFs with sentence and word embeddings from Word2Vec, FastText, or BERT. We observed that the LFs achieved promising results, outperforming BERT in some traits, such as gender. Moreover, the combination of the LFs with the rest of features usually results in better results than achieved with both features separately. Concerning the interpretability of the results, we observed that morphosyntax is the most relevant category for determining demographic traits. Besides,

we found correlations between the percentage of personal pronouns and verbs, the usage of colloquialisms, topics related to countries and languages.

4 Conclusions

In this doctoral thesis we have proven the effectiveness of the integration of LFs for conducting ADC (RH1) and their interpretability (RH2). Finally, it is worth mentioning that two software tools have been released to the Spanish research NLP community two software tools: UMUTextStats and UMUCorpusClassifier.

Acknowledgements

This PhD. Thesis has been partially supported by Banco Santander and the University of Murcia through the Doctorado Industrial programme, by the research project LaTe4PSP (PID2019-107652RB-I00/AEI/ 10.13039/501100011033) funded by MCIN/ AEI/10.13039/501100011033 and by the research project AIInFunds(PDC2021-121112-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR.

References

Garcí, F. A., J. W. Pennebaker, N. Ramírez-Esparza, and R. Suriá. 2007. La psicología del uso de las palabras: Un programa de computadora que analiza textos en español. *Revista mexicana de psicología*, 24(1):85–99.

García-Díaz, J. A., Á. Almela, G. Alcaraz-Mármol, and R. Valencia-García. 2020. Umucorpusclassifier: Compilation and evaluation of linguistic corpus for natural language processing tasks. *Procesamiento del Lenguaje Natural*, 65:139–142.

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

García-Díaz, J. A., M. Cánovas-García, and R. Valencia-García. 2020. Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in latin america. *Future Gener. Comput. Syst.*, 112:641–657.

García-Díaz, J. A., R. C. Palacios, and R. Valencia-García. 2022. Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020. *Future Gener. Comput. Syst.*, 130:59–74.

García-Díaz, J. A. and R. Valencia-García. 2022. Compilation and evaluation of the spanish saticorpus 2021 for satire identification using linguistic features and transformers. *Complex & Intelligent Systems*, 8(2):1723–1736.

García-Díaz, J. A., P. J. Vivancos-Vicente, A. Almela, and R. Valencia-García. 2022. Umutextstats: A linguistic feature extraction tool for spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6035–6044.

Tausczik, Y. R. and J. W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 231-234 recibido 12-12-2022 revisado 01-02-2023 aceptado 12-02-2023

Análisis y tipificación de errores lingüísticos para una propuesta de mejora de informes médicos en español

Analysis and classification of linguistic errors for a proposal to improve medical reports in Spanish

Jésica López Hernández
TECNOMOD, Departamento de Informática y Sistemas, Facultad de Informática, Campus de Espinardo, Universidad de Murcia, 30100, Murcia (España)
jesica.lopez@um.es

Resumen: Este es un resumen de la tesis doctoral realizada por Jésica López Hernández bajo la dirección del Dr. Rafael Valencia García, la Dra. Ángela Almela Sánchez-Lafuente y el Dr. Fernando Molina Molina. La defensa tuvo lugar el día 18 de mayo de 2022 en la Facultad de Letras de la Universidad de Murcia, con un tribunal compuesto por el Dr. Pascual Cantos Gómez (Universidad de Murcia), la Dra. Gema Alcaraz Mármol (Universidad de Castilla-La Mancha) y el Dr. Mario Andrés Paredes Valverde (Instituto Tecnológico Superior de Teziutlán). La tesis obtuvo la calificación de sobresaliente *cum laude* otorgada por unanimidad y mención de doctorado internacional.
Palabras clave: detección automática de errores, análisis de errores, informes médicos, lingüística computacional, procesamiento del lenguaje natural.

Abstract: This is a summary of the Ph.D. thesis written by Jésica López Hernández at University of Murcia under the supervision of Ph.D. Rafael Valencia García, Ph.D. Ángela Almela Sánchez-Lafuente and Ph.D. Fernando Molina Molina. The author was examined on May 18th, 2022 by a committee formed by Ph.D. Pascual Cantos Gómez (University of de Murcia), Ph.D. Gema Alcaraz Mármol (University of Castilla-La Mancha) and Ph.D. Mario Andrés Paredes Valverde (Higher Technological Institute of Teziutlán). The Ph.D. thesis was awarded an excellent grade and Cum Laude honours and the international mention.
Keywords: automatic error detection, error analysis, medical reports, computational linguistics, natural language processing.

1 Introducción

Uno de los principales campos de aplicación del procesamiento del lenguaje natural es el dominio biomédico. La documentación clínica contiene información de gran valor para la investigación y la práctica sanitaria, por tanto, resulta fundamental poder emplear en este campo tecnologías basadas en procesamiento automático de datos que permitan la extracción y clasificación de información, así como la anonimización de documentos clínicos, o la interoperabilidad semántica.

En el caso de los informes médicos, debido a sus características textuales y contextuales, la presencia de errores lingüísticos es común (Lai *et al.*, 2015), lo que dificulta su tratamiento automatizado. Como consecuencia, la corrección automática se convierte en un componente de gran importancia para el procesamiento de datos en informes médicos.

Los sistemas de corrección automática a la vanguardia, como las arquitecturas basadas en redes neuronales, necesitan grandes conjuntos de datos de entrenamiento para un rendimiento óptimo. Debido a la ausencia de corpus de dominio biomédico disponibles, ha ganado relevancia la recopilación y generación artificial de errores en corpus para el entrenamiento de estos sistemas. El desarrollo de una tipología de errores a partir del estudio empírico de un corpus de informes médicos va a permitir añadir nuevos patrones para la generación de errores de forma más exhaustiva y, con ello, la creación de modelos más robustos para el procesamiento de datos en medicina.

ISSN 1135-5948 DOI 10.26342/2023-70-19

Por tanto, el **propósito principal** de la tesis doctoral es la recopilación, clasificación y análisis de errores lingüísticos presentes en informes médicos en español. Mediante un estudio exploratorio con carácter descriptivo se pretende añadir otra capa de información a los métodos de detección y corrección automática disponibles para el dominio médico.

Este objetivo principal se desglosa a su vez en una serie de **objetivos específicos** entre los que se encuentran:

- Investigar sobre el estado actual del procesamiento del lenguaje natural en el dominio médico, así como la corrección automática, tanto en el ámbito general como en el dominio específico de la medicina.
- Compilar y preprocesar el corpus de estudio a partir de la recopilación de informes médicos digitalizados de varias especialidades médicas.
- Estudiar los principales métodos de detección y corrección de errores *non-word* y *real-word*. El error *non-word* genera una palabra incorrecta en el plano ortográfico; en cambio, el error *real-word* da lugar a una palabra existente y correcta idiomáticamente, pero errónea en el contexto, por tanto, este tipo de error se manifiesta en el plano semántico o sintáctico.
- Desarrollar una herramienta de cómputo y clasificación de errores.
- Identificar, analizar y clasificar de forma sistemática los errores lingüísticos presentes en informes médicos desde un enfoque cuantitativo y cualitativo.
- Comprobar si hay diferencias significativas entre las distintas especialidades y entre los errores presentes en el dominio médico y el español general.
- Contribuir a la creación de conjuntos de datos de entrenamiento más exhaustivos, que incorporen casuísticas de errores reales de informes médicos.

2 Estructura de la tesis

La tesis consta de una primera parte teórica dedicada a la investigación sobre el estado de la cuestión; y una segunda parte práctica, de carácter fundamentalmente descriptivo, que aborda el desarrollo metodológico y el análisis de los resultados. Estas dos partes se distribuyen en los siguientes capítulos:

En el **primer capítulo** se define el marco de referencia en el que se inserta la investigación y su finalidad. Por un lado, se exponen las razones que han motivado la realización de este trabajo y, por otro, se detalla la distribución de los distintos apartados que lo componen.

En el **segundo capítulo** se abordan los fundamentos teóricos y se documenta el estado de la cuestión en lo que respecta a los dos pilares que sustentan esta investigación: la corrección automática de errores y el lenguaje médico.

En el **tercer capítulo** se explica la propuesta metodológica empleada y los experimentos desarrollados. Se define el objetivo principal de la tesis y se formulan los objetivos específicos para dar respuesta al problema de investigación planteado. En segundo lugar, se presenta el corpus objeto de estudio, se proponen los criterios de análisis que se van a tener en cuenta y las distintas convenciones en cuanto al tratamiento de los datos. En la sección dedicada al procedimiento se describen las distintas fases del enfoque metodológico llevado a cabo, que incluye el preprocesamiento del corpus, la detección y corrección de errores *non-word* y *real-word* respectivamente, y el cómputo y clasificación de los errores detectados.

El **cuarto capítulo** comprende el análisis de datos a partir de los resultados obtenidos. Se realiza un análisis cuantitativo teniendo en cuenta la frecuencia, la distancia de edición, el tipo (omisión, sustitución, inserción y transposición) y subtipo de error, y la posición del error. Por su parte, en el análisis cualitativo se realiza un desglose pormenorizado de los distintos tipos de patrones de errores detectados, mencionando otros aspectos lingüísticos que pueden ser de utilidad para la finalidad del estudio.

El **quinto capítulo** incluye las conclusiones obtenidas, así como las limitaciones y desafíos presentes en la investigación y, por último, las sugerencias de líneas de trabajo futuras que servirán para mejorar y ampliar los resultados.

Finalmente, el **último capítulo** presenta las principales aportaciones científicas derivadas de esta tesis doctoral, incluyendo los artículos de investigación, los capítulos de libro y las comunicaciones en congresos.

3 Contribuciones más importantes

A continuación, se mencionan las principales contribuciones que emanan de la tesis:

Estado de la cuestión. Se ha aportado una revisión bibliográfica (López-Hernández, Almela y Valencia-García, 2019) en torno a la corrección automática y al análisis de errores en el ámbito biosanitario, que ha incluido el estudio de las principales técnicas de detección y recursos utilizados en el área. El fin principal fue conocer los desafíos y limitaciones actuales que presentaba la corrección automática específicamente en el lenguaje médico y proporcionar una síntesis de todas las investigaciones relevantes hasta la fecha.

Corpus. El corpus objeto de estudio está constituido por una recopilación de informes médicos electrónicos en español pertenecientes a las especialidades médicas de urgencias, unidad de cuidados intensivos (UCI), psiquiatría y cirugía general. El corpus, que está anonimizado, contiene un total de 2 321 826 *tokens* y ha sido sometido a un preprocesamiento y normalización para facilitar su tratamiento y análisis. Es un corpus privado, perteneciente a la empresa Vócali (https://vocali.net/), por lo que no puede ser distribuido.

Sistema. Se ha desarrollado un sistema para la detección y corrección de errores *non-word* (ortográficos), que incluye la comparación con un lexicón, la técnica de distancia de edición mínima para la generación de candidatos y la revisión manual asistida. Posteriormente, se ha trabajado en la detección de errores *real-word* (plano semántico o sintáctico). Para ello, se ha llevado a cabo la generación de modelos lingüísticos, la representación vectorial de las palabras del corpus a partir de Word2Vec y el etiquetado gramatical del corpus.

Por último, se ha desarrollado una herramienta de cómputo y clasificación con la que se ha efectuado una categorización sistemática de los errores detectados, junto con la creación de categorías adaptadas para estos errores. Esta herramienta permite la generación de matrices de confusión, que muestran qué carácter es sustituido por otro y con qué frecuencia. Como resultado, se han identificado los errores de sustitución más comunes y las combinaciones de caracteres involucradas.

Análisis. El análisis de resultados ha permitido recopilar los tipos de errores más frecuentes, conocer si existen diferencias significativas en los resultados de las especialidades médicas analizadas o si hay diferencias entre los errores detectados en el dominio médico y la tipificación existente sobre errores del español no especializado.

- **Análisis cuantitativo**: Se han detectado un total de 76 711 errores en un corpus formado por 2 321 826 *tokens*, lo que supone una tasa de error del 3,3 %. La especialidad con un porcentaje de errores más alto es urgencias. Los resultados indican que el tipo de error que ocurre con un porcentaje mayor en todas las especialidades es el de omisión de tilde y, en segundo lugar, el de omisión de letra, y la mayor parte de los errores se producen a distancia de edición 1. La mayoría de los errores se concentran en un número limitado de pares de caracteres. Entre ellos, destacan los pares de caracteres que generan confusión por su similitud fonética y por el desconocimiento de las normas académicas que regulan su uso. Por último, se observan casos cuyo error es motivado por las posiciones adyacentes de estas letras en el teclado y que evidencian que son errores de actuación o de tipo mecánico.

  Estos resultados se reflejan en López-Hernández, Almela y Valencia-García (2021), donde se muestra una primera aproximación, con un análisis de errores *non-word* en informes de la especialidad de urgencias; y en López-Hernández y Almela (2021), que presenta los resultados tras ampliar la variabilidad del corpus e incorporar informes de UCI, cirugía general y psiquiatría, aportando un análisis cuantitativo de los tipos de errores detectados.
- **Análisis cualitativo**: Se ha realizado una catalogación y descripción cualitativa de los errores detectados, que incluye una explicación de las posibles causas de aparición. En esta sección se incluye toda aquella información lingüística complementaria que puede ser útil para el desarrollo de un módulo basado en conocimiento lingüístico y el tratamiento automatizado de los errores.

  Entre los patrones de errores *non-word* detectados destacan: el uso erróneo de tildes, la formación errónea de palabras mediante derivación y composición, la escritura errónea de extranjerismos y

nombres propios, la simplificación de grupos consonánticos, la representación gráfica de fonemas errónea, la analogía con otras formas, el uso equivocado de minúsculas y mayúsculas, la creación y uso incorrecto de abreviaturas, y el tratamiento erróneo de siglas y símbolos. En el caso de los errores *real-word*, se detectan errores de paronimia, de ausencia de concordancia gramatical, de formación errónea de palabras por fenómenos de composición y prefijación, y la presencia de formas verbales anómalas en el dominio.

En Bravo-Candel *et al.* (2021) se introducen errores sintéticos en un corpus mediante reglas, para el entrenamiento de un modelo de traducción automática neuronal *Seq2seq*. En esta publicación se utilizaron dos corpus para entrenar y probar el sistema: un corpus general con 611 millones de palabras extraídas de artículos de Wikipedia en español, y un corpus de casos clínicos recopilados a partir de tres fuentes diferentes (CodiEsp, MEDDOCAN, SPACCC) y compuesto por aproximadamente 2 millones de palabras. En López-Hernández, Molina-Molina y Almela (2022) se presentan los resultados tras haber llevado a cabo la identificación, análisis y clasificación sistemática de errores *real-word* en el corpus estudiado.

Módulo basado en conocimiento lingüístico. Incorporamos un listado con los patrones detectados y la información recopilada, que puede emplearse en distintas partes del proceso de detección y corrección automática. Esta información es utilizada para desarrollar nuevas reglas formalizables por el sistema que imiten los errores detectados y, con ello, se contribuye a la creación de conjuntos de datos sintéticos para entrenamiento. En el caso de los errores *real-word* es especialmente interesante contar con este repertorio, pues son errores que suelen pasar desapercibidos en los procesos de detección. Por tanto, al entrenar el sistema para que aprenda de la casuística de errores que hemos recopilado, este será más robusto.

Además de en la fase de generación de errores para el aumento de datos de entrenamiento, la recopilación de los fenómenos que más frecuentemente constituyen errores permite aportar información en la arquitectura de decisión y ponderación de alternativas de corrección.

Agradecimientos

Esta tesis doctoral ha sido financiada por el Ministerio de Educación, Cultura y Deporte de España a través de las Ayudas para la formación de profesorado universitario (FPU), del Programa Estatal de Promoción del Talento y su Empleabilidad, con referencia FPU16/03324. También ha sido apoyada por la Agencia Estatal de Investigación (AEI) a través del proyecto LaTe4PSP (PID2019-107652RB-I00/AEI/10.13039/501100011033).

Bibliografía

Bravo-Candel, D., J. López-Hernández, J. A. García-Díaz, F. Molina-Molina y F. García-Sánchez. 2021. Automatic correction of real-word errors in Spanish clinical texts. *Sensors*, 21(9):2893.

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

López-Hernández, J. y Á. Almela. 2021. Detección automática de errores lingüísticos en textos clínicos: análisis de patrones de error en varias especialidades médicas. *Panace@. Revista de medicina, lenguaje y traducción*, 22(53):96-108.

López-Hernández, J., Á. Almela y R. Valencia-García. 2019. Automatic spelling detection and correction in the medical domain: A systematic literature review. En *Technologies and Innovation. CITI 2019. Communications in Computer and Information Science* (vol. 1124, pp. 104-117). Cham: Springer.

López-Hernández, J., Á. Almela y R. Valencia-García. 2021. Linguistic errors in the biomedical domain: Towards an error typology for Spanish. *Sintagma: revista de lingüística*, 33:83-100.

López-Hernández, J., F. Molina-Molina y Á. Almela. 2022. Analysis of context-dependent errors in the medical domain in Spanish: a corpus-based study. *Sage Open*, 13(1).

Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 235-238 recibido 16-12-2022 revisado 01-02-2023 aceptado 12-02-2023

Machine Learning approaches for Topic and Sentiment Analysis in multilingual opinions and low-resource languages: From English to Guarani

Enfoques de aprendizaje automático para el análisis de sentimientos y temas en opiniones multilingües y en idiomas con escasez de recursos: Del inglés al guaraní

Marvin Matías Agüero-Torales
University of Granada, Spain
maguero@correo.ugr.es

Abstract: The following is a summary of a Ph.D. thesis written by Marvin Matías Agüero-Torales at the University of Granada under the supervision of Ph.D. Antonio Gabriel López-Herrera. The author was examined on Friday, February 4th, 2022 by a committee composed of Ph.D. Enrique Herrera-Viedma and Ph.D. Carlos Gustavo Porcel Gallego from the University of Granada, Ph.D. María José del Jesús Díaz and Ph.D. Salud María Jiménez-Zafra from the University of Jaén, and Ph.D. Jesús Serrano-Guerrero from the University of Castilla-La Mancha. The Ph.D. thesis was awarded the Summa Cum Laude mention.
Keywords: natural language processing (NLP), machine learning, code-switching, low-resource languages.

Resumen: Este es un resumen de la tesis doctoral realizada por Marvin Matías Agüero-Torales en la Universidad de Granada bajo la dirección del doctor D. Antonio Gabriel López-Herrera. La defensa de la tesis se llevó a cabo el viernes 4 de febrero de 2022 ante un tribunal formado por los doctores D. Enrique Herrera-Viedma y D. Carlos Gustavo Porcel Gallego de la Universidad de Granada, Dña. María José del Jesús Díaz y Dña. Salud María Jiménez-Zafra de la Universidad de Jaén, y D. Jesús Serrano-Guerrero de la Universidad de Castilla-La Mancha. La tesis recibió la calificación de Sobresaliente Cum Laude por unanimidad.
Palabras clave: procesamiento de lenguaje natural (PLN), aprendizaje automático, code-switching, idiomas con escasez de recursos.

1 Introduction

In recent years, the internet, especially social media, has become the main source of information, with people sharing their opinions, beliefs, emotions, and experiences online. Researchers from various fields, particularly Natural Language Processing (NLP), have been interested in analyzing this web content. NLP involves using computational techniques to analyze and synthesize natural language, especially text found on the web.

This doctoral thesis proposes using machine learning techniques to analyze opinions in low-resource languages, including Spanish, Guarani, and Jopara, in monolingual, multilingual, and code-switching settings. These techniques include sentiment analysis, which aims to identify the sentiment expressed in a text, and topic modeling, which aims to identify the main topics in a collection of texts.

In this thesis, we followed the path of text mining and NLP at the intersection of computation, artificial intelligence, and computational linguistics, focusing on multilingualism in low-resource languages. Our objectives are to (i) investigate different machine learning approaches that can handle multilingual opinions written in social media (even code-switching), particularly those based on neural networks, (ii) create new linguistic resources for analyzing text in low-resource languages and dialects, particularly those found on social media, and (iii) develop machine learning models for NLP in low-resource languages and dialects in monolingual, multilingual, and code-switching settings. The research aims to gain a deeper understanding

ISSN 1135-5948 DOI 10.26342/2023-70-20

of this problem and provide a comprehensive analysis.

There have been relatively few studies on tasks involving multilingualism with truly low-resource languages (such as Guarani/Jopara) or specific dialects (such as Spanish from Spain) in the literature. This is likely because most research in this area has focused on languages with more resources available, such as a sufficient number of Wikipedia or Common Crawl pages. It is important to carefully study the behavior of state-of-the-art machine learning models, as well as traditional models, to determine which is best suited for addressing the problem of multilingualism, particularly in low-resource languages, and under what conditions.

This thesis may be beneficial to both Spanish-speaking communities (especially in Spain) and Guarani-speaking communities (in Paraguay and surrounding countries such as Argentina, Bolivia, and Brazil) because most NLP systems are designed for use with rich-resource languages. The approaches presented in this work could be applied in various fields and disciplines, including marketing, psychology, sociology, politics, tourism, health informatics, and more, in order to extract insights from written opinions in these languages in various dimensions (such as sentiments, affections, and language type). It is important to have adequate resources for accurately and fairly analyzing written opinions in these languages.

2 Structure

This thesis consists of eight chapters and three appendices, which are described below.

Chapter 1 Introduces the research being conducted, its background and context, as well as the motivation, objectives, and methods of the research.

Chapter 2 Aims to present a thorough review of the use of deep learning to address the problem of multilingual sentiment analysis in social media to the research community. It provides a comprehensive overview of the field and highlights common ideas and issues that have been addressed in the implementation of multilingual sentiment analysis. It also offers a clear summary and discussion to identify potential areas for further research. This chapter is an expansion of a paper published in the journal *Applied Soft Computing* in the special issue 'Soft Computing for Recommender Systems and Sentiment Analysis' (Agüero-Torales, Abreu Salas, and López-Herrera, 2021).

Chapter 3 Focuses on the creation of corpora for low-resource languages and code-switched languages and is divided into the following sections: (i) collection of Spanish COVID-19-related tweets using keywords, language identification tools, and geolocated data for Spanish cities and regions; (ii) collection of Guarani-Spanish (also known as Jopara) Twitter text data for sentiment analysis, which includes challenges such as unbalanced classes due to the limited number of tweets written in Guarani-dominant; (iii) collection of three new, multi-annotated corpora of Jopara Guarani-dominant tweets for affect detection: (a) emotion recognition, (b) humor detection, and (c) identification of offensive and toxic language. The content of this chapter has been adapted from papers published in *Procesamiento Del Lenguaje Natural* (Agüero-Torales, Vilares, and López-Herrera, 2021) and CALCS 2021 (co-located with NAACL 2021) (Agüero-Torales, Vilares, and López-Herrera, 2021), as well as a work submitted to a journal.

Chapter 4 We used NLP techniques to study the discussions taking place on Twitter in Spain at the start of the COVID-19 pandemic. We analyzed the tweets and tracked the evolution of the topics by comparing them to newspaper articles. We also developed a small evaluation framework that involved human judgment. We used both a generative approach and a discriminative approach, which involves identifying the most important keywords and phrases, to represent the topics. The results of this research have been published in the journal *Procesamiento Del Lenguaje Natural* (Agüero-Torales, Vilares, and López-Herrera, 2021).

Chapter 5 Here, various machine learning methods, ranging from traditional approaches to more advanced transformer-based techniques, were applied to the low-resource language Guarani and to a combination of Guarani and Spanish (called Jopara). The performance of the different models was compared and error analysis was conducted to gain further insight into the classifiers' performance in this particular low-resource set-

ting. This chapter is an extension of a previously published paper in CALCS 2021 (co-located with NAACL 2021) (Agüero-Torales, Vilares, and López-Herrera, 2021).

Chapter 6 We describe our efforts to build and pre-train transformer-based language models (Vaswani et al., 2017) using Wikipedia data in the low-resource language Guarani, which faces challenges due to the presence of code-switching. We present a summary of the approaches we took to train a set of BERT models (Devlin et al., 2019) for Guarani and Jopara and evaluate them on tasks related to sentiment analysis. These models overall outperformed the mBERT (multilingual BERT) and Spanish BERT (Cañete et al., 2020, BETO), which do not include Guarani during pre-training, on tasks such as, (i) emotion recognition, (ii) humor detection, (iii) identification of offensive language, and (iv) polarity classification in F1-score and accuracy metrics.

Chapter 7 Summarizes the publications, contributions, and findings of the thesis.

Chapter 8 Presents the conclusions of the thesis. Additionally, we suggest potential areas for further investigation based on the results we have obtained.

Appendices *Appendix A* contains the publications that are part of the thesis. *Appendix B* explains the quantitative analysis for the topic modeling discussed in Chapter 4 and provides information about the annotation guidelines for the Guarani-dominant Jopara corpora used in Chapters 5 and 6. *Appendix C* provides details about the implementation and hyperparameter optimization of the machine learning models used in Chapters 5 and 6, as well as information about the scraped Twitter accounts mentioned in Chapter 3.

3 Contributions

This section provides an overview of the key findings, results, and contributions of the thesis.

3.1 Software prototype

`Gastro-miner`[1] is a cloud-based tool that allows the analysis of users' reviews and opinions written in English about restaurants on social media platforms such as TripAdvisor.com. It allows the collection, storage, cleaning, preprocessing, sentiment analysis, and visualization of review data. The tool was developed using Python, including Scrapy for web scraping, NLTK for NLP, Matplotlib for data visualization, and Django as a web framework. The tool was implemented using virtualization technology such as Vagrant, VirtualBox, and the Docker stack, and data was stored using MongoDB. The sentiment analysis stage used the VADER tool.[2] `Gastro-miner` can be customized for use on other social media platforms, languages, or settings, and the methodology and architecture of the tool are considered a contribution of the thesis.

The results of this study have been published at *Procedia Computer Science* (Agüero-Torales et al., 2019) and were presented at the *Proceedings of the ITQM 2019* and as a poster (Agüero-Torales, López-Herrera, and Cobo, 2018) at the 'Jornadas Científicas de Ciencia de Datos' organized by the *Universidad Comunera* (Asunción, Paraguay), and was awarded first place in the *i-Data* applied data science contest.

3.2 Contributions and resources

The main contributions and resources that have been made available to the research community are listed below:

1. Several corpora for low-resource languages: (a) an unlabeled Spanish Twitter corpus ($\sim$ 1M) about the COVID-19 pandemic outbreak,[3] (b) the first Guarani-dominant Jopara corpus (3,491 tweets) for sentiment analysis,[4] annotated according to a trinary scale (positive, negative, neutral), and (c) the first Guarani-dominant Jopara text-based dataset (2,364 tweets) for affect detection;[5] which includes three multi-annotated corpora: (i) emotion recognition annotated according to four mood categories (happy, sad, angry, other), (ii) humor detection, and (iii) identification of offensive and toxic language.

2. A small evaluation framework with a small guide, that outlines the process followed by native Spanish-speaking anno-

[1] `https://github.com/mmaguero/cloud-based-tool-SA`

[2] `https://github.com/cjhutto/vaderSentiment`

[3] `https://doi.org/10.7910/DVN/6PPSAZ`

[4] `https://doi.org/10.7910/DVN/GLDX14`

[5] `https://github.com/mmaguero/guarani-multi-affective-analysis`

tators to evaluate a sample of the topics discovered in Chapter 4.

3. An annotation mini-guidelines document that outlines the process followed by the bilingual annotators (Guarani-Spanish) as they manually annotated the Guarani-dominant Jopara corpora.

4. A customized tool for language identification.[6] This tool is made up of multiple other tools.

5. A method[7] for discovering topics in Spanish tweets (spoken in Spain) that combines linguistic knowledge with generative and discriminative approaches using the LDA (Latent Dirichlet Allocation) algorithm.

6. A detailed and well-organized set of procedures for creating a Twitter dataset for code-switching and low-resource languages, which outlines the limitations and difficulties encountered during the data-gathering process.

7. A Guarani tokenizer and a set of pre-trained Guarani language models, based on BERT, a widely used transformer-based model, that can be used for a variety of NLP tasks in Guarani or Jopara, such as sentiment analysis. These models were trained with data from Wikipedia in Guarani, including:[8] (a) three monolingual BERT models for the Guarani language and (b) two language models fine-tuned with Guarani (BETO and mBERT respectively).

References

Agüero-Torales, M., D. Vilares, and A. López-Herrera. 2021. On the logistical difficulties and findings of jopara sentiment analysis. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 95–102, Online, June. Association for Computational Linguistics.

Agüero-Torales, M. M., J. I. Abreu Salas, and A. G. López-Herrera. 2021. Deep learning and multilingual sentiment analysis on social media data: An overview. *Applied Soft Computing*, 107:107373.

Agüero-Torales, M. M., A. G. López-Herrera, and M. J. Cobo. 2018. Gastro-miner: Una Herramienta Basada en la Nube para el Análisis de Sentimientos en Opiniones sobre Restaurantes en TripAdvisor. Caso de Estudio sobre Restaurantes de la Provincia de Granada. In *I Jornadas Científicas en Ciencia de Datos*, page 34, Asunción, Paraguay, October. Universidad Comunera. Abstract (Poster).

Agüero-Torales, M. M., D. Vilares, and A. G. López-Herrera. 2021. Discovering topics in twitter about the covid-19 outbreak in spain. *Procesamiento del Lenguaje Natural*, 66(0):177–190.

Agüero-Torales, M., M. Cobo, E. Herrera-Viedma, and A. López-Herrera. 2019. A cloud-based tool for sentiment analysis in reviews about restaurants on tripadvisor. *Procedia Computer Science*, 162:392–399. 7th International Conference on Information Technology and Quantitative Management (ITQM 2019): Information technology and quantitative management based on Artificial Intelligence.

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*.

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.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[6] `https://github.com/mmaguero/lang-detection`

[7] `https://github.com/mmaguero/twitter-analysis`

[8] `https://huggingface.co/mmaguero`

Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 239-242 recibido 16-12-2022 revisado 01-02-2023 aceptado 11-02-2023

Sarcasm and Implicitness in Abusive Language Detection: A Multilingual Perspective

Sarcasmo e implícidad en el reconocimiento automático del lenguaje abusivo en una perspectiva multilingüe

Simona Frenda[1,2]
[1]PRHLT Research Center, Universitat Politècnica de València, Spain
[2]Dipartimento di Informatica, Università degli Studi di Torino, Italy
simona.frenda@unito.it

Abstract: PhD thesis in Computer Science focused on Natural Language Processing, written by Simona Frenda under the supervision of Prof. Viviana Patti and Prof. Paolo Rosso. This thesis was developed in a co-tutelle program between the PRHLT Research Center of the Universitat Politècnica de València (Spain) and the Computer Science Department of the University of Turin (Italy). In this work, we analysed, linguistically and computationally, the characteristics of the implicit abusive language, especially when it is masked as sarcastic. The thesis defence was held in Torino on June 6th, 2022. The doctoral committee was composed by: Prof. Liviu Petrisor Dinu (University of Bucharest, Romania), Prof. Els Lefever (Ghent University, Belgium) and Prof. Elena Cabrio (Université Côte d'Azur, France). An international mention was achieved, and the work was graded as excellent and awarded Cum Laude.
Keywords: Natural Language Processing, Computational Linguistics, Abusive Language Detection, Irony Detection, Stance Detection.

Resumen: Tesis doctoral en Informática con tema en Procesamiento del Lenguaje Natural realizada por Simona Frenda y dirigida por la Profa. Viviana Patti y el Prof. Paolo Rosso en el marco de un convenio de cotutela entre el PRHLT Research Center de la Universitat Politècnica de València (España) y el Departamento de Informática de la Universidad de Turín (Italia). En esta tesis se analiza a nivel lingüístico y computacional las características del lenguaje abusivo implícito, especialmente cuando está disfrazado como sarcástico. La defensa de la tesis fue en Turín el 6 de junio de 2022 ante un tribunal compuesto por el Prof. Liviu Petrisor Dinu (Universidad de Bucarest, Rumania), la Profa. Els Lefever (Universidad de Ghent, Bélgica), y la Profa. Elena Cabrio (Universidad de Côte d'Azur, Francia). Se obtuvo la mención internacional y una calificación de sobresaliente cum laude.
Palabras clave: Procesamiento del lenguaje natural, Lingüística computacional, Detección del lenguaje abusivo, Detección de la ironía, Detección de la stance.

1 Introduction

The possibility to monitor hateful content online on the basis of what people write is becoming an important topic for several actors such as governments, ICT companies, and NGO's operators conducting active campaigns in response to the worrying rise of online abuse and hate speech. Hand in hand, abusive language detection turns into a task of growing interest in Natural Language Processing (NLP), especially when applied to the recognition of various forms of hatred in social media posts. Abusive language is a broad umbrella term which is commonly used for denoting different kinds of hostile user-generated contents that intimidate or incite to violence and hatred, targeting many vulnerable groups in social platforms (Poletto et al., 2021). Such hateful contents are pervasive nowadays and can also be detected even in other kinds of texts, such as online newspapers. The importance of understanding and automatically detecting abusive language is due to the observation of real manifestations

ISSN 1135-5948 DOI 10.26342/2023-70-21

of violent acts connected to negative behaviours online in its various forms, such as cyberbullying, racism, sexism, or homophobia. Various approaches have been proposed in the last years to support the identification and monitoring of these phenomena, but unfortunately, they are far from solving the problem due to the inner complexity of abusive language, and to the difficulties to detecting its implicit forms (Wiegand, Ruppenhofer, and Eder, 2021).

In our doctoral investigation, we have studied the issues related to automatic identification of abusive language online, investigating various forms of hostility against women, immigrants and cultural minority communities in languages such as Italian, English, and Spanish. The analysis of the results of different methods of classification of hateful and non-hateful messages revealed important challenges that lie principally on the implicitness of some manifestations of abusive language expressed through the use of figurative devices (i.e., irony and sarcasm) (Frenda et al., 2022), recall of inner ideologies (i.e., sexist ideology) (Frenda et al., 2019a) or cognitive schemas (i.e., stereotypes) (Frenda, Patti, and Rosso, 2022), and expression of unfavourable stance (Frenda et al., 2019b).

To face these challenges, we have proposed distinct solutions applicable also to different textual genres. We observed that, in particular, cognitive (i.e., stereotypes) and creative aspects (i.e., sarcasm) of abusive language are harder to infer automatically from texts. Sarcasm, for instance, is a recurrent element in this kind of texts, and tends to affect the accuracy of the systems of recognition (Frenda, 2018). Indeed, for its peculiarities, sarcasm is apt to disguise hurtful messages, especially in short and informal texts such as the ones posted on Twitter. Its ironic sharpness and its echoic function of recalling a meaning that is the opposite or an extension of the literal one, make sarcasm appropriate to lower tones without losing the hurtfulness of the message. Moreover, funny messages are more likely to be accepted and shared by the community, making the abuse viral. Therefore, our hypothesis is that **information about the presence of sarcasm could help to improve the detection of hateful messages, even when they are camouflaged as sarcastic**. To verify it, we elaborated specific research questions:

RQ1 How to make abusive language detection systems sensitive to implicit manifestations of hate?
RQ2 What is the role played by sarcasm in hateful messages online?
RQ3 Could the awareness of the presence of sarcasm increase the performance of abusive language detection systems?

Focusing on these questions, 1) we investigated the characteristics of implicit manifestations of hate speech and examined, in terms of performance, the techniques that could help systems to infer them, such as the use of lexical resources, specific models to capture semantic relations, and the use of transfer learning techniques combined with linguistic features; 2) we analysed the role of ironic language in hateful texts, observing the multilingual characteristics of irony and especially of sarcasm, validating, with experiments of classification, these traits in terms of features; 3) we evaluated the benefits of ironic awareness in hate speech detection exploiting computational techniques that make systems aware of ironic language, such as the multi-task learning approach. This technique enables systems of abusive language detection to acquire specific knowledge about ironic language. Finally, we measured the significance of the obtained results in comparison to existing approaches and baseline models.

The corpora used in our experiments have been exploited as benchmark datasets within the EVALITA evaluation campaign for NLP tools for Italian, contributing to creating a new state of the art for these tasks in Italian: IronITA 2018 (Cignarella et al., 2018) and HaSpeeDe 2020 (Sanguinetti et al., 2020). Moreover, the multidisciplinary and multilingual frame of our analyses allowed us to reflect on the boundaries between dimensions and topical focuses that often overlap in computational approaches to detect abusive language and related phenomena.

2 Thesis Overview

The work presented in the thesis has been organized in 7 chapters grouped in 3 principal parts.

I part: Abusive Language Detection Chapter 1. The first chapter is the introductory section, where we described the social problems related to the new technologies, introducing the issue of the *abusive language* and the difficulties to detect it automatically.

Chapter 2 In the second chapter, we defined the concept of *abusive language*, looking at the juridical and linguistic theories. Moreover, we resumed the state of the art from a computational perspective, focusing especially on the open challenge of implicit abusive language detection.

Chapter 3 In the third chapter, we reported the linguistic, statistical and computational analysis performed on benchmark datasets to individuate the characteristics of the explicit and implicit manifestations of hate speech. Additionally, we described the linguistic resources created manually, and the designed approaches that make systems able to infer indirect abusive messages such as negative stereotypes (**RQ1**). Finally, we presented the second edition of the HaSpeeDe[1] shared task organized at EVALITA 2020 on hate speech and stereotypes detection in Italian tweets and news headlines.

II part: Irony and Sarcasm Detection

Chapter 4 In the fourth chapter, we defined what is *ironic language*, looking at the linguistic theories stretching from pragmatic to cognitive studies. In addition, we introduced the state of the art on irony and sarcasm detection, focusing especially on studies that analysed the peculiarities of sarcasm.

Chapter 5 In the fifth chapter, we proposed statistical and computational analysis to individuate the characteristics of irony and sarcasm. We observed linguistic traits of irony from a mono and multilingual perspective, and emotional and aggressive language involved in the expression of irony and sarcasm, especially when the topic of the text regards controversial issues such as the integration of cultural minorities (**RQ2**). In this chapter, we described also our experience as organizers of the IronITA[2] shared task at EVALITA 2018 on irony and sarcasm detection.

III part: Abusive and Ironic Language

Chapter 6 Taking into account the findings emerged from previous chapters, in the sixth one, we proposed a new computational approach that exploits the simultaneous learning from abusive and ironic language to detect hate speech in Italian tweets and news headlines. The results showed an improvement of the performance (33 % Δ), especially in hate speech detection in tweets, evaluated as significant (below a cut-off of 0,05) for all the metrics by means of a bootstrap sampling significance test (**RQ3**).

Chapter 7 In the last chapter, we reported the obtained results and the observations emerged from our analyses. We individuated the remaining challenges that we plan to address in further works, and we summarized the contributions to the NLP community in terms of findings, methodologies, resources, and publications.

3 Conclusions and Contributions

Various scholars in linguistics stress the mutual relation between language and society, composed of the speakers of that language. Actually, with *words* we are not just speaking, but we do things, things that could help people or things that could marginalize or hurt people. Our investigation aimed at contributing to the comprehension of how abuses, such as misogyny and racism, are expressed directly and indirectly, and how they could be recognized by machines.

The corpora-based analysis, the statistical tests and computational experiments on various benchmark datasets showed that abusive language towards women and immigrants involves important social biases that appear to be pervasive even in discussions that involve these targets. Another recurrent element is the presence of irony in these messages, used to lessen the social cost of their meaning. To make the systems of abusive language detection aware of stereotypes or prejudices, we experimented various approaches, discovering that especially lexica-based features are very useful even in the systems with neural architectures. Approaching abusive language detection as a classification problem, we noticed that one of the points that remained unsolved was related to the presence of ironic devices. Irony, in fact, is used to mask the purpose of haters to insult specific vulnerable targets. Ironic texts have been found to be aggressive, above all when the sarcastic form of irony is employed; proving, therefore, some arguments in favour of linguistic and pragmatic theories (Bowes and Katz, 2011).

Considering that, we designed a new approach of detection, exploiting the presence of irony in manual annotated texts. We designed a system that fine-tune Italian language models simultaneously on the tasks of hateful and ironic language recognition in a multi-task framework. We compared its re-

[1] http://di.unito.it/haspeede2020

[2] http://di.unito.it/ironita2018

sults with the one obtained with the previous approach that combines general knowledge, coming from language models, and linguistic information, provided by means of specific features. We discovered that the awareness of sarcasm helps the system to retrieve correctly hate speech in social media texts, such as tweets; and that linguistic features make the system sensible to stereotypes in both tweets and news headlines. Finally, our research questions encouraged also the investigation about irony and its manifestations in various contexts. Therefore, our analyses contributed also to the more theoretical and linguistic discussion on: 1) the peculiarities of sarcasm compared to other forms of irony, and 2) mono and multilingual characteristics of irony. Sarcasm, defined in literature as a sharp form of irony with the intent of scorning a victim, proved to be characterized by: hurtful language, explicit contradictions marked with adverbial locutions, semantic and polarity shifts, and false assertions and euphemistic forms. The computational experiments carried out on irony detection revealed, instead, that negative emotions are involved in the expression of irony, regardless of the language, the context, and the genre.

Although some important issues in Abusive Language detection have been addressed in this work, other challenges remain open for further investigation, such as: the misclassification of texts containing swear words used with non-abusive intent (surprise, friendly nicknames); and the processing of texts only at message level leaving unexplored the contextual information that could help to give a more informed perspective to interpret them as abuses or not.

Acknowledgments

The work in this thesis was partially supported by various financial projects. Among them: the Spanish research project SomEMBED funded by Ministerio de Economía y Sostenibilidad, the NII International Internship Program funded by JSPS KAKENHI, the Italian project M.EMO.RAI funded by RAI - Radiotelevisione Italiana Spa, the Italian project IhatePrejudice funded by Compagnia di San Paolo, and the European project "STERHEOTYPES" funded by Compagnia di San Paolo Foundation, Volkswagen Stiftung and Carlsberg Fondation.

References

Bowes, A. and A. Katz. 2011. When sarcasm stings. *Discourse Processes: A Multidisciplinary Journal*, 48(4):215–236.

Cignarella, A. T., S. Frenda, V. Basile, C. Bosco, V. Patti, and P. Rosso. 2018. Overview of the EVALITA 2018 task on irony detection in Italian tweets (IronITA). In *EVALITA 2018*, volume 2263. CEUR-WS.

Frenda, S. 2018. The role of sarcasm in hate speech: A multilingual perspective. In *Proceedings of Doctoral Symposium at SEPLN 2018*. CEUR-WS.

Frenda, S., A. T. Cignarella, V. Basile, C. Bosco, V. Patti, and P. Rosso. 2022. The unbearable hurtfulness of sarcasm. *Expert Systems with Applications*, 193:116398.

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

Frenda, S., K. Noriko, V. Patti, P. Rosso, et al. 2019b. Stance or insults? In *Ninth International Workshop on Evaluating Information Access*, pages 15–22. National Institute of Informatics.

Frenda, S., V. Patti, and P. Rosso. 2022. Killing me softly: Creative and cognitive aspects of implicitness in abusive language online. *Natural Language Engineering*, page 1–22.

Poletto, F., V. Basile, M. Sanguinetti, C. Bosco, and V. Patti. 2021. Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, 55:477–523.

Sanguinetti, M., G. Comandini, E. Di Nuovo, S. Frenda, M. A. Stranisci, C. Bosco, T. Caselli, V. Patti, and I. Russo. 2020. Haspeede 2@ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task. In *EVALITA 2020*. CEUR.

Wiegand, M., J. Ruppenhofer, and E. Eder. 2021. Implicitly Abusive Language–What does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the NAACL: Human Language Technologies*, pages 576–587.

Información General

SEPLN 2023

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

27-29 de septiembre 2023
http://sepln2023.sepln.org/

1 Presentación

La XXXIX edición del Congreso Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) se celebrará los días 27, 28 y 29 de septiembre de 2023, y estará precedido por una jornada de talleres de trabajo el día 26 de septiembre de 2023.

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:

- Desarrollo de recursos y herramientas lingüísticas.
- Análisis morfológico y sintáctico.
- Semántica, pragmática y discurso.
- Resolución de ambigüedad léxico-semántica.
- Generación de texto monolingüe y multilingüe.
- Traducción automática.
- Multimodalidad.
- Procesamiento del habla.
- Sistemas de diálogo / asistentes conversacionales.
- Indexación y recuperación de información multimedia.
- Recuperación y extracción de información monolingüe y multilingüe.
- Sistemas de búsqueda de respuestas.
- Evaluación de sistemas de PLN.
- Análisis automático de contenido textual.
- Análisis de opiniones y minería de la argumentación.
- Detección de plagio.

- Procesamiento de la negación y la especulación.
- Minería de texto en redes sociales.
- Resumen automático de texto.
- Simplificación de texto.
- Conocimiento y sentido común.
- PLN en el ámbito biomédico.
- Generación de recursos didácticos basada en PLN.
- PLN para lenguas con recursos limitados.
- Aplicaciones industriales del PLN.
- Aspectos éticos del PLN.
- Interpretabililidad y análisis de modelos para 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 quinta edición de IberLEF el día 26 de septiembre.

5 Comité ejecutivo SEPLN 2023

Presidencia del Comité Organizador

- L. Alfonso Ureña López (Universidad de Jaén).
- Mª. Teresa Martín Valdivia (Universidad de Jaén).

Coordinación:

- Eugenio Martínez Cámara (Universidad de Jaén).

Miembros:

- M. Carlos Díaz Galiano (Universidad de Jaén).
- Miguel Ángel García Cumbreras. (Universidad de Jaén).
- Manuel García Vega. (Universidad de Jaén).
- Salud María Jiménez Zafra. (Universidad de Jaén).
- Fernando Martínez Santiago. (Universidad de Jaén).
- M. Dolores Molina González. (Universidad de Jaén).
- Arturo Montejo Ráez. (Universidad de Jaén).
- Flor Miriam Plaza del Arco (Università Bocconi).

Colaboradores:

- Alba María Mármol Romero (Universidad de Jaén).
- Estrella Vallecillo Rodríguez (Universidad de Jaén).
- Mariia Chizhikova (Universidad de Jaén).
- Alberto José Gutiérrez Megías. (Universidad de Jaén).
- Jaime Collado Montañez. (Universidad de Jaén).

6 Consejo Asesor

Miembros:

- Xabier Arregi (Universidad del País Vasco, España)
- Aitziber Atutxa (Universidad del País Vasco, España)
- Miguel Ángel Alonso Pardo (Universidad de La Coruña, España)
- Manuel de Buenaga (Universidad de Alcalá, España)
- Jose Camacho Collados (Universidad de Cardiff, Reino Unido)
- Sylviane Cardey-Greenfield (Centre de recherche en linguistique et traitement automatique des langues, Francia)
- Irene Castellón (Universidad de Barcelona, España)
- Arantza Díaz de Ilarraza (Universidad del País Vasco, España)
- Antonio Ferrández (Universidad de Alicante, España)
- Koldo Gojenola (Universidad del País Vasco, España)
- José Miguel Goñi (Universidad Politécnica de Madrid, España)
- Inma Hernaez (Universidad del País Vasco, España)
- Elena Lloret (Universidad de Alicante, España)
- Ramón López-Cózar Delgado (Universidad de Granada, España)
- Bernardo Magnini Fondazione (Bruno Kessler, Italia)
- Nuno J. Mamede (Instituto de Engenharia de Sistemas e Computadores, Portugal)
- M. Teresa Martín Valdivia (Universidad de Jaén, España)
- Patricio Martínez-Barco (Universidad de Alicante, España)
- Eugenio Martínez Cámara (Universidad de Jaén, 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)
- Arturo Montejo Ráez (Universidad de Jaén, España)
- Manuel Montes y Gómez (Instituto Nacional de Astrofísica, Óptica y Electrónica, México)
- Mariana Neves (German Federal Institute for Risk Assessment, Alemania)
- Lluís Padró Universidad (Politécnica de Cataluña, España)
- Manuel Palomar (Universidad de Alicante, España)
- Ferrán Pla (Universidad Politécnica de Valencia, España)
- German Rigau (Universidad del País Vasco, España)
- Álvaro Rodrigo Yuste (Universidad Nacional de Educación a Distancia, España).
- Paolo Rosso (Universidad Politécnica de Valencia, España)
- Leonel Ruiz Miyares (Centro de Lingüística Aplicada de Santiago de Cuba, Cuba)
- Horacio Saggion (Universidad Pompeu Fabra, España)
- Emilio Sanchís (Universidad Politécnica de Valencia, España)
- Encarna Segarra (Universidad Politécnica de Valencia, España)
- Thamar Solorio (University of Houston, Estados Unidos de América)
- Maite Taboada (Simon Fraser University, Canadá)
- Mariona Taulé (Universidad de Barcelona, España)
- Juan-Manuel Torres-Moreno (Laboratoire Informatique d'Avignon / Université d'Avignon, Francia)
- José Antonio Troyano Jiménez (Universidad de Sevilla, España)
- L. Alfonso Ureña López (Universidad de Jaén, España)
- Rafael Valencia García (Universidad de Murcia, España)
- René Venegas Velásques (Pontificia Universidad Católica de Valparaíso, Chile)
- Felisa Verdejo Maíllo (Universidad Nacional de Educación a Distancia, España)
- Manuel Vilares (Universidad de la Coruña, España)
- Luis Villaseñor-Pineda (Instituto Nacional de Astrofísica, Óptica y Electrónica, México)

7 Fechas importantes

Fechas para la presentación y aceptación de comunicaciones:

- Fecha límite para la entrega de comunicaciones: 31 de marzo de 2023.
- Notificación de aceptación: 16 de mayo de 2023.
- Fecha límite para entrega de la versión definitiva: 31 de mayo de 2023.

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.), además de referencias y figuras.
- Los artículos pueden estar escritos en inglés o español. El título, resumen y palabras clave deben escribirse en ambas lenguas.
- El formato será en Word ó LaTeX

Envío de los Trabajos

- El envío de los trabajos se realizará electrónicamente a través de la plataforma de envío publicada en: http://www.sepln.org/la-revista/informacion-para-autores.
- Para los trabajos con formato LaTeX se enviará el archivo PDF.
- 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:

Xabier Arregi	Universidad del País Vasco (España)
Aitziber Atutxa	Universidad del País Vasco (España)
Miguel Ángel Alonso Pardo	Universidad de La Coruña (España)
Manuel de Buenaga	Universidad de Alcalá (España)
Jose Camacho Collados	Universidad de Cardiff (Reino Unido)
Sylviane Cardey-Greenfield	Centre de recherche en linguistique et traitement automatique des langues (Francia)
Irene Castellón	Universidad de Barcelona (España)
Arantza Díaz de Ilarraza	Universidad del País Vasco (España)
Antonio Ferrández	Universidad de Alicante (España)
Koldo Gojenola	Universidad del País Vasco (España)
José Miguel Goñi	Universidad Politécnica de Madrid (España)
Inma Hernaez	Universidad del País Vasco (España)
Elena Lloret	Universidad de Alicante (España)
Ramón López-Cózar Delgado	Universidad de Granada (España)
Bernardo Magnini	Fondazione Bruno Kessler (Italia)

Cartas al director

Sociedad Española para el Procesamiento del Lenguaje Natural
Departamento de Informática. Universidad de Jaén
Campus Las Lagunillas, Edificio A3. Despacho 127. 23071 Jaén
secretaria.sepln@ujaen.es

Más información

Para más información sobre la Sociedad Española del Procesamiento del Lenguaje Natural puede consultar la página web http://www.sepln.org.

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