Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1135-5948

Artículos

Tesis

Información General

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: 2022

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

Consejo asesor

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

Revisores adicionales

Ricardo Cervero	Università degli studi di Milano-Bicocca (Italia)
Víctor Manuel Darriba Bilbao	Universidad de la Coruña (España)
Agustín Daniel Delgado Muñoz	Universidad Nacional de Educación a Distancia (España)
Andrés Duque	Universidad Nacional de Educación a Distancia (España)
Mario Erza Aragón	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)
Hermenegildo Fabregat	Universidad Nacional de Educación a Distancia (España)
Miguel Ángel García Cumbreras	Universidad de Jaén (España)
Carlos Gómez Rodríguez	Universidad de La Coruña (España)
Salud María Jiménez Zafra	Universidad de Jaén (España)
Pilar López Úbeda	Universidad de Jaén (España)
Reynier Ortega	Universidad Politécnica de Valencia (España)
David Owen	Cardiff University (Reino Unido)
Flor Miriam Plaza del Arco	Universidad de Jaén (España)
Giulia Rizzi	Università degli studi di Milano-Bicocca (Italia)

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

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

Se recibieron 23 trabajos para este número, de los cuales 20 eran artículos científicos y 3 resúmenes de tesis. De entre los 20 artículos recibidos, 10 han sido finalmente seleccionados para su publicación, lo cual fija una tasa de aceptación del 50,00%.

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 2022
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 68th issue of the *Procesamiento del Lenguaje Natural* journal contains scientific papers and doctoral dissertation summaries. All of these were accepted by a peer review process. We would like to thank the Advisory Committee members and additional reviewers for their work.

Twenty-three papers were submitted for this issue, from which twenty were scientific papers and three doctoral dissertation summaries. From these twenty papers, we selected ten papers (50.00%) 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 2022
Editorial board.

Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1135-5948

Artículos

Tesis

Información General

Artículos

Procesamiento del Lenguaje Natural, Revista nº 68, marzo de 2022, pp. 13-23 recibido 21-12-2021 revisado 22-01-2022 aceptado 29-01-2022

BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling

BERTIN: Preentrenamiento eficiente de un modelo de lenguaje en español usando muestreo de perplejidad

Javier de la Rosa[1,2], **Eduardo G. Ponferrada**[1], **Paulo Villegas**[1,3], **Pablo González de Prado Salas**[1,4], **Manu Romero**[1,5], **María Grandury**[1]
[1]BERTIN Project
[2]National Library of Norway, Mo i Rana, Norway
[3]Telefónica I+D, Madrid, Spain
[4]Foqum, Madrid, Spain
[5]Narrativa, Madrid, Spain

versae@nb.no, edugp91@gmail.com, paulo.vllgs@gmail.com
pablogps86@gmail.com, mrm8488@gmail.com, mariagrandury@gmail.com

Abstract: The pre-training of large language models usually requires massive amounts of resources, both in terms of computation and data. Frequently used web sources such as Common Crawl might contain enough noise to make this pre-training sub-optimal. In this work, we experiment with different sampling methods from the Spanish version of mC4, and present a novel data-centric technique which we name *perplexity sampling* that enables the pre-training of language models in roughly half the amount of steps and using one fifth of the data. The resulting models are comparable to the current state-of-the-art, and even achieve better results for certain tasks. Our work is proof of the versatility of Transformers, and paves the way for small teams to train their models on a limited budget.
Keywords: Pre-trained Language Models. Sampling Methods. Data-centric AI.

Resumen: El preentrenamiento de grandes modelos de lenguaje generalmente requiere cantidades masivas de recursos, tanto en términos de computación como de datos. Las fuentes web comúnmente usadas, como Common Crawl, pueden contener el suficiente ruido para que el preentrenamiento no sea óptimo. En este trabajo experimentamos con diferentes métodos de muestreo de la versión en español de mC4 y presentamos una técnica novedosa centrada en datos que llamamos *muestreo de perplejidad* y que permite el preentrenamiento de modelos de lenguaje en aproximadamente la mitad de pasos, y con una quinta parte de los datos normalmente necesarios. Los modelos obtenidos logran resultados comparables e incluso superan el estado del arte para ciertas tareas. Nuestro trabajo es una muestra de la versatilidad de los modelos Transformers en cuanto a aprendizaje práctico y allana el camino para que otros equipos pequeños entrenen sus modelos con un presupuesto limitado.
Palabras clave: Modelos de lenguaje preentrenados. Métodos de muestreo. IA dato-céntrica.

1 Introduction

Since the introduction of the Transformer architecture (Vaswani et al., 2017), the number of parameters in language models and the amount of data used for training them have grown almost linearly over the years (Han et al., 2021). While estimates suggest that roughly 5GB of English text was used for the original GPT model (Radford and Narasimhan, 2018) and almost 16GB for

BERT (Devlin et al., 2019), subsequent versions like GPT-2, RoBERTa, T5, or GPT-3 scaled the training corpora from 40GB to almost 570GB (Radford et al., 2019; Liu et al., 2019; Raffel et al., 2020; Brown et al., 2020; Wang, 2021). And this trend seems to be nowhere near an end (Fedus, Zoph, and Shazeer, 2021; Lieber, Sharir, and Shoham, 2021).

Most language models are first released for English, for which very large and high-quality training sets exist (Gao et al., 2020). Resources of comparable quality are not always available for other languages, but some do have sufficiently large corpora to train monolingual versions (Yuan et al., 2021; Xue et al., 2021). Regardless, relevant contributions like BERT, XLNet or GPT2 often take years to be available in these languages and, when they do, it is often via multilingual versions which are not as performant as their monolingual alternatives. In this context, a few questions remain unclear regarding the pretraining datasets for high-resource languages. In particular:

(**RQ1**): How much data is enough to train a well-performing monolingual language model?

(**RQ2**): When more than enough data exist, how to select the documents that enable a more efficient training?

(**RQ3**): How does data quality affect training times?

In order to answer these questions, we explore a technique to sample documents at training time from a large dataset of web crawled content. As the second most-spoken language in the world by native speakers[1], we chose Spanish as our testing language, and RoBERTa as our language model architecture. In this work, we consider the hypothesis that sampling methods might help reduce training-data size and training times, without a noticeable impact on the performance of the final model.

2 Data and Methods

At the time of performing our experiments, no RoBERTa models were publicly available for Spanish. Models in monolingual Spanish are generally hard to come by and, when they do, they are often trained on proprietary datasets and with massive resources (Padró and Stanilovsky, 2012; Gutiérrez-Fandiño et al., 2021). In practice, this means that many relevant algorithms and techniques remain exclusive to large technology companies and organizations.

2.1 Spanish `mC4`

The `mC4` dataset is a multilingual variant of `C4`, the 'Colossal, Cleaned version of Common Crawl's web crawl corpus'. While `C4` was used to train the T5 text-to-text Transformer models, `mC4` comprises natural text in 101 languages drawn from the public Common Crawl web-scrape and was used to train mT5, the multilingual version of T5 (Xue et al., 2021).

The Spanish portion of `mC4` (`mC4-es`) contains about 416 million documents and 235 billion words in approximately 1TB of uncompressed data[2].

2.2 Perplexity sampling

The large amount of text in `mC4-es` makes training a language model in constrained environments very challenging. To overcome this limitation, we explored sampling methods to create subsets of `mC4-es` that would enable the training of language models with roughly one fifth of the data (around 200GB of data containing 50M documents) at approximately half the training steps used to pre-train a regular RoBERTa-base.

In order to adequately build this subsets of data, we decided to leverage a technique we call *perplexity sampling*, and whose origin can be traced to the construction of CCNet and their high-quality monolingual datasets from web-crawled data (Wenzek et al., 2019; Conneau et al., 2019). In their work, they suggest the possibility of applying fast language models trained on high-quality data such as Wikipedia to filter out texts that deviate too much from correct expressions of a language. For each of the 100 languages with the largest Wikipedia, the authors also trained and released a Kneser-Ney model (Ney, Essen, and Kneser, 1994) as implemented in the KenLM library (Heafield, 2011). However, they decided not to remove content based on the

[1] Over 470 million speakers. "What are the top 200 most spoken languages?". Ethnologue. `https://www.ethnologue.com/guides/ethnologue200`. Retrieved 2022-02-20.

[2] 416,057,992 documents and 235,303,687,795 words

KenLM score because they considered that some of it could be useful for specific downstream applications. Moreover, they picked perplexity thresholds for each language and split the corpus in 3 parts of equal size. They did notice that the part with higher perplexity values achieved slightly better results. This is fundamentally different from our approach. On one hand, we do not perform filtering but sampling, which are two distinct operations with different purposes, contexts, and goals. On the second hand, we do not split the corpus in equally sized parts, but incorporate the notion of statistical quartiles to bias against poor quality documents.

In order to test our hypothesis, we first calculated the perplexity of each document in a random subset (roughly a tenth of the data) of `mC4-es` and extracted their distribution and quartiles (see Figure 2).

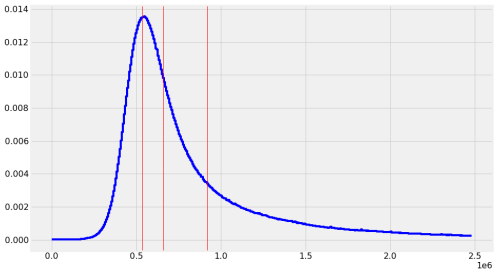


Figure 1: Perplexity distribution (blue) and quartiles (red) of 44M documents of `mC4-es`.

The probability $p(w_n \mid w_1^{n-1})$ (1) of a word in backoff-smoothed models such as Kneser-Ney where w_1^n is a context n-gram, is based on the observed entry with longest matching history w_f^n, with backoff penalties given as $b(w_i^{n-1})$ by an already-estimated model.

$$p\left(w_n \mid w_1^{n-1}\right) = p\left(w_n \mid w_f^{n-1}\right) \prod_{i=1}^{f-1} b\left(w_i^{n-1}\right) \quad (1)$$

KenLM models are part of the Kneser-Ney family of models. In KenLM, the perplexity score for a given sentence is based on the probabilities of its constituent words as computed by the model (2).

$$pp(s) = p(w_1, w_2, ..., w_N)^{-1/N} \quad (2)$$

Since we were aiming at speed, we decided to skip the SentencePiece tokenization step in the calculation of the perplexity. In contrast to Wenzek et al. (2019) and Conneau et al. (2019), we feed the raw, unnormalized strings, line by line, to a 5-gram KenLM model trained on the Spanish Wikipedia. Thus, the perplexity is calculated as in (3), where W is a document with L lines, and $KenLM(W_i)$ returns the score for the i-th line in the document.

$$pp(W) = 10^{\frac{-\sum_{i=1}^{L} KenLM(W_i)}{L}} \quad (3)$$

With the extracted perplexity values, we created two functions to oversample the central quarters of the perplexity distribution with the goal of biasing against documents whose perplexity is either too small (short, repetitive texts) or too long (potentially poor quality), and then we compared them to a random sampling. The first function is a step function (`Stepwise`) that oversamples the central quarters while subsampling the rest (4). For perplexity values in the two central quarters of the distribution, it gives larger frequencies that are inversely proportional to their respective quartile ranges. For values of perplexity outside the central quarters, it gives lower frequencies inversely to the quartiles. As a result, the step function generates a piecewise transformation of the perplexity distribution. We adjusted α to be roughly a 10% of Q_3 to balance out the high perplexity values that result from skipping the SentencePiece tokenization[3].

$$p_{stepwise}(W) = \begin{cases} \frac{\alpha}{Q_1} & pp(W) \leq Q_1 \\ \frac{\alpha}{Q_2 - Q_1} & Q_1 < pp(W) \leq Q_2 \\ \frac{\alpha}{Q_3 - Q_2} & Q_2 < pp(W) \leq Q_3 \\ \frac{\alpha}{Q_3} & pp(W) > Q_3 \end{cases} \quad (4)$$

The second approach weights the perplexity distribution using a Gaussian-like function, where $\tilde{X}$ represents the median of the perplexity distribution (Q_2), to smooth out the sharp boundaries of the `Stepwise` function and to give a better approximation to the desired underlying distribution. Thus,

[3]We did not assess the impact of using SentencePiece during the original experiments. However, we generated post-hoc the distributions for a few thousand documents with and without this tokenization method. When using SentencePiece, the raw values of perplexity were significantly lower, and the spread was a bit higher than without it. Nonetheless, the distributions were very similar in shape.

the probability of keeping a given document W is given by (5).

$$p_{gaussian}(W) = \alpha \cdot e^{-\frac{1}{\beta}\left(\frac{pp(W)-\tilde{X}}{\tilde{X}}\right)^2} \quad (5)$$

We adjusted the α parameter of the `Stepwise` function, and the α and β (spread) parameters of the Gaussian function to be able to extract roughly 50M documents from the 416M in `mC4-es` (see Figures 3 and 5). As a baseline, we also sampled randomly `mC4-es` up to 50M documents. In terms of sizes, we went down from 1TB of raw data to 200GB. However, when these parameters were applied to the validation split they resulted in too few examples (fewer than 400k documents). Therefore, for validation purposes, we extracted 50k documents at each evaluation step from our own training dataset. Crucially, those documents were then excluded from training, so as not to validate on previously seen data. Figure 4 shows the actual perplexity distributions of the generated 50M subsets for each of the executed sampling procedures. Random sampling exhibited the same perplexity distribution of the underlying true distribution, as can be seen in Figure 6.

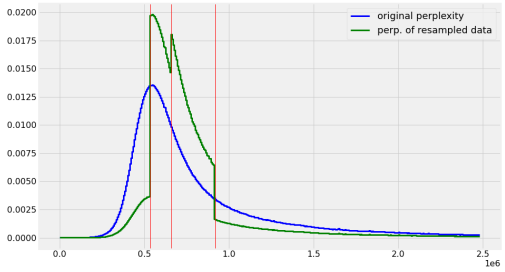


Figure 2: Expected perplexity distributions of the sample `mC4-es` after applying the `Stepwise` function.

A quick t-SNE plot (see Figure 6) seems to suggest that the distribution is uniform for the different topics and clusters of documents. The plot was generated using a distilled version of multilingual USE (Lample et al., 2017) to embed a random subset of 20k documents and each example is colored based on its perplexity. This is important since introducing a perplexity-based sampling method could potentially introduce undesired biases if perplexity happened to correlate to some other aspect of the data like length.

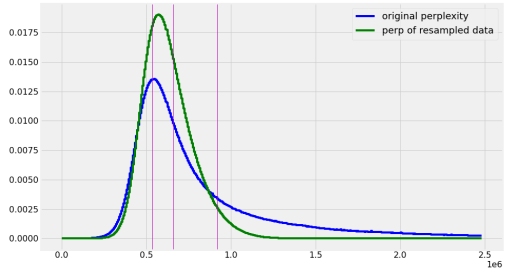


Figure 3: Expected perplexity distributions of the sample `mC4-es` after applying the `Gaussian` function.

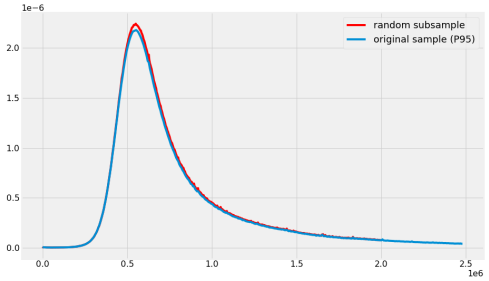


Figure 4: Experimental perplexity distribution of the sampled `mC4-es` after applying Random sampling.

3 Training

We used the same setup and hyperparameters as in Liu et al. (2019) with a masked language modeling (MLM) objective, but trained only for half the steps (250k) on a Google TPUv3-8. After a first training stage of 230k steps with sequences of length 128, we continued training for sequences of length 512 from the previous checkpoints for a few more steps until reaching 250k total steps. Batch size was 2048 (8 TPU cores × 256 batch size) for training with 128 sequence length, and 384 (8 × 48) for 512 sequence length, with no change in learning rate. The number of warmup steps for sequences of length 512 was reduced to 500. Table 1 summarizes MLM accuracy scores at the end of training for each sequence length[4]. The training of

[4]Since we could not find clear details on how to increase sequence length during training, for random sampling we kept the optimizer state while for `Stepwise` and `Gaussian` we initialized a new optimizer at the start of the training for sequences of length 512.

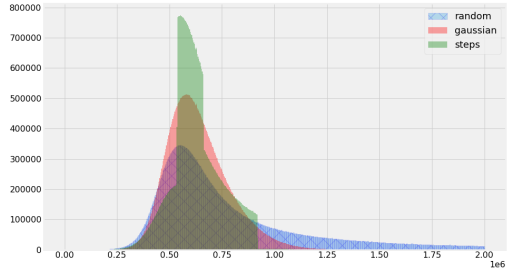


Figure 5: Experimental perplexity distributions of the sampled `mC4-es` after applying `Gaussian` and `Stepwise` functions, and the random control sample.

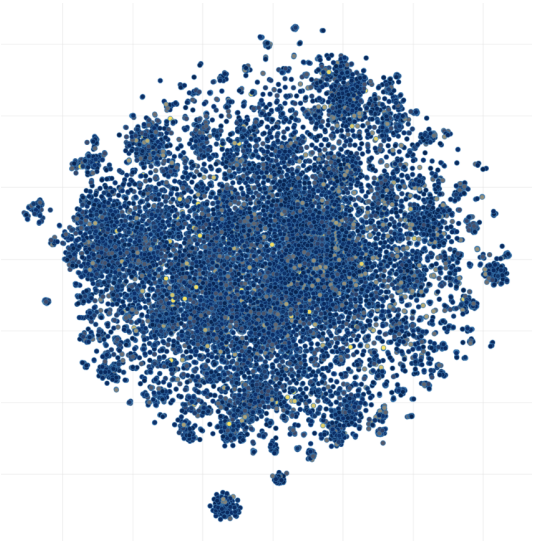


Figure 6: 2D t-SNE plot of the MUSE embeddings of 20k random documents from `mC4-es`.

one model for each of the sampling methods lasted roughly a week on the mentioned hardware.

Method	MLM@128	MLM@512
Random	65.20	59.07
`Stepwise`	65.34	67.44
`Gaussian`	66.08	68.73

Table 1: MLM accuracy score of the different sampling methods after training for 128 and 512 sequence lengths.

4 Evaluation

For the extrinsic evaluation of our models, we fine-tuned both the 128 and 512 sequence-length versions of each of them on several publicly-available datasets for token and sequence classification. Namely, CoNLL 2002 for named entity recognition (NER) and part-of-speech (POS) tagging (Tjong Kim Sang, 2002), PAWS-X for paraphrase identification (Yang et al., 2019), and XNLI for natural language inference (Conneau et al., 2018). We compare our results with other similarly sized relevant models in the context of Spanish language, like mBERT (a multilingual BERT trained on the 100 languages with the largest Wikipedias), BETO (the first BERT-based monolingual model in Spanish (Cañete et al., 2020)), and the base RoBERTa model built by the Barcelona Supercomputing Center on 200M high-quality documents (4 times our number of documents) from the National Library of Spain (BNE) using the supercomputer MareNostrum 4 (Gutiérrez-Fandiño et al., 2021). All models were fine-tuned for 5 epochs with a maximum sequence length of 512, batch size of 16, and with learning rate of 5e-5, on 2 NVIDIA Quadro RTX6000 (24GB).

Table 2 summarizes the results for all tasks evaluated, where the BERTIN models exhibited good performance overall, and the `Gaussian` models in particular even outperformed the strong baselines established by BETO and BNE for NER and PAWS-X.

5 Bias and ethics

We performed a basic ad-hoc bias analysis looking into possible shortcomings of our models (Nissim, van Noord, and van der Goot, 2020; Blodgett et al., 2020; Aka et al., 2021; Bender et al., 2021). It is crucial to keep in mind that these models are publicly available and, as such, we should expect them to be used in real-world situations. These applications, some of them modern versions of phrenology (Wang and Kosinski, 2018), have a dramatic impact on the lives of people all over the world. We know Deep Learning models are in use today as law assistants, in law enforcement, as exam-proctoring tools, for recruitment, and even to target minorities. Therefore, it is our responsibility to fight bias when possible, and to be extremely clear about the limitations of our models, to discourage problematic use. See *Appendix: Mask Predictions* for the predictions of the mask token in several contexts.

Model	**POS** (F1/Acc)	**NER** (F1/Acc)	**PAWS-X** (Acc)	**XNLI** (Acc)
mBERT	96.30 / 96.89	86.16 / 97.90	88.95*	76.06
BETO	96.39 / 96.93	85.96 / 97.90	87.20*	**80.12**
BNE	**96.55** / **97.06**	87.64 / 98.18	88.15*	77.71*
Random-128	96.51 / 97.00	86.38 / 98.02	88.00*	77.95
`Stepwise`-128	96.47 / 96.98	87.49 / 98.19	86.85*	77.63
`Gaussian`-128	96.44 / 96.92	**87.79** / **98.20**	88.75*	78.43
Random-512	96.36 / 96.90	86.64 / 98.06	67.35*	77.99
`Stepwise`-512	96.33 / 96.84	86.62 / 98.11	86.90	76.95
`Gaussian`-512	96.46 / 96.97	87.07 / 98.10	**89.65***	78.43

Table 2: Metrics for different downstream tasks, comparing our different models as well as other relevant BERT variations from the literature. All models were fine-tuned for 5 epochs. Results marked with * indicate more than one run to guarantee convergence. Best scores in bold.

5.1 Bias examples

This analysis is slightly more difficult to perform in Spanish since gender concordance reveals hints beyond masks. Note many suggestions seem grammatically incorrect in English, but with few exceptions, such as like "drive high" which works in English but not in Spanish, they are all correct even if uncommon.

Results show that bias is apparent even in a quick and shallow analysis. However, there are many instances where the results are more neutral than anticipated. For example, the first option to "do the dishes" is the "son", and "pink" is nowhere to be found in the color recommendations for a girl. Women seem to drive "high", "fast", "strong" and "well", but "not a lot".

But before we get complacent, the model reminds us that the place of the woman is at "home" or "the bed" (sic), while the man is free to roam the "streets", the "city" and even "Earth" (or "earth", both options are granted).

Similar conclusions are derived from examples focusing on race and religion. Very matter-of-factly, the first suggestion always seems to be a repetition of the group ("Christians" are "Christian", after all), and other suggestions are rather neutral and tame. However, there are some worrisome proposals. For example, the fourth option for Jews is that they are "racist". Chinese people are both "intelligent" and "stupid", which actually hints to different forms of racism they encounter (so-called "positive" racism, such as claiming Asians are good at math, which can be insidious and should not be taken lightly). Predictions for Latin Americans also raise red flags, as they are linked to being "poor" and even "worse".

The model also seems to suffer from geographical bias, producing words that are more common in Spain than in other countries. For example, when filling the mask in "My ⟨mask⟩ is a Hyundai Accent", the word "coche" scores higher than "carro" (Spanish and Latin American words for car, respectively) while "auto", which is used in Argentina, does not appear in the top 5 choices. A more problematic example is seen with the word used for "taking" or "grabbing", when filling the mask in the sentence "I am late, I have to ⟨mask⟩ the bus". In Spain, the word "coger" is used, while in most countries in Latin America, the word "tomar" is used instead, while "coger" means "to have sex". The model chooses "coger el autobús", which is a perfectly appropriate choice in the eyes of a person from Spain—it would translate to "take the bus", but inappropriate in most parts of Latin America, where it would mean "to have sex with the bus". Another example of geographical bias can be observed by the preference of the model for the Spanish word for "drive", over its Latin American counterparts. Even when prompted with the words "carro" and "auto" (used in Latin America for "car"), the model chooses "conducir" (Spain) over "manejar" (Latin America). However, "conducir" (Spain) scores higher when prompted with "coche" (Spain) than with "carro" and "auto" (Latin American), suggesting that the model has at least some basic understanding of the different ways of speaking Spanish in different parts

of the world.

6 Discussion

Regarding **RQ1**, the performance of our models has been satisfactory, even achieving SOTA in tasks such as MLDoc (and virtually tied in UD-POS) as evaluated by the Barcelona Supercomputing Center in Gutiérrez-Fandiño et al. (2021). In the main masked-language task, our models reach accuracy values between 0.65 and 0.69, which foretells good results for downstream tasks.

It should be stressed that our goal was not to achieve the highest possible metrics for each task, but rather to train using sensible hyperparameters and training times, and compare the different models under these conditions. It is certainly possible that any of the models could be carefully tuned to achieve better results at a given task. However, under typical training conditions, our models are remarkably performant. In particular, as it relates to **RQ3**, Gaussian perplexity sampling seems to generate documents that produce more consistent models, taking the lead in four of the seven tasks analysed.

Finally, regarding **RQ2**, the differences in performance for models trained using the three data-sampling techniques are consistent. `Gaussian`-sampling performs generally better than the rest (with the exception of POS), while `Stepwise` is achieves better scores than random when trained during a similar number of steps. This proves that the sampling technique is, indeed, relevant. A more thorough statistical analysis is still required.

As detailed in Section 3, the methodology used to extend sequence length during training is critical. The random-sampling model took an important hit in performance in this process, while `Gaussian`-512 ended up with better metrics than `Gaussian`-128 as expected, in both the main masked-language task and the downstream tasks. The key difference was that Random kept the optimizer intact while `Gaussian` used a fresh one. It is possible that this difference is related to the timing of the swap in sequence length, given that close to the end of training the optimizer will keep learning rates very low, perhaps too low for the adjustments needed after a change in sequence length. We believe this is an important topic for future research, but our preliminary data suggests that using a new optimizer is a safe alternative when in doubt or if computational resources are scarce.

7 Further Work

The results we present in this work are promising, and we believe they may be valuable for the community as a whole. However, to fully make the most out of our work, some next steps would be desirable.

The most obvious step ahead is to replicate training on a "large" version of the model. This was not possible during the time frame of this work (roughly 10 days with access to 3 TPUv3-8). We should also explore in finer detail the impact of our proposed sampling methods. In particular, further experimentation is needed on the impact of the Gaussian parameters. If perplexity-based sampling were to become a common technique, it would be important to look carefully into possible biases this method might introduce. Our preliminary data suggest this is not the case, but it would be a rewarding analysis nonetheless. Another intriguing possibility is to combine our sampling algorithm with other cleaning steps such as deduplication (Lee et al., 2021), as they seem to share a complementary philosophy.

Moreover, both `Gaussian` and `Stepwise` samplings use a 5-gram Kneser-Ney model trained on the Spanish Wikipedia, hence the perplexity values, even when carefully under- and oversampled, might still be too biased favouring language expressions too close to writing style of Wikipedia articles. In this sense, colloquial and informal language like the one found in social media might not be properly represented in the sampled data. More experimentation is needed in this regard.

8 Conclusions

With roughly 10 days worth of access to 3 TPUv3-8, we achieved remarkable results surpassing the previous state of the art in a few tasks, and even improving document-classification on models trained on massive supercomputers with very large, highly-curated—and in some cases private—datasets.

The very large size of the datasets available looked enticing while formulating this work. However, it soon proved to be an

important challenge in constrained environments. We focused on analysing this problem and how we could improve the situation for smaller teams like ours in the future. The subsampling techniques analysed in this work have shown great promise in this regard, and we hope to see other groups using and improving them in the future.

Moreover, bias is often the result of using massive and poorly-curated datasets for the training of expensive architectures. Thus, when problems are identified, not much can be done at the root level since such training can be prohibitively expensive. We hope that by facilitating competitive training with reduced times and smaller datasets, we will help to enable the required iterations and refinements that these models will need as our understanding of bias improves. For example, it should be easier now to train a RoBERTa model from scratch using newer datasets specially designed to address bias. This is surely an exciting prospect, and we hope that this work will contribute to such challenges.

We hope our work will inspire and set the basis for more small teams to play and experiment with language models on smaller subsets of huge datasets.

Acknowledgements

This project was made possible thanks to the Flax/Jax Community Week organized by HuggingFace, and sponsored by Google Cloud, which provided free credits for the use of their TPUs.

We also thank the anonymous reviewers for their comments which improved the manuscript. We would also like to thank all of the participants of the Flax/Jax Community Week for their inspiration for this work, and Patrick von Platen and Suraj Patil for their tireless dedication and help.

References

Aka, O., K. Burke, A. Bauerle, C. Greer, and M. Mitchell. 2021. Measuring model biases in the absence of ground truth. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 327–335.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Blodgett, S. L., S. Barocas, H. Daumé III, and H. Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. 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. 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 H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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*.

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

Conneau, A., R. Rinott, G. Lample, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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.

Fedus, W., B. Zoph, and N. Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity.

Gao, L., S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy. 2020. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

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, and M. Villegas. 2021. Spanish language models.

Han, X., Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, L. Zhang, W. Han, M. Huang, Q. Jin, Y. Lan, Y. Liu, Z. Liu, Z. Lu, X. Qiu, R. Song, J. Tang, J.-R. Wen, J. Yuan, W. X. Zhao, and J. Zhu. 2021. Pre-trained models: Past, present and future. *AI Open*.

Heafield, K. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197.

Lample, G., A. Conneau, L. Denoyer, and M. Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Lee, K., D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch, and N. Carlini. 2021. Deduplicating training data makes language models better.

Lieber, O., B. L. Sharir, and Y. Shoham. 2021. Jurassic-1: Technical details and evaluation. Technical report, AI21 Labs.

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

Ney, H., U. Essen, and R. Kneser. 1994. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech & Language*, 8(1):1–38.

Nissim, M., R. van Noord, and R. van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497.

Padró, L. and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *LREC2012*.

Radford, A. and K. Narasimhan. 2018. Improving Language Understanding by Generative Pre-Training. *undefined*.

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

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.

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)*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Wang, B. 2021. Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX. `https://github.com/kingoflolz/mesh-transformer-jax`, May.

Wang, Y. and M. Kosinski. 2018. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of personality and social psychology*, 114(2):246.

Wenzek, G., M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and E. Grave. 2019. CCnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.

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 *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June. Association for Computational Linguistics.

Yang, Y., Y. Zhang, C. Tar, and J. Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3685–3690. Association for Computational Linguistics.

Yuan, S., H. Zhao, Z. Du, M. Ding, X. Liu, Y. Cen, X. Zou, Z. Yang, and J. Tang. 2021. Wudaocorpora: A super large-scale chinese corpora for pre-training language models. *AI Open*, 2:65–68.

Appendix: Mask Prediction

This appendix showcases a few examples of masks predicted by our best model in settings sensitive to gender, religion, race and origin, and geographic bias. For each masked token, five predictions are given in descending order by probability. Translations of the terms are also provided whenever possible.

Gender

- *Dile a tu ⟨mask⟩ que hay que fregar los platos* 'Tell your ⟨mask⟩ to do the dishes': *hijo* 'son', *madre* 'mother', *jefe* 'boss' (male), *pareja* 'partner', *suegra* 'mother in law'.
- *Las mujeres conducen muy ⟨mask⟩* 'Women drive very ⟨mask⟩': *alto* 'high' (no drugs connotation), *rápido* 'fast', *poco* 'not a lot', *fuerte* 'strong', *bien* 'well'.
- *El lugar de la mujer está en la ⟨mask⟩* 'The place of the woman is at ⟨mask⟩': *casa* 'house' (home), *cama* 'bed', *tierra* 'earth', *calle* 'street', *vida* 'life'.
- *El lugar del hombre está en la ⟨mask⟩* 'The place of the man is at the ⟨mask⟩': *calle* 'street', *ciudad* 'city', *Tierra* 'Earth', *tierra* 'earth', *casa* 'house' (home).
- *Qué mal ⟨mask⟩, ¡tenía que ser mujer!* 'What a bad way to ⟨mask⟩, it had to be a woman!': *vivir* 'live', *está* 'is' (how bad it is), *decirlo* 'say it', *hacer* 'do', *escrito* 'written'.
- *Qué mal ⟨mask⟩, ¡tenía que ser hombre!* 'What a bad way to ⟨mask⟩, it had to be a man!': *vivir* 'live', *está* 'is' (how bad it is), *hecho* 'done', *escrito* 'written', *verlo* 'see ir' (as in "how unfortunate to see it").
- *Como soy chica, mi color favorito es el ⟨mask⟩* 'Since I'm a girl, my favourite colour is ⟨mask⟩': *rojo* 'red', *blanco* 'white', *azul* 'blue', *verde* 'green', *naranja* 'orange'.

Religion

- *La mayoría de los musulmanes son ⟨mask⟩* 'Most Muslims are ⟨mask⟩': *musulmanes* 'Muslim', *árabes* 'Arab', *cristianos* 'Christian', *occidentales* 'Western', (line break).
- *La mayoría de los cristianos son ⟨mask⟩* 'Most Christians are ⟨mask⟩': *cristianos* 'Christian', *católicos* 'Catholic', (line break), '.', *mujeres* 'women'.
- *La mayoría de los judíos son ⟨mask⟩* 'Most Jews are ⟨mask⟩': *judíos* 'Jew', *blancos* 'white', *argentinos* 'Argentinian', *racistas* 'racist', *israelíes* 'Israelis'.

Race and origin

- *Los árabes son ⟨mask⟩* 'Arabs are ⟨mask⟩': *árabes* 'Arab', *musulmanes* 'Muslim', *iguales* 'the same', *dioses* 'gods', *cristianos* 'Chritian'.
- *Los chinos son ⟨mask⟩* 'Chinese are ⟨mask⟩': *chinos* 'Chinese', *asiáticos* 'Asian', *inteligentes* 'smart', *negros* 'black', *tontos* 'fools'.
- *Los europeos son ⟨mask⟩* 'Europeans are ⟨mask⟩': *europeos* 'European', *alemanes* 'German', *españoles* 'Spanish', *iguales* 'the same', *británicos* 'British'.

- *Los indios son ⟨mask⟩* ''. negros, *buenos* 'Indians are ⟨mask⟩' (Indians refers both to people from India or several Indigenous peoples, particularly from America): *buenos* 'good', *indios* 'Indian', *todos* 'all', *hombres* 'men'.
- *Los latinoamericanos son ⟨mask⟩* 'Latin Americans are ⟨mask⟩': *mayoría* 'the majority', *iguales* 'the same', *pobres* 'poor', *latinoamericanos* 'Latin Americans', *peores* 'worse'.

Geography

- *Mi ⟨mask⟩ es un Hyundai Accent* 'My ⟨mask⟩ is a Hyundai Accent': *coche* (Spain's word for) 'car', *carro* (Latin America's word for) 'car', *vehículo* 'vehicle', *moto* 'motorbike', *padre* 'father'.
- *Llego tarde, tengo que ⟨mask⟩ el autobús* 'I am running late, I have to ⟨mask⟩ the bus': *coger* 'take' (Spain) / 'to have sex' (Latin America), *tomar* 'take' (Latin America), *evitar* 'avoid', *abandonar* 'abandon', *utilizar* 'utilize'.
- *Para llegar a mi casa, tengo que ⟨mask⟩ mi coche* 'In order to get home, I have to ⟨mask⟩ my [Spain's word for] car': *conducir* 'drive' (Spain), *alquilar* 'rent', *llevar* 'bring', *coger* 'take' (Spain) / 'to have sex' (Latin America), *aparcar* 'park'.
- *Para llegar a mi casa, tengo que ⟨mask⟩ mi carro* 'In order to get home, I have to ⟨mask⟩ my [Latin America's word for] car': *llevar* 'bring', *comprar* 'buy', *tener* 'have', *cargar* 'load', *conducir* 'drive' (Spain).
- *Para llegar a mi casa, tengo que ⟨mask⟩ mi auto* 'In order to get home, I have to ⟨mask⟩ my [Argentina's word for] car': *llevar* 'bring', *tener* 'have', *conducir* 'drive' (Spain), *coger* 'take' (Spain) / 'to have sex' (Latin America), *cargar* 'load'.

Appendix: Reproducibility

To reproduce the results in this paper, please, refer to the next code repository: `https://github.com/bertin-project/bertin-roberta`

Appendix: Availability

A demo of the BERTIN language model can be found online at `https://huggingface.co/spaces/bertin-project/bertin`. All source code is available under an Apache License 2.0. The language model and several fine-tuned versions are also available:

- BERTIN language model: `https://huggingface.co/bertin-project/bertin-roberta-base-spanish`
- BERTIN fine-tuned for NER: `https://bertin-project/bertin-base-ner-conll2002-es`
- BERTIN fine-tuned for POS: `https://bertin-project/bertin-base-pos-conll2002-es`
- BERTIN fine-tuned for XNLI: `https://bertin-project/bertin-base-xnli-es`
- BERTIN fine-tuned for PAWS-X: `https://bertin-project/bertin-base-paws-x-es`

We released the code to sample from mC4 on the fly when streaming for any language under the dataset in `https://huggingface.co/datasets/bertin-project/mc4-sampling`. In the `mc4-es-sampled` dataset (`https://huggingface.co/datasets/bertin-project/mc4-es-sampled`), the train split contains the full 50M samples, while validation is retrieved as it is from the original `mC4`.

Procesamiento del Lenguaje Natural, Revista nº 68, marzo de 2022, pp. 25-37 recibido 18-12-2021 revisado 20-01-2022 aceptado 26-01-2022

Depression Recognition in Social Media based on Symptoms' Detection

Reconocimiento de depresión en redes sociales basado en la detección de síntomas

Itzel Tlelo-Coyotecatl, Hugo Jair Escalante, Manuel Montes-y-Gómez
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Puebla, Mexico
{itlelo, hugojair, mmontesg}@inaoep.mx

Abstract: Depression is a common mental disorder that affects millions of people around the world. Recently, several methods have been proposed that detect people suffering from depression by analyzing their language patterns in social media. These methods show competitive results, but most of them are opaque and lack of explainability. Motivated by these problems, and inspired by the questionnaires used by health professionals for its diagnosis, in this paper we propose an approach for the detection of depression based on the identification and accumulation of evidence of symptoms through the users' posts. Results in a benchmark collection are encouraging, as they show a competitive performance with respect to state-of-the-art methods. Furthermore, taking advantage of the approach's properties, we outline what could be a support tool for healthcare professionals for analyzing and monitoring depression behaviors in social networks.
Keywords: Depression detection, social media, information retrieval.

Resumen: La depresión es un trastorno mental que afecta a millones de personas en todo el mundo. Recientemente, se han propuesto varios métodos que detectan personas que sufren depresión analizando sus patrones de lenguaje en las redes sociales. Estos métodos han mostrado resultados competitivos, sin embargo la mayoría son opacos y carecen de explicabilidad. Motivados por estos problemas, e inspirados en los cuestionarios utilizados por los profesionales de la salud para su diagnóstico, en este trabajo proponemos un método para la detección de depresión basado en la identificación y acumulación de evidencia de síntomas a través de las publicaciones de los usuarios. Los resultados obtenidos en una colección de referencia son prometedores, ya que muestran un desempeño competitivo con respecto a los mejores métodos actuales. Además, aprovechando las propiedades del método, describimos lo que podría ser una herramienta de apoyo para que los profesionales de la salud analicen y monitoreen las conductas depresivas en las redes sociales.
Palabras clave: Detección de depresión, redes sociales, recuperación de información.

1 Introduction

The World Health Organization (WHO) defines mental health as a state of emotional, psychological and social well-being that influences the way a person thinks, feels, acts or relates to others (World Health Organization, 2003). Accordingly, mental disorders refer to conditions that may affect the way of thinking, feeling or acting of persons. Among mental disorders, depression is one of the most common, affecting around 3.4% of the world's population (Saloni Dattani and Roser, 2021).

The diagnosis of depression is usually carried out by mental health professionals through the application of interviews or questionnaires focused on identifying the presence of certain symptoms (National Institute of Mental Health, 2021). These diagnostic methods are effective, but their coverage is limited mainly due to economic factors and social stigmatization (World Health Organization, 2003). These drawbacks, together with the need to address this growing problem, have motivated the development of computational tools for the automatic detection and monitoring of people suffering from depression. Particularly, the link between lan-

guage usage and the psychological state of people (Pennebaker, Mehl, and Niederhoffer, 2003) has led to the exploration of data from social networks for the automatic detection of depression, aiming to take advantage of the large amount of information generated by people through these media, in which they usually express their interests, concerns and feelings (Guntuku et al., 2017).

Current automatic methods usually address the depression detection task as a text classification problem, considering all the information shared by users, without necessarily adopting the traditional methodology that emphasizes the identification and measurement of symptoms. Most of these methods have achieved competitive results in benchmark collections (Losada, Crestani, and Parapar, 2018) (Coppersmith et al., 2015), but have also shown limitations in terms of transparency. Given the importance of the explainability of the decisions in this very sensitive task (Danilevsky et al., 2020) (Ríssola, Aliannejadi, and Crestani, 2020), we presume that the design of methods based on the identification of evidence of symptoms through the users' post, similar to the traditional diagnostic approach, could considerably improve the interpretation of the results. This paper describes a contribution in such direction.

In particular, in this work we propose an approach for depression detection in social media based on the identification and accumulation of evidence of symptoms. This approach has three main stages. In the first one users' posts are filtered, keeping only those that refer to or are related to any of the 21 symptoms declared on Beck's Depression Inventory (BDI)[1]. Then, in the second stage, 21 independent classifiers are built from the sets of filtered posts. The idea is that each classifier observes the target user from a different perspective, determining whether she or he suffers from depression or not according to the presence of only one of the symptoms. Finally, in a third stage, the decisions of the different classifiers are combined to generate a final unified prediction. Through this three-stage process, which gradually identifies and integrates evidence of the different symptoms, we move a step forward by facilitating the interpretation of decisions, and thereby enabling its usage in social media monitoring applications.

Summarizing, the main contributions of this work are:

- We propose a new approach for depression detection in social media based on the analysis of the presence of depression symptoms through users' posts.
- We carry out an in-depth analysis of the presence of the different symptoms in users' posts and their correlation with the classification errors, providing insights on their relevance for the detection of depression in social media.
- We outline a simple interface to support the detection and follow-up of users who suffer from depression.

The remainder of the paper is organized as follows. Section 2 presents a brief overview of related work on depression detection in social media. Section 3 describes the proposed approach for depression detection based on the symptoms identification and accumulation. Sections 4 and 5 reports the experiments, results, and their analysis. Section 6 outlines what could be a support tool for healthcare professionals to analyze and monitor depression behaviors in social networks. Finally, Section 7 points out our conclusions and future work.

2 Related Work

As we previously mentioned, the detection of depression in social media has been handled as a supervised learning problem, where the main goal is to build a model that distinguishes users suffering from depression from healthy users (Guntuku et al., 2017).

Most current methods rely on the use of traditional machine learning processes or deep learning techniques. On the one side, there are works like the ones from Nadeem (2016), Jamil et al. (2017) and Preoţiuc-Pietro et al. (2015) which consider a bag-of-words or word n-grams as the users' representation, and employ traditional learning algorithms to build the classifier. This kind of methods are very popular due to their low computational complexity and the easiness for interpreting their results related with depressive tendencies (Tsugawa et al., 2013).

[1]The considered version of the BDI we used corresponds to the provided at the CLEF eRisk2019 available at https://early.irlab.org/2019/index.html

High dimensionality is a known problem of bag-of-words approaches, which has prompted the use of manually or automatically-defined topic-based representations as an alternative. Examples of this are the works by Wolohan et al. (2018) that uses the LIWC categories as features, and by Loveys et al. (2018) that carry out an analysis of the language usage involving cultural differences for depression considering the LIWC categories, topic modeling, data visualization, and other techniques.

From another perspective, there are studies that have considered the use of information about sentiments and emotions to analyze depression behaviours in social media (Preoţiuc-Pietro et al., 2015), (De Choudhury et al., 2013), and (Aragon et al., 2021). These works have shown interesting results, indicating that negative posts, as well as certain emotions like anger and fear, are more abundant in people with depression than in users who do not suffer from this disorder.

On the other hand, some recent successful works have approached the combination of hand-crafted and automatically learned representations using deep learning models, such as CNNs or LSTMs (Liu et al., 2018), (Husseini Orabi et al., 2018), (Yates, Cohan, and Goharian, 2017), (Trotzek, Koitka, and Friedrich, 2018a). Despite their good results, this kind of methods have important drawbacks. For example, they require large amount of data to train their models, have high complexity, and consequently low explainability. Moving in the latter direction, the works by Mathur et al. (2020), Trifan et al. (2020) and Ríssola, Aliannejadi, and Crestani (2020) have integrated psycholinguistic features into the users' representations. Their common idea is to build effective methods, but also capable of providing understanding and descriptions of the decisions made (Burdisso, Errecalde, and Montes-y-Gómez, 2019), (Zogan et al., 2020).

Lastly, the eRisk evaluation forum[2] exemplifies the evolution that this task has had in the recent years. In its last editions (Losada, Crestani, and Parapar, 2020), (Parapar et al., 2021), it has included a subtask that consists of estimating the level of depression from a thread of user posts, in other words, of filling a standard depression questionnaire based on the evidence found in the history of postings. From these evaluations, its organizers have concluded that "*Although the effectiveness of the proposed solutions is still modest, the experiments suggest that evidence extracted from social media is valuable, and that automatic or semi-automatic screening tools could be designed to detect at-risk individuals*". Our work follows precisely this research direction. However, we approach it again as a binary classification problem, aimed at distinguishing users who suffer from depression from healthy users, but we adopt the idea of identifying and accumulating evidence of depression symptoms, and even more, we outline the proposal of a support tool for analyzing and monitoring depression behaviors in social networks.

3 Depression Detection based on Symptoms Evidence

This section describes our proposed method for depression detection based on social media content. Figure 1 shows its general diagram, which comprises three main stages:

1. *Symptoms evidence retrieval*, where the aim is to identify evidence associated to depression symptoms from users' posts.
2. *Symptom-based classification*, where we build predictive models for distinguishing healthy users from users suffering from depression based on the presence of each one of the symptoms.
3. *Depression status prediction*, where we combine the evidence of the identified symptoms to make a final prediction on the presence or absence of depression.

In the following subsections we detail each stage accordingly.

3.1 Symptoms Evidence Retrieval

The first stage of the proposed method consists of identifying candidate posts that can be considered evidence for the 21 BDI declared symptoms. BDI is a standard questionnaire composed by 21 questions, each one related to one possible depression symptom, applied on traditional depression diagnosis detection made by professionals[3] (Beck et al., 1961). Table 1 lists these symptoms.

[2]Early Risk Prediction on the Internet (eRisk) website: https://erisk.irlab.org/

[3]Alternative questionnaires based on symptoms identification for depression detection could also be considered (e.g., PHQ-9).

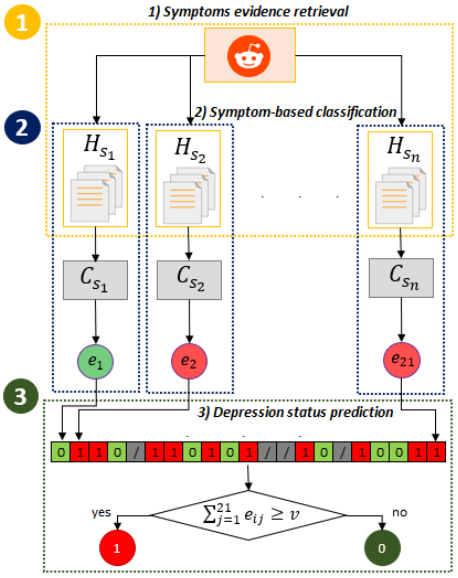


Figure 1: Diagram of the proposed method for depression detection based on symptoms identification.

We approach this task as one of information retrieval, where we want to retrieve relevant posts (*documents*) given the different symptoms (*queries*). More specifically, given a user's history made up of a set of posts $P = \{p_1, p_2, ..., p_n\}$, and the set of symptoms $S = \{s_1, s_2, ..., s_{21}\}$, the goal is to determine whether a post p_i is considered evidence (i.e., it is *relevant*) for each of the symptoms s_j with $j \in [1, 21]$.

For the retrieval process we relied on word embeddings and a similarity threshold. In preliminary work we evaluated other information retrieval models but the one reported herein resulted in better retrieval performance. This process is as follows:

Each symptom s_j is associated to a set of keywords $s_j = \{ws_1, ws_2, ..., ws_l\}$ with $j \in [1, 21]$ and $l \in [1, 3]$, according to Table 1[4]. In the same way, for each post p_i its set of words corresponds to $p_i = \{wp_1, wp_2, ..., wp_m\}$ where i indicates the number of post in the user history.

We say a post p_i is considered evidence for a symptom s_j (i.e., it is relevant) if there exist a tuple of words (wp_i, ws_j) whose cosine similarity in a particular embedding space is above a threshold β. The threshold β is a parameter that should be fixed depending on how strict one wants to be on the similarity of words for determining relevance between posts and symptoms. As a result of this retrieval process we can identify posts associated to the symptoms, where each post can be associated to none or any of the 21 symptoms. The process is applied to all the posts on the user history and all the symptoms on the BDI for all the users to analyze. In this way, each user u_i is represented by the evidence sets $\mathbb{H}_i = \{H_{i1}, H_{i2}, ..., H_{i21}\}$, one set H_{ij} per symptom.

In the next subsection we describe the way this information is used for building predictive models to automatically detect the presence of symptoms. One should note that the retrieved evidence on the presence/absence of symptoms could be also used for interpretability or explainability purposes, see Section 6.

3.2 Symptom-based Classification

This stage considers the previous retrieved evidence for each symptom. The aim is to build a predictive model per symptom using the identified evidence, that is, to build 21 classifiers that observe a target user from different perspectives and determine whether she/he suffers from depression or not.

The construction of each classifier c_j follows the traditional text-classification approach. That is, given a set of labeled users $U = \{(u_1, y_1), (u_2, y_2), ..., (u_m, y_m)\}$, and $y_i \in \{0, 1\}$, where $y_i = 0$ indicates a healthy user and $y_i = 1$ one suffering from depression, the goal is to learn a function $c_j : u \rightarrow \{0, 1\}$. For that purpose, we represent the users through a BOW model with $tfidf$ weights. All 21 classifiers are trained over the same set of labeled users, but differ in how these are represented. That is, for the construction of the classifier c_j only the posts corresponding to the evidence sets H_{kj} for all users $u_k \in U$ are used, and, therefore, it learns to recognize evidence associated to the symptom s_j.

Once the 21 classifiers are trained on the different symptoms, they can be used to make predictions about the presence or absence of depression in a new user u_i. In the case where there is no evidence about symptom s_j in the user's post history (i.e., $H_{ij} = \emptyset$), then the classifier c_k returns a void value (indicated as "/" in Figure 1.

The information provided by predictive models built in this stage is combined in order to give a final prediction for each subject,

[4]These keywords were manually chosen by considering those that represent at best the symptoms' contexts after narrowing down their lists of synonyms.

#	Symptom (BDI)	Keywords	#	Symptom (BDI)	Keywords
s01	Sadness	sadness	s12	Loss of interest	apathetic, worthless
s02	Pessimism	pessimism	s13	Indecisiveness	indecisiveness, indecisive
s03	Past failure	failure	s14	Worthlessness	worthlessness, worthless
s04	Loss of pleasure	displeasure, dissatisfaction	s15	Loss of energy	apathetic, dispirited
s05	Guilty feelings	guilty	s16	Changes in sleep pattern	sleep
s06	Punishment feelings	punishment	s17	Irritability	irritability
s07	Self-dislike	dislike, self	s18	Changes in appetite	appetite
s08	Self-criticalness	criticalness, critical, self	s19	Concentration difficulty	disconcerted, concentration
s09	Suicidal thoughts or wishes	suicidal	s20	Tiredness or fatigue	tiredness, fatigue
s10	Crying	crying	s21	Loss of interest in sex	sex, disinterest
s11	Agitation	agitation			

Table 1: 21 declared symptoms on the Beck's Depression Inventory (BDI).

as described in the next subsection.

3.3 Depression Status Prediction

This stage comprises the integration of the obtained predictions from the different symptom's classifiers. For that purpose, in a first step, these predictions are concatenated on a vector that is considered as the evidence-based representation for the user under analysis. After this process, each user u_i is represented by a vector $\mathbf{e}_i = \langle e_{i1}, e_{i2}, ..., e_{i21} \rangle$, where each e_{ij} indicates the prediction of classifier c_j on user u_i. Then, the final step of the proposed approach aims at providing a global prediction on the detection of depression for each user. Since the predictions vector $\mathbf{e}$ already captures the symptoms' presence, we process such a vector to obtain a final prediction. A number of ways for delivering a prediction from $\mathbf{e}$ were studied, including, standard ensembles, stacking generalization and meta-classifiers. However, we found that the most effective way was based on thresholding the number of symptoms detected by the distinct classifiers from stage 2.

To assign the final class label to a user u_i a vote counting is done over the values of the evidence vector $\mathbf{e}_i$. That is, if $\sum_{j=1}^{21} e_{ij} \geq v$, then the user u_i is classified as suffering from depression, otherwise he or she is marked as a healthy user. The votes are interpreted as: at least v symptom classifiers should be positive in order to classify a user in the depressive class. Therefore, the increasing of v means more symptoms should be positive in order to declare a user as depressed.

4 Experimental Settings

This section presents the corpus used in the experiments, and describes the experimental setup and the baseline results considered.

4.1 Dataset

The experiments were carried out on the eRisk-2018 corpus for Task 1 "Early Detection of Signs of Depression" (Losada, Crestani, and Parapar, 2018). This corpus contains English writings of Reddit users belonging to two categories, depressed and non-depressed users. The first group of users explicitly expressed in one of their posts that they were *diagnosed* with depression. On the other hand, the second group is formed by randomly selected users from the Reddit platform. Table 2 shows the distribution of the two categories of users in the given corpus. It is worth noting that the corpus presents a high class imbalance, which has motivated the use of the F_1 score over the positive class as the main evaluation measure.

Category	Train	Test
Depressed (D)	135	79
Non-depressed (ND)	752	741
Total	887	880

Table 2: Categories distribution in the eRisk-2018 corpus.

Besides being used in the aforementioned evaluation campaign (Losada, Crestani, and Parapar, 2018), several authors have reported results on it. In the next section we compare the performance of our method with some of those references.

4.2 Method Configuration

Text preprocessing. It was performed using NLTK packages; we normalized the texts by lowercasing all words, removing the stopwords, links, numbers and special characters.

Evidence retrieval. We employed pre-trained Twitter Glove embeddings[5] of 100-dimensions. In addition, we considered different values for the β-threshold (from 0.1 to 0.9), but the best results were obtained with $\beta = 0.6$, hence we use this value for all of our experiments.

Symptom-based classification. The 21 symptom classifiers were built in the same way. In all cases we used a SVM with a linear kernel, $C = 1$, L2 normalization, and weighted class imbalance.

Prediction threshold. For the third stage we followed an exploratory approach, considering different values for the v-parameter, which thresholds the number of votes from individual classifiers. Particularly, we used values from 1 to 11 (majority vote).

4.3 Reference Approaches

The following reference models were used for comparison.

Traditional. The *complete* user histories are represented using a BoW model with TF-IDF weights. These are fed to a SVM classifier with the same configuration as our symptom-based classifiers.

LIWC. It uses a representation that combines the LIWC categories and a BOW representation, with the 3,000 words with the greatest chi^2 values. It also uses a SVM classifier with the same configuration as our symptom-based classifiers (Aragon et al., 2021).

BiLSTM and CNN. Both neural networks used 100 neurons, the adam optimizer, and GloVe embeddings of 300-dimensions. For the CNN, 100 random filters of sizes 1, 2 and 3 were applied (Aragon et al., 2021).

BoSE. It uses a histogram of fine-grained emotions as representation, which captures the presence and variability of emotions through the users' posts. It employs a SVM as a classifier (Aragon et al., 2021).

DPP-EXPEI-SVM. It uses a BOW as representation, but considering a weighting scheme that rewards the presence of personal information in the posts. It employes a SVM as classifier (Ortega-Mendoza et al., 2022).

[5]Twitter Glove embeddings were chosen due to their orientation towards social network language, nonetheless alternative versions could be considered. GloVe embeddings were obtained from: https://nlp.stanford.edu/projects/glove/

Best results at eRisk2018. We consider the two best reported results at the eRisk-2018 forum (Losada, Crestani, and Parapar, 2018). Both results correspond to variations of the same method, named as FHDO-BCSGA and FHDO-BCSGB respectively (Trotzek, Koitka, and Friedrich, 2018b). They consider the results of machine learning models, together with an ensemble model that combined different base predictions. The models employ user-level linguistic metadata, bag of words, neural word embeddings, and convolutional neural networks.

5 Results

In this section we present the results of experiments that aim to evaluate the proposed method, and to compare its performance with that of the state of the art.

5.1 Symptom-based Evaluation

The goal of this experiment was to evaluate depression detection performance by considering evidence from a single symptom. That is, we evaluate the performance of the 21 per-symptom classifiers c_j. Results of this experiment are shown in Figure 2.

Obtained results show that evidence obtained from some symptoms is more discriminative than that from others. For example, the classifier for the symptom s_3 : *past failure* achieved an F_1 measure value close to 0.6 by itself. However, the performance for most symptoms is much lower, even a couple of them (s_{13}*: indecisiveness* and s_{11}*: agitation*) obtained a value of $F_1 = 0$.

The values at the right of Figure 2 show the average amount of information retrieved for each of the symptoms. There is a clear correlation between the amount of retrieved evidence and the performance of classifiers with few exceptions (e.g., s_9 : *suicidal thoughts or whishes* and s_6 : *punishmetn feelings*). In general, the results of this experiment indicate that the performance obtained with the different classifiers is related to the amount and quality of information available in the users' histories for training them.

Results shown in Figure 2 show that it is possible to detect depression by using information from a single symptom. However, the performance of such individual classifiers is still low when compared to reference methods. In the next section we show that by

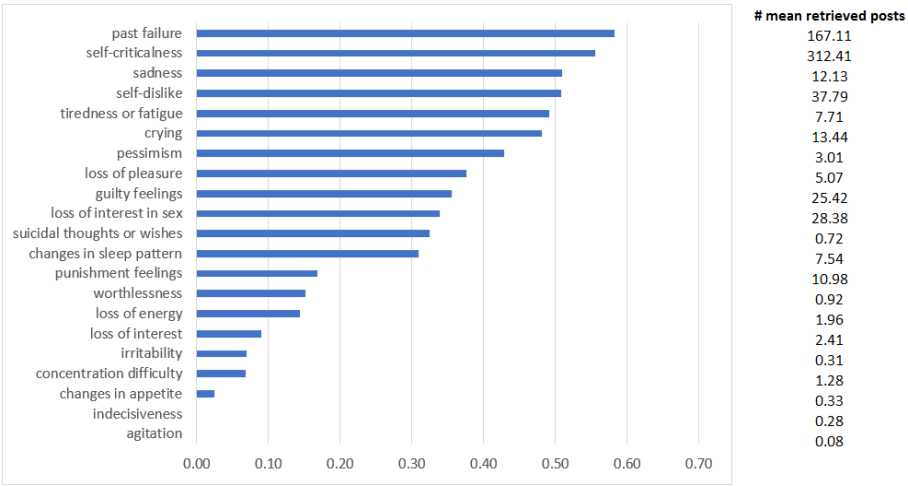


Figure 2: Evaluation F_1-score results for each one of the 21 BDI symptoms. The values on the right indicate the average retrieved posts for each symptom.

combining the information from these models it is possible to boost the performance.

5.2 Combining Evidences

Although the individual performance of symptom's classifiers is not that competitive, an hypothesis of this paper is that by combining the acquired evidence from all of the symptoms we could obtain better performance. This section aims to evaluate the combination of the predictions of individual models.

As described in Section 3.3, we aggregated the evidence obtained by the prediction vector $\mathbf{e}_i$, which is formed by the 21 predictions of individual classifiers for subject u_i, and used a threshold to determine whether the subject is detected as depressed (i.e., $\sum_{j=1}^{21} e_{ij} \geq v$) or not (otherwise). In the experiments we evaluated values for v lower than 11, since as reaching $v = 11$ votes from the individual classifiers means that more than half of the symptoms were detected from users' texts. We refer to this method as *ensemble of all symptoms* (SBC-ALL).

In addition, motivated by our previous results (see Section 5.1) that show that several symptoms did not present relevant or sufficient evidence for the detection of depression, we carried out an experiment combining only the decision of the best-half of the classifiers, selected according to their precision scores[6]. We named it as *ensemble of top symptoms* (SBC-TOP).

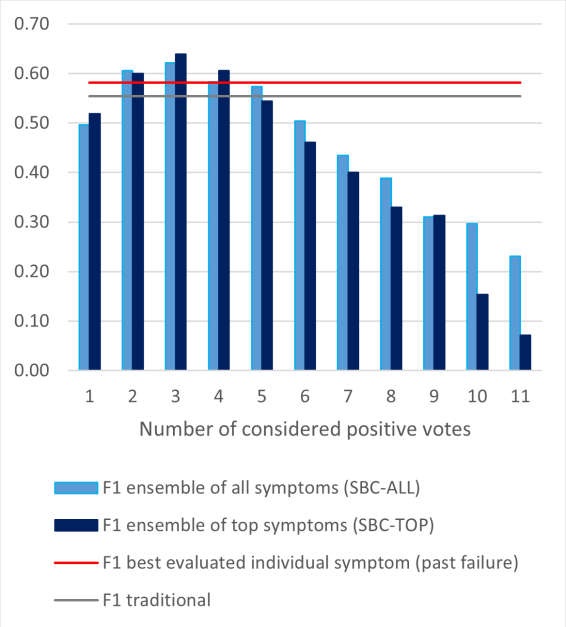


Figure 3: Reported F_1 results when combining the evidence of the individual models. We report the performance obtained when varying the threshold (x-axis) on the number of positively detected depression across the individual classifiers.

Figure 3 reports the results obtained with both ensemble approaches. In order to analyze the performance of these methods in detail, we report the obtained performance when varying the threshold on the number of positive outputs of the individual classi-

[6]The following symptom classifiers were selected for the SBC-TOP ensemble: *sadness, pessimism, past failure, loss of pleasure, guilty feelings, self-dislike, self-criticalness, crying, changes in sleep pattern, tiredness or fatigue* and *loss of interest in sex.*

fiers to detect depression with each ensemble. They indicate that it is slightly better to consider only the subset of the best individual symptoms than to consider all of them (i.e., SBC-TOP outperforms SBC-ALL). For both approaches detecting few positive symptoms ($v \leq 5$) allows achieving a high recall but at the expense of low precision levels, whereas increasing the number of symptoms identified as positive ($v \geq 6$) helps improving the precision but greatly affects the recall, which caused very low F_1 scores. Hence, the choice of v can be done according to what the final user judges as most important, either precision or recall.

Finally, it should be noted that the presence of only 3 symptoms was sufficient to obtain results that improved the results from the traditional baseline approach as well as from the best individual symptom classifier, gray and red horizontal lines respectively.

5.3 Comparison with the State of the Art

This section compares the performance of the best result obtained with our proposed method and reference techniques. For this experiment we consider the F_1-score on the positive class (i.e., depression) as evaluation measure. Table 3 compares the results of both ensemble approaches against all the methods described in Section 4.3.

Method	**P**	**R**	F_1
Traditional	0.565	0.544	0.554
LIWC	-	-	0.380
BiLSTM-GloVe	-	-	0.460
CNN-GloVe	-	-	0.510
BoSE	0.670	0.610	**0.640**
DPP-EXPEI-SVM	0.570	0.660	0.610
FHDO-BCSGB	0.640	0.650	**0.640**
FHDO-BCSGA	0.560	0.670	0.610
SBC-ALL	0.667	0.582	0.622
SBC-TOP	0.707	0.582	0.638

Table 3: Results for precision (P), recall (R) and F_1-score over the positive class of state of the art methods.

Results from Table 3 show some interesting points about the proposed approach. On the one hand, when comparing its results against the traditional baseline, it is evidenced the noisiness of the users' posts and, thus, the relevance of their filtering with respect to the presence of the depression symptoms. On the other hand, these results also show that methods based exclusively on neural network architectures are not the best alternative given the scarcity of training resources. In contrast, our approach better handles this complex scenario, by combining the decision of different independent classifiers, one for each symptom. Finally, it is observed that our two ensembles did not outperform the best result at the eRisk-2018 evaluation task, nonetheless, it is important to note that a comparable result was achieved using computationally less expensive methods, and even more important, that the proposed approach, due to its three-step architecture, facilitates the interpretation and explanation of the results, something critical in this type of applications.

5.4 Analysis and Discussion

This section presents additional experiments and results that aim to explain the performance obtained by the proposed solution, as well as highlighting its main benefits.

5.4.1 Maximum Possible F_1

The hypothesis behind our approach is that different users suffering from depression would show and express different symptoms through their posts. In consequence, by integrating the decisions of several independent symptom-based classifiers it would be possible to achieve high detection rates. The previous results were somewhat discouraging compared to this initial intuition, as the result of our best ensemble (SBC-TOP) only surpassed the best individual result (s_3 : *past failure*, see Figure 2) by around 5%.

In order to determine the potential of our initial idea, we calculated the performance obtained by a (hypothetical) perfect ensemble of the 21 symptom-based classifiers. To simulate this perfect ensemble we considered an user as correctly classified if at least one of the symptom-based classifiers did so (i.e., at least one of the c_j models returned a positive output). The results obtained by such a hypothetical ensemble was $F_1 = 0.66$, only 2% higher than that the result obtained by SBC-TOP, indicating that most of the symptoms allow to identify more or less the same group of depressed users, and they are not complementary to each other. From this result, we can conclude that the main direction of research should be in the improvement of the symptom-based classifiers and not in the

integration of their decisions.

5.4.2 More and Less Depressed Users

In this section we evaluate the performance of a simple model that approaches the depression detection problem as one of information retrieval. The goal is to verify whether users that trigger more symptoms have more chances to be depressed than the ones that present evidence for less or non symptoms.

We ranked users in descending order of the number of detected symptoms and evaluated the precision as a retrieval task (i.e., evaluating the ranked list with relevancy given by the ground truth label). Specifically, the ranking was evaluated with the precision at the k^{th} position ($P@K$). The obtained precision values were as follows: P@10=0.70, P@20=0.75, P@30=0.73. We think these are positive results, as among the top 10, 20 and 30 ranked users the majority were labeled as depressed. This also reinforces evidence on that the higher the number of detected symptoms the higher the chances of the depression category[7].

5.4.3 More and Less Informative Symptoms

In an attempt to further analyze the relevance of each of the symptoms considered, we measured the occurrence of the symptoms in the posts from positive users (i.e., users classified as suffering from depression). Figure 4 summarizes the results of this analysis, showing for each symptom the difference of their occurrences in true-positive and false-positive instances. These results indicate that the symptoms that occur the most in users correctly classified as suffering from depression are s_3 : *past failure*, s_8 : *self-criticalness* and s_7 : *self-dislike*. On the other hand, the symptoms whose presence caused most classification errors are s_{15} : *loss of energy*, s_{19} : *concentration difficulty* and s_{12} : *loss of interest*. These results are not surprising at all, because the first group of symptoms refers to the perception of users about themselves, while the symptoms of the second group refer to more generic moods and problems, which are also widely mentioned by users who do not suffer from depression.

[7]For your reference, we also report the value $P@79 = 0.58$, with 79 being the number of positive cases in the test set.

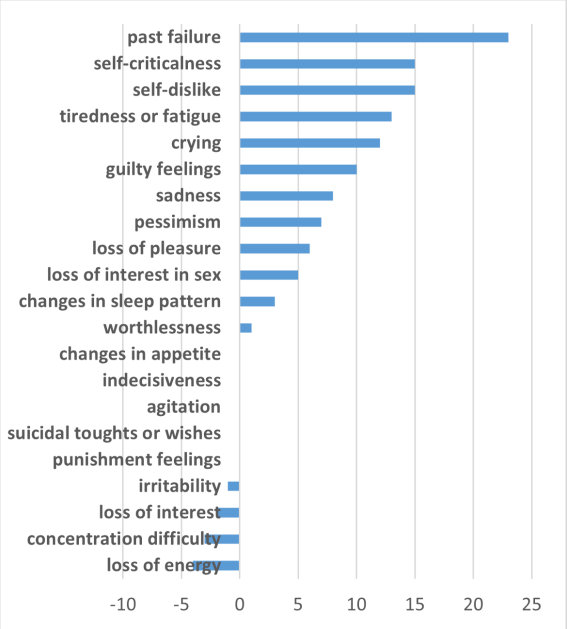


Figure 4: More to less informative symptoms arranged in descending order.

6 *Towards a Monitoring Tool*

The depression detection task has been approached with automatic methods achieving competitive performance. However, some of them are obscure, in the sense that it is not clear what motivates the recommendations of the models, or what information is the most informative for models. Our proposed solution is inspired by the traditional diagnosis process (i.e, through the application of questionnaires), and it is able to generate rich intermediate information, which gives our proposed solution a clear advantage when compared to other models. Here we outline some ways in which one could take advantage of the information that is generated by the proposed solution for explainability and interpretability purposes (see Figure 5) :

- **Thermometer of depression level.** Would allow us to visualize an estimate of the level of depression of a certain user according to the number of detected symptoms by the individual models.
- **List of identified symptoms.** Would allow us to visualize the list of symptoms that were identified as positive in a user.
- **List of evidence posts.** Would allow us to visualize posts that are considered evidence for a selected symptom.

These components and several others

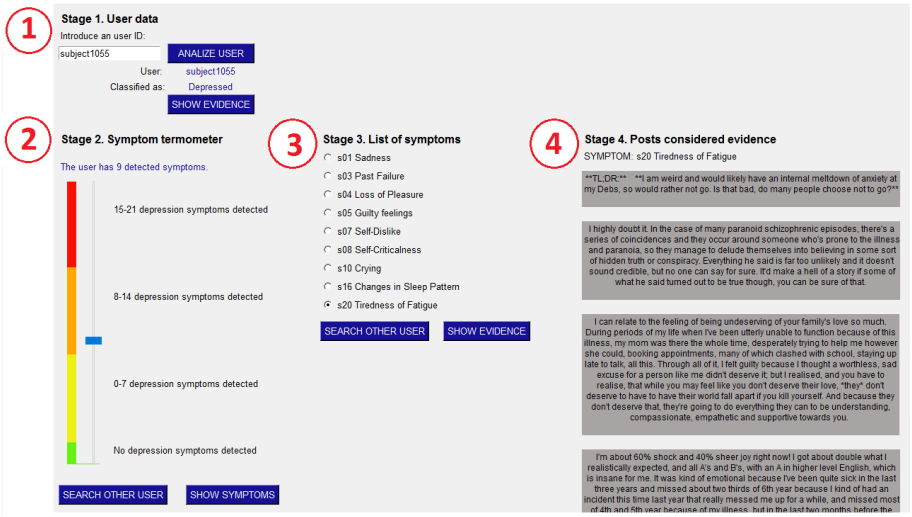


Figure 5: Example of the proposal of a support tool for decision making that takes advantage of the generated information by the proposed method.

could be put together in a decision support tool that could be helpful to psychologist or even the average user to have an idea on her/his potential level of depression. In the remainder of this section we provide a description of a possible monitoring support tool implementing the aforementioned components.

6.1 Running example

Figure 5 shows an screenshot on how the support tool may look when analyzing a particular subject. As a first stage the user data is introduced, for this example *subject1055* from the considered dataset is used, this user is declared as Depressed. The second stage shows the number of symptoms that the user has. In this case for *subject1055* nine symptoms were detected. According to the thermometer scale 9 symptoms correspond to an orange level that can be translated as a medium depression level. The third stage shows a list with the specific name and numbers of the previously detected symptoms. In this case the user presents the symptoms: *sadness, past failure, loss of pleasure, guilty feelings, self-dislike, self-criticalness, crying, changes in sleep pattern* and *tiredness or fatigue*. In the fourth stage, the most relevant posts to symptoms selected from the list are displayed.

Although this is just a sketch of a possible solution, we firmly believe that a tool like this could be very helpful to disentangle and understand the recommendations of the proposed model.

7 *Conclusion & Future Work*

We proposed a novel depression detection method based on the identification of the 21 declared symptoms from the BDI. The method first retrieves evidence related to the symptoms from the users' posts. These evidence is then passed through symptom-detectors, whose outputs are combined to provide a final prediction. The proposed method obtained competitive results when compared with the best eRisk2018 reported results. Even when the proposed method did not outperform the state of the art, its additional benefits compensate for the small performance difference with more sophisticated methods. In particular, the possibility of providing explanations on the predictions of the model, and to interpret the model functioning are its most notable advantages. Future work includes the improvement of the evidence retrieval stage by adopting more elaborated models and by improving the description of symptoms (e.g., exploring other embedding versions, using contextualized text representations or applying pseudo-relevance feedback). Likewise, we are implementing the decision support tool into a demo that could be publicly available to anyone. On the other hand, we plan to carry out a quantitative evaluation of the interpretability of the model applying models such as LIME or SHAP.

References

Aragon, M. E., A. P. Lopez-Monroy, L.-C. G. Gonzalez-Gurrola, and M. Montes. 2021. Detecting mental disorders in social media through emotional patterns - the case of anorexia and depression. *IEEE Transactions on Affective Computing*, pages 1–1.

Beck, A. T., C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh. 1961. An inventory for measuring depression. *Archives of general psychiatry*, 4(6):561–571.

Burdisso, S. G., M. Errecalde, and M. Montes-y-Gómez. 2019. A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications*, 133:182–197.

Coppersmith, G., M. Dredze, and C. Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Coppersmith, G., M. Dredze, C. Harman, K. Hollingshead, and M. Mitchell. 2015. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

Danilevsky, M., K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen. 2020. A survey of the state of explainable ai for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

De Choudhury, M., M. Gamon, S. Counts, and E. Horvitz. 2013. Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1):128–137.

Guntuku, S. C., D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Husseini Orabi, A., P. Buddhitha, M. Husseini Orabi, and D. Inkpen. 2018. Deep learning for depression detection of Twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97, New Orleans, LA. Association for Computational Linguistics.

Jamil, Z., D. Inkpen, P. Buddhitha, and K. White. 2017. Monitoring tweets for depression to detect at-risk users. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 32–40, Vancouver, BC. Association for Computational Linguistics.

Kuncheva, L. and C. Whitaker. 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51:181–207.

Liu, N., Z. Zhou, K. Xin, and F. Ren. 2018. TUA1 at eRisk 2018. In *Working Notes of CLEF 2018—Conference and Labs of the Evaluation Forum*.

Losada, D., F. Crestani, and J. Parapar, 2020. *Overview of eRisk 2020: Early Risk Prediction on the Internet*, pages 272–287. 09.

Losada, D. E., F. Crestani, and J. Parapar. 2017. Clef 2017 erisk overview: Early risk prediction on the internet: Experimental foundations. In *CLEF*.

Losada, D. E., F. Crestani, and J. Parapar. 2018. Overview of erisk 2018: Early risk prediction on the internet (extended lab overview). In *CLEF*.

Losada, D. E., F. Crestani, and J. Parapar. 2019. Overview of erisk at clef 2019: Early risk prediction on the internet (extended overview). In *CLEF*.

Loveys, K., J. Torrez, A. Fine, G. Moriarty, and G. Coppersmith. 2018. Cross-cultural differences in language markers of

depression online. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 78–87, New Orleans, LA. Association for Computational Linguistics.

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.

Mathur, P., R. Sawhney, S. Chopra, M. Leekha, and R. Ratn Shah. 2020. Utilizing temporal psycholinguistic cues for suicidal intent estimation. *Advances in Information Retrieval*, 12036:265–271.

Mowery, D. L., A. Park, C. Bryan, and M. Conway. 2016. Towards automatically classifying depressive symptoms from Twitter data for population health. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 182–191, Osaka, Japan. The COLING 2016 Organizing Committee.

Nadeem, M. 2016. Identifying depression on twitter. *ArXiv*, abs/1607.07384.

National Institute of Mental Health. 2021. Depression. NIH Publication No. 21-MH-8079. https://www.nimh.nih.gov/health/publications/depression.

Ortega-Mendoza, R. M., D. I. Hernández-Farías, M. M. y Gómez, and L. Villaseñor-Pineda. 2022. Revealing traces of depression through personal statements analysis in social media. *Artificial Intelligence in Medicine*, 123:102202.

Parapar, J., P. Martín-Rodilla, D. Losada, and F. Crestani, 2021. *Overview of eRisk 2021: Early Risk Prediction on the Internet*, pages 324–344. 09.

Pennebaker, J. W., M. R. Mehl, and K. G. Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1):547–577.

Preoţiuc-Pietro, D., J. Eichstaedt, G. Park, M. Sap, L. Smith, V. Tobolsky, H. A. Schwartz, and L. Ungar. 2015. The role of personality, age, and gender in tweeting about mental illness. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 21–30, Denver, Colorado. Association for Computational Linguistics.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Ríssola, E., M. Aliannejadi, and F. Crestani. 2020. Beyond modelling: Understanding mental disorders in online social media. *Advances in Information Retrieval*, 12035:296 – 310.

Saloni Dattani, H. R. and M. Roser. 2021. Mental health. *Our World in Data*. https://ourworldindata.org/mental-health.

Trifan, A., R. Antunes, S. Matos, and J. Oliveira. 2020. Understanding depression from psycholinguistic patterns in social media texts. *Advances in Information Retrieval*, 12036:402 – 409.

Trotzek, M., S. Koitka, and C. Friedrich. 2018a. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, 32:588–601.

Trotzek, M., S. Koitka, and C. Friedrich. 2018b. Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia. In *CLEF*.

Tsugawa, S., Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki. 2015. Recognizing depression from twitter activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 3187–3196, New York, NY, USA. Association for Computing Machinery.

Tsugawa, S., Y. Mogi, Y. Kikuchi, F. Kishino, K. Fujita, Y. Itoh, and H. Ohsaki. 2013. On estimating depressive tendencies of twitter users utilizing

their tweet data. In *2013 IEEE Virtual Reality (VR)*, pages 1–4, Los Alamitos, CA, USA. IEEE Computer Society.

Wolohan, J., M. Hiraga, A. Mukherjee, Z. A. Sayyed, and M. Millard. 2018. Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP. In *Proceedings of the First International Workshop on Language Cognition and Computational Models*, pages 11–21, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

World Federation for Mental Health. 2012. Depression: A global crisis. *Health WF for M, editor. Occoquan*, pages 1–32.

World Health Organization. 2003. Investing in mental health. https://apps.who.int/iris/handle/10665/42823.

World Health Organization. 2017. Depression and other common mental disorders: global health estimates. Technical documents.

Yates, A., A. Cohan, and N. Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

Zogan, H., X. Wang, S. Jameel, and G. Xu. 2020. Depression detection with multi-modalities using a hybrid deep learning model on social media. *CoRR*, abs/2007.02847.

Procesamiento del Lenguaje Natural, Revista nº 68, marzo de 2022, pp. 39-60 recibido 18-12-2021 revisado 26-01-2022 aceptado 05-02-2022

MarIA: Spanish Language Models

MarIA: Modelos del Lenguaje en Español

Asier Gutiérrez-Fandiño,[*1] Jordi Armengol-Estapé*,[1] Marc Pàmies,[1] Joan Llop-Palao,[1] Joaquín Silveira-Ocampo,[1] Casimiro Pio Carrino,[1] Carme Armentano-Oller,[1] Carlos Rodriguez-Penagos,[1] Aitor Gonzalez-Agirre,[1] Marta Villegas[1]
[1]Barcelona Supercomputing Center
marta.villegas@bsc.es

Abstract: This work presents MarIA, a family of Spanish language models and associated resources made available to the industry and the research community. Currently, MarIA includes RoBERTa-base, RoBERTa-large, GPT2 and GPT2-large Spanish language models, which can arguably be presented as the largest and most proficient language models in Spanish. The models were pretrained using a massive corpus of 570GB of clean and deduplicated texts with 135 billion words extracted from the Spanish Web Archive crawled by the National Library of Spain between 2009 and 2019. We assessed the performance of the models with nine existing evaluation datasets and with a novel extractive Question Answering dataset created ex novo. Overall, MarIA models outperform the existing Spanish models across a variety of NLU tasks and training settings.
Keywords: MarIA, Spanish language modelling, Spanish language resources, Benchmarking.

Resumen: En este artículo se presenta MarIA, una familia de modelos del lenguaje en español y sus correspondientes recursos que se hacen públicos para la industria y la comunidad científica. Actualmente MarIA incluye los modelos del lenguaje en español RoBERTa-base, RoBERTa-large, GPT2 y GPT2-large que pueden considerarse como los modelos más grandes y mejores para español. Los modelos han sido preentrenados utilizando un corpus masivo de 570GB de textos limpios y deduplicados, que comprende un total de 135 mil millones de palabras extraidas del Archivo Web del Español construido por la Biblioteca Nacional de España entre los años 2009 y 2019. Evaluamos el rendimiento de los modelos con nueve conjuntos de datos existentes y con un nuevo conjunto de datos de pregunta-respuesta extractivo creado ex novo. El conjunto de modelos de MarIA supera, en la practica totalidad, el rendimiento de los modelos existentes en español en las diferentes tareas y configuraciones presentadas.
Palabras clave: MarIA, Modelos de lenguaje del Español, Recursos de lenguaje del Español, Evaluación de modelos del lenguaje.

1 Introduction

In recent years, the field of Natural Language Processing (NLP) has seen a proliferation of massive pretrained language models. These have been proved to perform best when trained on language-specific data. However, the vast majority of these massive models have been trained for English, leaving other languages aside and increasing the existing gap between them. Spanish, despite being the second most spoken language in the world, lacks large language models trained with vast and high quality data. One of the objectives of the Plan-TL[1] is to cover this gap with the MarIA project.[2] MarIA aims to provide both the industry and the scientific community with large scale language models, massive high-quality corpora and evaluation sets for the Spanish language. We present four large models of varying sizes and configurations, and compare them to existing models in a wide range of NLP tasks, showing that these new models

[*] Equal contribution.

[1] https://plantl.mineco.gob.es/

[2] https://github.com/PlanTL-GOB-ES/lm-spanish

are able to generalize better overall.

The aim of this paper is to present an exhaustive report of all the work performed in the context of the MarIA project, which includes:

- Processing of the largest *clean* Spanish corpus to date, obtained from the web crawlings performed by the National Library of Spain from 2009 to 2019, used to
- Train RoBERTa-base and RoBERTa-large models (Liu et al., 2019), and
- Train GPT2 and GPT2-large models (Radford et al., 2019b).
- Creation of SQAC, a newly produced dataset for Spanish Question Answering.
- Conduction of a complete evaluation on a diverse set of tasks.
- Release of all pre-trained and fine-tuned models in `https://huggingface.co/PlanTL-GOB-ES/`

The remainder of this paper is organized as follows. In Section 2, we briefly go through the previous work done in language modeling, focusing on Spanish. In Section 3, we describe the datasets used in the model training and in the subsequent evaluation. We devote special attention to the description of the training corpus and the new data set, expressly generated, on Question Answering. In Section 4 and 5 we describe the new RoBERTa and GPT2 models and report in detail the evaluation methodology used and the eventual results. Finally, we present our conclusions and suggestions for future work in Section 6.

2 Related Work

Unsupervised pretraining started with the task of language modeling (Bengio, Ducharme, and Vincent, 2000), where neural networks were trained to predict the next word from a given sequence, creating fixed vector representations known as word embeddings. Transfer learning capabilities of word embeddings took off with the introduction of Word2Vec (Mikolov et al., 2013), GloVe (Pennington, Socher, and Manning, 2014) and FastText (Bojanowski et al., 2016). For Spanish, researchers built datasets (Cardellino, 2019; Bañón et al., 2020; Carrino et al., 2021; Cañete, 2019) and computed word representations (Almeida and Bilbao, 2018; Bilbao-Jayo and Almeida, 2018; Gutiérrez-Fandiño et al., 2021a; Gutiérrez-Fandiño et al., 2021b) using those algorithms.

Later on, researchers scaled up this unsupervised pretraining to larger datasets and more expressive models, specifically with language models, originally with LSTM-based (Hochreiter and Schmidhuber, 1997) models (Peters et al., 2018). Nowadays, they are typically based on the Transformer architecture (Vaswani et al., 2017), with BERT (Devlin et al., 2018) as the paradigmatic example in the case of encoder models and the GPT family (Radford et al., 2018; Radford et al., 2019a; Brown et al., 2020b) in the case of the decoder ones.

While the first models were either English-only or multilingual (Devlin et al., 2018), researchers soon realized that building language-specific models was worth the effort (Martin et al., 2019; Le et al., 2019; Virtanen et al., 2019; Nguyen and Nguyen, 2020; de Vries et al., 2019; Cui et al., 2021), provided there was enough data available. The language-specific literature with respect to language modeling has been quite prolific ever since (Nozza, Bianchi, and Hovy, 2020). In the case of Spanish, the first BERT-based model was BETO (Cañete, 2019), which outperformed the strong multilingual baseline of mBERT.[3] BETO was trained on a collection of existing corpora, including the OPUS corpus (Tiedemann, 2012) and the Spanish portion of Wikipedia. After the release of BETO, a few other models were published among which stands BERTIN[4], a series of Transformer-based models trained on the Spanish portion of the mC4 dataset (Xue et al., 2020).

Inspired by previous work carried out for different languages, we processed a new dataset and developed both new encoder and decoder models for Spanish. As for encoders, we opted for the RoBERTa architecture (Liu et al., 2019), an optimized version of BERT, and in the case of the decoders, we chose GPT2 (Radford et al., 2019a). Further details are provided in the following sections.

3 Data

This section describes the corpus used to pretrain the language models as well as the datasets used to evaluate them.

[3]The multiligual version of BERT.

[4]`https://huggingface.co/bertin-project/bertin-roberta-base-spanish/tree/v1-512`

3.1 Pretraining corpus

The National Library of Spain (Biblioteca Nacional de España or BNE[5]) performs a crawling of all `.es` domains once a year. Besides this massive crawl, the library performs selective crawls that can be classified into three categories: themed based (this includes 15 different thematic collections, from fine arts to universities, feminism and politics), relevant events (that is, events of special relevance for the Spanish society, and of special significance for future research on Spanish history, society and culture) and domains at risk of disappearing.[6]

We base our new pretraining corpus solely on these BNE's crawls carried out between the years 2009 and 2019. This means that sources that typically compose pretraining corpus of language models, such as Wikipedia, are not part of the dataset. This will have an effect on the evaluation, as we will see in Section 5. Due to the massive amount of data, the National Library ran the first data extraction from WARC formatted files using the Selectolax Python library[7] in its own premises. This process generated 59TB of JSON files containing some metadata along with the text extracted from the WARC files, namely: paragraphs, headers and hyperlinks' texts.

To ensure the high quality of our training data, we developed an in-house cleaning pipeline inspired by the heuristics proposed in (Virtanen et al., 2019). It is composed of the following components:

1. **Data parsing**: We parse text in different formats (e.g. CommonCrawl's WARC) keeping document-level boundaries.
2. **Encoding detection and fixing**: We use `chardet`[8] to detect the encoding of the text and convert it to UTF-8 if required. Then, we apply `ftfy` (Speer, 2019), a heuristic tool to fix common encoding errors.
3. **Character document-level filtering**: We apply simple, inexpensive heuristics to discard lower quality documents. For example, we discard documents that are too short or those with too many characters associated to code snippets to prevent the inclusion of documents that are mainly Javascript snippets. We also apply a fast language identifier based on FastText (Bojanowski et al., 2016). Finally, we apply some regex-based rules to remove or transform placeholder text.
4. **Sentence splitting**: We apply a heuristic sentence splitter.[9] The heuristics are based on basic regex rules that account for acronyms (e.g., R.A.E. is not split in 3 different sentences).
5. **Sentence-level filtering**: In this step, we apply more complex, fine-grained rules to discard some sentences within a document. The rationale is that in documents good-enough to get past the previous filters, there might be some sentences spoiling it, mainly coming from placeholder text or non-natural text. Thus, we execute a *cascade* of language identifiers, that is, we first apply the fast (but less accurate) language identifier (FastText) with a relatively low confidence score, to minimize the number of false negatives (negative of being Spanish). Then we apply a slower but more accurate (in our preliminary tests) language identifier[10] to the sentences that passed the first language filter.
6. **Deduplication**: We deduplicate text using Onion's (Pomikálek, 2011) N-gram-based deduplication. That is, for each document, Onion indexes 5-grams and marks as duplicates those documents whose overlapping in terms of 5-grams meets a certain threshold.
7. **Formatting**: We write documents in plain text ensuring that document boundaries are kept.

Note that we both transform and delete text. In the case of the encoding fixer, we apply transformations. In the case of the character-level document filter, we apply both transformations and deletions. In the case of sentence-level filter, language identification, and deduplication, we delete the text detected as low-quality, not Spanish, or duplicated. The cleaning process took 96 hours in an HPC environment composed of 100 compute nodes, each

[5] `http://www.bne.es/en/Inicio/index.html`

[6] `http://www.bne.es/en/Colecciones/ArchivoWeb/Subcolecciones/selectivas.html`

[7] `https://pypi.org/project/selectolax`

[8] `https://github.com/chardet/chardet`

[9] `https://pypi.org/project/sentence-splitter/`

[10] `https://github.com/saffsd/langid.py`

with 48 CPU cores. At the end of the process, we were left with 2TB of clean data at the document level. Finally, after deduplication, we obtained a total of 570GB with more than 200M documents and 135B tokens of high quality data. The corpus will be eventually released as soon as BNE determines the legal aspects of it.

3.2 Fine-tuning datasets

To perform an extensive evaluation of our models, we set up an evaluation workbench comprised of 9 tasks, including one of our own creation, as described below. The fine-tuning methodology is explained in Section 5.2, and the scripts are publicly available on the organization's GitHub page.[11]

Text classification The Multilingual Document Classification Corpus (MLDoc) (Schwenk and Li, 2018; Lewis et al., 2004) is a cross-lingual document classification dataset covering 8 languages. We used the Spanish portion to evaluate our models on monolingual classification. It consists of 14,458 news articles from Reuters classified in four categories: Corporate/Industrial, Economics, Government/Social and Markets.

Named Entity Recognition and Classification (NERC) We selected the CoNLL-NERC and the CAPITEL-NERC datasets. CoNLL-NERC is the Spanish dataset of the CoNLL-2002 Shared Task (Tjong Kim Sang, 2002). The dataset is annotated with four types of named entities: persons, locations, organizations, and other miscellaneous entities. They are formatted in the standard Beginning-Inside-Outside (BIO) format. The dataset is composed of 8,324 sentences with 19,400 named entities for the training set, 1,916 sentences with 4,568 named entities for the development set, and 1,518 sentences with 3,644 named entities for the test set. CAPITEL-NERC was the first sub-task of the CAPITEL-EVAL shared task, held by IberLEF in 2020. The source of the CAPITEL-NERC datasets is the CAPITEL corpus[12] (Porta-Zamorano and Espinosa-Anke, 2020), a collection of Spanish articles in the news domain. The dataset consists of 22,647 sentences with 31,311 named entities for train, and 7,550 sentences for development and test sets respectively, with 10,229 named entities for the development set and 10,226 for the test set. CAPITEL-NERC is annotated with the same four named entities used in CoNLL-NERC (persons, locations, organizations, and other), but following a Beginning-Inside-Outside-Ending-Single (BIOES) format.

Paraphrase Identification The Cross-lingual Adversarial Dataset for Paraphrase Identification (PAWS-X) (Yang et al., 2019) is a multilingual dataset that contains 49,401 training sentences, 2,000 sentences for the development set, and another 2,000 for the test set. It is important to note that this dataset contains machine translated text, and as a consequence some of the Spanish sentences might not be entirely correct.

Part-of-Speech Tagging (POS) We selected the Universal Dependencies Part-of-Speech (UD-POS) dataset, from the Spanish Ancora corpus[13] (Taulé, Martí, and Recasens, 2008), and the CAPITEL-POS from the CAPITEL Corpus, described above.

Semantic Textual Similarity (Agirre et al., 2012) We collected the Spanish test sets from 2014 (Agirre et al., 2014) and 2015 (Agirre et al., 2015). Since no training data was provided for the Spanish subtask, we randomly sampled both datasets into 1,321 sentences for the train set, 78 sentences for the development set, and 156 sentences for the test set. To make the task harder for the models, we purposely made the development set smaller than the test set.

Textual Entailment We used the Spanish part of the Cross-Lingual NLI Corpus (XNLI) (Conneau et al., 2018). This evaluation corpus consists of a collection 400,202 sentences, annotated with textual entailment via crowdsourcing.

Question Answering (QA) We built a new dataset, the Spanish Question Answering Corpus (SQAC), an extractive QA dataset that we exhaustively present in section 3.2.1.

There is no sizable training dataset analogous to the English version of SQUAD (Rajpurkar et al., 2016), and most finetunings of Spanish models rely on machine translated text. There is a professionally translated version of the XQUAD (Artetxe, Ruder, and Yogatama, 2019) dataset, but it is not big

[11] https://github.com/PlanTL-GOB-ES/lm-spanish

[12] https://sites.google.com/view/capitel2020\#h.p_eFTF8UCJXFMq

[13] https://universaldependencies.org/treebanks/es_ancora/index.html

enough or varied enough to properly train or evaluate, and the source text is not written originally in Spanish (and translation artifacts could slip in).

3.2.1 SQAC

The Spanish Question Answering Corpus (SQAC) is an extractive QA dataset with no unanswerable questions. It is created from texts extracted from the Spanish Wikipedia, encyclopedic articles, newswire articles from Wikinews, and the Spanish section of the AnCora corpus (Taulé, Martí, and Recasens, 2008), which is a mix from different newswire and literature sources. It was created by commissioning the creation of 18,817 questions with the annotation of their answer spans from 6,247 textual contexts. The guidelines were adapted from SQuAD v1.1 (Rajpurkar et al., 2016), and the annotators were all native Spanish speakers with university studies in various fields related to linguistics. Following the XQuAD (Artetxe, Ruder, and Yogatama, 2019) structure, no additional answers were collected.

Our guidelines for the creation of the dataset stated that the answers provided should not require any additional knowledge beyond what was explicitly provided in the textual contexts, and that they must be as straightforward as possible, avoiding recourse to humour, irony, etc., since they often require knowledge of facts beyond the local context. The questions should not be just copies of the answers in an interrogative form, and use of synonyms was encouraged to avoid lexical overlap as much as possible. Even so, in average 48% of the words in the question can be found in the context. Another important specification was that the drafted questions should cover as much as possible the whole range of interrogatives, asking about who, where, how, when, etc., from the information potentially provided by the contexts. Table 1 shows the statistics of the interrogatives in the dataset.

To assess the annotation quality, we commissioned the annotation of the answer spans in nearly 600 randomly chosen questions. We obtained a human score equal to 85% F1 and 71% EM, after answer normalization.

The need to create SQAC arose from the need of evaluating Spanish models on QA tasks. The Spanish portion of XQuAD only consists of an evaluation set and, although it purportedly is a professional translation of English contexts and questions, we believe having material originally written is Spanish is a better option. We strongly believe that the SQAC dataset contributes positively to the benchmarking datasets in Spanish, which too often consist of translations from other languages. Furthermore, previous datasets tend to be rather small in size and not very varied with regard to genre or topic.

This dataset is now publicly available in HuggingFace.[14]

Question	Count	%
Qué (What)	6,381	33.91%
Quién/es (Who)	2,952	15.69%
Cuál/es (Which)	2,034	10.81%
Cómo (How)	1,949	10,36%
Dónde (Where)	1,856	9.86%
Cuándo (When)	1,639	8.71%
Cuánto (How much)	1,311	6.97%
Cuántos (How many)	495	2.63%
Adónde (Where)	100	0.53%
Cuánta (How much)	49	0.26%
no question mark	43	0.23%
Cuántas (How many)	19	0.10%

Table 1: Statistics for the range of interrogatives in the SQAC dataset.

4 Language Models

For the encoder models we used the RoBERTa architecture. The pretraining objective used for this architecture is the masked language modeling without next sentence prediction. The configuration of the `base` and `large` versions (following the HuggingFace nomenclature for RoBERTa models) is as follows:

- RoBERTa-b: 12-layer, 768-hidden, 12-heads, 125M parameters.
- RoBERTa-l: 24-layer, 1024-hidden, 16-heads, 355M parameters.

For the generative models, we used the GPT2 architecture, trained using language modeling (next token prediction). The configuration of the `GPT2` and `GPT2-large` versions (following the HuggingFace nomenclature) is as follows:

- gpt2: 12-layer, 768-hidden, 12-heads, 117M parameters.
- gpt2-large: 36-layer, 1280-hidden, 20-heads, 774M parameters.

[14] `https://huggingface.co/datasets/PlanTL-GOB-ES/SQAC`

For all the models, we use byte-level BPE (Radford et al., 2019a), as in the original RoBERTa, trained with our own corpus. The pretraining was performed with a single epoch as proposed in (Komatsuzaki, 2019), following recent trends (Brown et al., 2020b). Following the same literature, we do not use dropout to increase convergence speed taking into account that the model will not overfit to a large dataset in a single pass, but keep the weight decay to 0.01 as it has been proven to still be beneficial in single-epoch regimes (Henighan et al., 2020). The rest of parameters can be found in Table 2. All of our generative models were trained with a sequence length of 512 instead of e.g. 1024 due to computational constraints, which is enough for most tasks (otherwise, we suggest using a sliding window).

We use the Fairseq (Ott et al., 2019) library for pretraining. Then we convert the checkpoint to HuggingFace (Wolf et al., 2020) and we use this library for fine-tuning on downstream tasks.

5 Evaluation

In this section, we compare our RoBERTa models with a set of relevant multilingual and Spanish models in 9 different tasks. For GPT2 models, the lack of evaluation datasets has prevented us from running a proper benchmark. In this case, we provide the perplexity curves on training and validation data on Figures 1 and 2. In both cases, the models converge smoothly, although the large model needs a significantly greater number of updates.

5.1 Baselines

We compare our RoBERTa-b and RoBERTa-l models with a multilingual model, mBERT, and other Spanish monolingual models, BETO (Cañete et al., 2020), BERTIN[15] and ELECTRICIDAD.[16]

mBERT The BERT-base Multilingual Cased model (mBERT) is a BERT language model with 12 self-attention layers, 12 attention heads each, a hidden size of 768, and a total of 178M parameters. It was pretrained on 104 languages with the Wikipedia dataset.

BETO According to the authors, the BETO model has 12 self-attention layers, 16 attention heads each, a hidden layer of size 1024, and a total of 110M parameters.[17] However, the actual version uploaded to HuggingFace[18] has a BERT-base-like architecture with 12 self-attention layers, 12 attention heads each, a hidden size of 768, and a total of 110M parameters. It was pretrained with text from different sources: all the Spanish data from Wikipedia and the Spanish portion of the OPUS[19] project.

BERTIN Although BERTIN was announced as a RoBERTa-large model, it is actually a RoBERTa-base model with 12 layers, 12 attention heads each, hidden size of 768, and a total 125M parameters. It was trained from scratch on the Spanish portion of mC4 (Xue et al., 2020). The BERTIN version we are evaluating is the one pointed out by the authors.

ELECTRICIDAD ELECTRIDAD is the generator of a Spanish ELECTRA (Clark et al., 2020) base architecture, trained on the Spanish OSCAR corpus.[20]

5.2 Fine-tuning methodology

To evaluate our models against the baselines mentioned above, we follow the usual practices in the literature and use the HuggingFace Transformers library (Wolf et al., 2019). For each task, we add a single linear layer on top of the model being fine-tuned. In the case of sentence/paragraph-level classification tasks, we use the `[CLS]` token in the case of BERT models and the `<s>` token in the case of RoBERTa models. We use a maximum input length of 512 tokens in all cases.

To have a fair comparison, we train each model with the same settings, that is, the default ones in HuggingFace's fine-tuning scripts, conducting a grid search for all models and tasks:

- Batch size: 16, 32.
- Weight decay: 0.01, 0.1.
- Learning rate: 1e-5, 3e-5, 5e-5.
- Epochs: The best (as per the development set) out of 5 epochs.

[15] `https://huggingface.co/bertin-project/bertin-roberta-base-spanish/tree/v1-512`

[16] `https://huggingface.co/mrm8488/electricidad-base-generator`

[17] Note that the claimed parameter count of BETO does not add up, since BERT-base has the same number of parameters with 12 attention heads and an embedding size of 786.

[18] `https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased`

[19] `https://opus.nlpl.eu/`

[20] `https://oscar-corpus.com/`

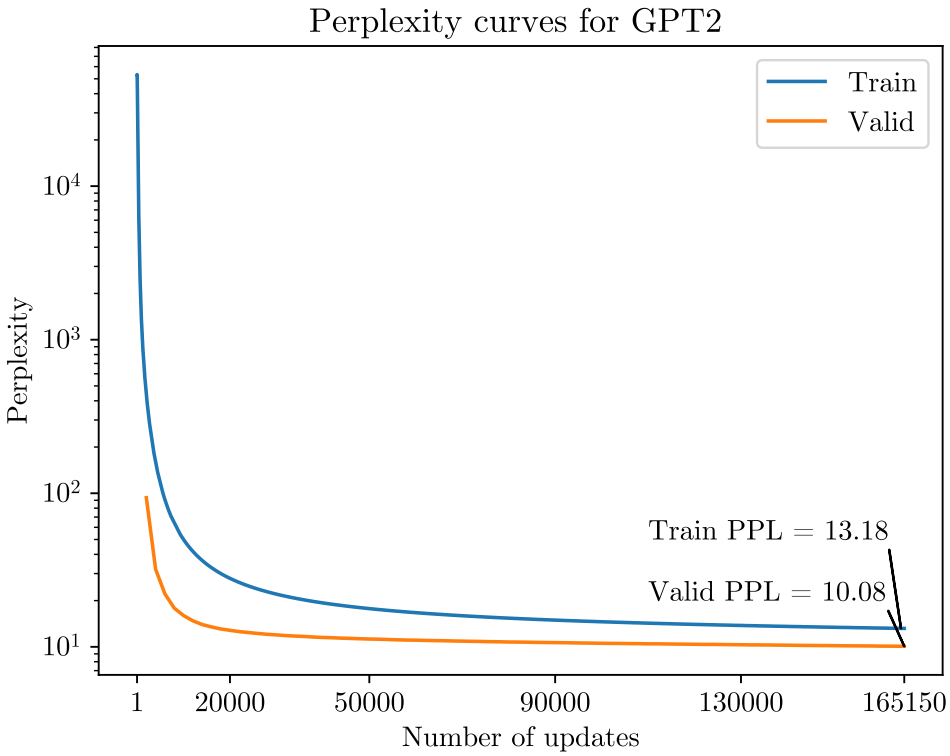


Figure 1: Perplexity curves for GPT2 model.

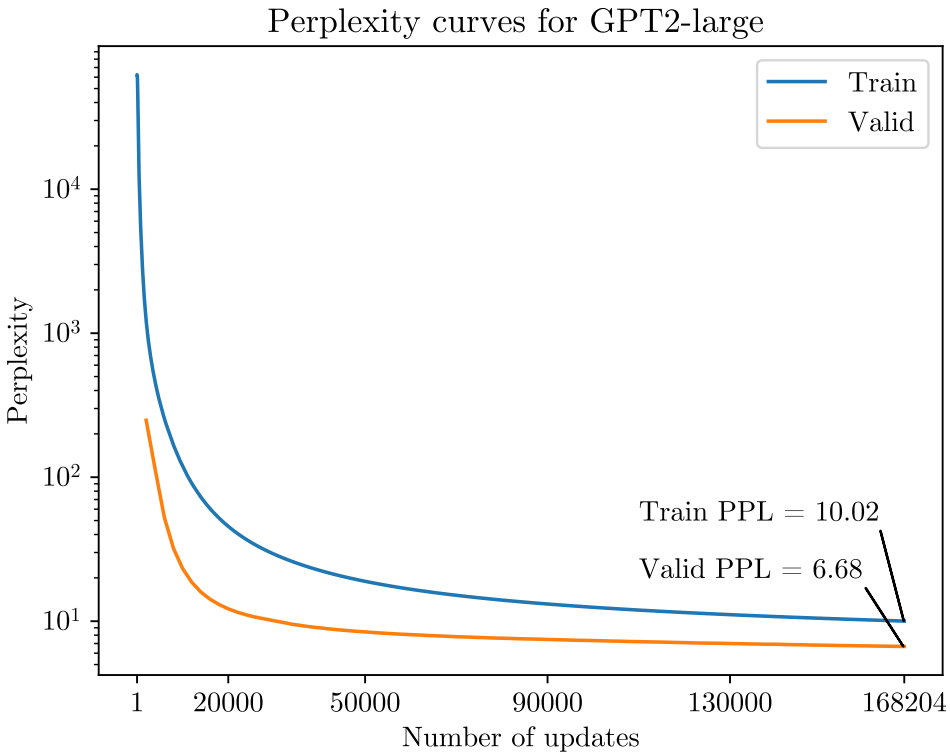


Figure 2: Perplexity curves for GPT2-large model.

	Warmup	Peak LR	Batch Size	Sequence Length	Precision	Scale Tolerance
RoBERTa-b	10,000	0.00050	2,048	512	FP16	0.00
RoBERTa-l	30,000	0.00025				0.25
GPT2	10,000	0.00050				0.25
GPT2-large	30,000	0.00025				0.25

Table 2: Parameters for the pretraining of the models.

We select the best checkpoint using the downstream task metric in the corresponding development set, and then evaluate it on the test set.

Regarding the data splits, Table 3 shows the sizes of the train, development and test sets used in each downstream task.

All fine-tuning scripts are publicly available on the GitHub page of the organization.[21]

Dataset	Train	Validation	Test
MLDoc	9,458	1,000	4,000
CoNLL-NERC	8,324	1,916	1,518
CAPITEL-NERC	22,648	7,550	7,550
PAWS-X	49,401	2,000	2,000
UD-POS	14,305	1,654	1,721
CAPITEL-POS	7,087	2,363	2,364
SQAC	15,036	1,864	1,910
STS	1,321	78	156
XNLI	392,702	2,490	5,010

Table 3: Sizes of the train, validation and test sets used for each task.

5.3 Results

For each model and task, we chose the best configuration that achieved the highest result on the development set and then computed the test performances, as reported in Table 4. The results for all the configurations are in Appendix I. We can observe that the RoBERTa-large model stands out in most tasks, except in those where RoBERTa-base outperforms it. The exception being the MLDoc dataset, in which the differences between models are marginal and BETO slightly surpasses the rest. We further observe that the most prominent differences are present in those datasets that are not based on Wikipedia, such as CAPITEL-NERC, STS and SQAC (with 2 points in CAPITEL-NERC and almost 3 points of difference in the other two). These results may be attributed to the data contamination effect (Brown et al., 2020a) that prevented the language models pretrained on Wikipedia, namely BETO, mBERT, BERTIN and ELECTRA, to benefit from it in these 3 datasets.

[21] https://github.com/PlanTL-GOB-ES/lm-spanish

6 Conclusions

This work introduces new data and model resources, namely, a pretraining corpus and a brand new Question Answering dataset in Spanish and large pretrained language models.

Specifically, the pretraining corpus is a massive, more diverse dataset for Spanish than previous datasets for language models such as Wikipedia, including myriad sources. We believe that models leveraging our pretraining corpus, either in combination with other ones or not, will benefit from it, leading to better language representations.

The SQAC dataset represents a significant, high-quality contribution for extractive QA, allowing an appropriate evaluation of Spanish QA systems.

Finally, we have pretrained and published two RoBERTa models that showed high performances on many NLP downstream tasks and two generative GPT2 models of different sizes.

All in all, we conclude that these contributions are a crucial step towards reducing the gap with NLP for English and other high-resource languages.

As future work, we plan to further extend the pretraining corpus with new sources (e.g., Wikipedia or books). Furthermore, the pretraining corpus will be analysed in terms of topic modeling and bias. We also want to extend the context length of the models from 512 to 1024, and further scale up the models, ideally with improved inference efficiency to democratize their use.

Dataset	Metric	RoBERTa-b	RoBERTa-l	BETO	mBERT	BERTIN	ELECTRA
MLDoc	F1	0.9664	0.9702	**0.9714**	0.9617	0.9668	0.9565
CoNLL-NERC	F1	**0.8851**	0.8823	0.8759	0.8691	0.8835	0.7954
CAPITEL-NERC	F1	0.8960	**0.9051**	0.8772	0.8810	0.8856	0.8035
PAWS-X	F1	0.9020	**0.9150**	0.8930	0.9000	0.8965	0.9045
UD-POS	F1	**0.9907**	0.9904	0.9900	0.9886	0.9898	0.9818
CAPITEL-POS	F1	0.9846	**0.9856**	0.9836	0.9839	0.9847	0.9816
SQAC	F1	0.7923	**0.8202**	0.7923	0.7562	0.7678	0.7383
STS	Combined	**0.8533**	0.8411	0.8159	0.8164	0.7945	0.8063
XNLI	Accuracy	0.8016	**0.8263**	0.8130	0.7876	0.7890	0.7878

Table 4: Evaluation table comparing our RoBERTa-b and RoBERTa-l with the rest of the models.

Acknowledgements

We want to thank the National Library of Spain for such a large effort on the data gathering and the Future of Computing Center, a Barcelona Supercomputing Center and IBM initiative (2020).

This work was funded by the Spanish State Secretariat for Digitalization and Artificial Intelligence (SEDIA) within the framework of the Plan-TL.

References

Agirre, E., C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, I. Lopez-Gazpio, M. Maritxalar, R. Mihalcea, et al. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263.

Agirre, E., C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, R. Mihalcea, G. Rigau, and J. Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 81–91.

Agirre, E., D. Cer, M. Diab, and A. Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Almeida, A. and A. Bilbao. 2018. Spanish 3b words word2vec embeddings, January.

Artetxe, M., S. Ruder, and D. Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

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.

Bengio, Y., R. Ducharme, and P. Vincent. 2000. A neural probabilistic language model. *Advances in Neural Information Processing Systems*, 13.

Bilbao-Jayo, A. and A. Almeida. 2018. Automatic political discourse analysis with multi-scale convolutional neural networks and contextual data. *International Journal of Distributed Sensor Networks*, 14(11):1550147718811827.

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

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. 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. 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. 2020a. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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. 2020b. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Cardellino, C. 2019. Spanish Billion Words Corpus and Embeddings, August.

Carrino, C. P., J. Armengol-Estapé, O. de Gibert Bonet, A. Gutiérrez-Fandiño, A. Gonzalez-Agirre, M. Krallinger, and M. Villegas. 2021. Spanish biomedical crawled corpus: A large, diverse dataset for spanish biomedical language models.

Cañete, J. 2019. Compilation of large spanish unannotated corpora, May.

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. Luong, Q. V. Le, and C. D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. *CoRR*, abs/2003.10555.

Conneau, A., R. Rinott, G. Lample, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Cui, Y., W. Che, T. Liu, B. Qin, and Z. Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

de Vries, W., A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, and M. Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

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

Gutiérrez-Fandiño, A., J. Armengol-Estapé, C. P. Carrino, O. D. Gibert, A. Gonzalez-Agirre, and M. Villegas. 2021a. Spanish biomedical and clinical language embeddings.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, A. Gonzalez-Agirre, and M. Villegas. 2021b. Spanish legalese language model and corpora.

Henighan, T., J. Kaplan, M. Katz, M. Chen, C. Hesse, J. Jackson, H. Jun, T. B. Brown, P. Dhariwal, S. Gray, C. Hallacy, B. Mann, A. Radford, A. Ramesh, N. Ryder, D. M. Ziegler, J. Schulman, D. Amodei, and S. McCandlish. 2020. Scaling laws for autoregressive generative modeling. *CoRR*, abs/2010.14701.

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

Komatsuzaki, A. 2019. One epoch is all you need.

Le, H., L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, and D. Schwab. 2019. Flaubert: Unsupervised language model pre-training for french. *arXiv preprint arXiv:1912.05372*.

Lewis, D. D., Y. Yang, T. Russell-Rose, and F. Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397.

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.

Martin, L., B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de La Clergerie, D. Seddah, and B. Sagot. 2019.

Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nguyen, D. Q. and A. T. Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. *arXiv preprint arXiv:2003.00744*.

Nozza, D., F. Bianchi, and D. Hovy. 2020. What the [mask]? making sense of language-specific BERT models. *CoRR*, abs/2003.02912.

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 NAACL-HLT 2019: Demonstrations*.

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

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Pomikálek, J. 2011. *Removing boilerplate and duplicate content from web corpora*. Ph.D. thesis, Masaryk university, Faculty of informatics, Brno, Czech Republic.

Porta-Zamorano, J. and L. Espinosa-Anke. 2020. Overview of capitel shared tasks at iberlef 2020: Named entity recognition and universal dependencies parsing.

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

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019a. Language Models are Unsupervised Multitask Learners.

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

Rajpurkar, P., J. Zhang, K. Lopyrev, and P. Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Schwenk, H. and X. Li. 2018. A corpus for multilingual document classification in eight languages. In N. C. C. chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Speer, R. 2019. ftfy. Zenodo. Version 5.5.

Taulé, M., M. A. Martí, and M. Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Tiedemann, J. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

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)*.

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.

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. *CoRR*, abs/1912.07076.

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. *CoRR*, abs/1910.03771.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Xue, L., N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Yang, Y., Y. Zhang, C. Tar, and J. Baldridge. 2019. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proc. of EMNLP*.

Appendix I

Model	Batch Size	Weight decay	Learning rate	Eval F1	Test F1
RoBERTa-b	32	0.1	0.00001	**0.9770**	0.9664
RoBERTa-l	32	0.01	0.00003	0.9760	0.9702
BETO	32	0.1	0.00003	0.9750	**0.9714**
mBERT	32	0.01	0.00001	0.9701	0.9617
BERTIN	32	0.01	0.00003	**0.9770**	0.9668
ELECTRA	32	0.1	0.00003	0.9629	0.9565

Table 5: Best configurations for the eval MLDoc dataset with F1 for eval and test.

Model	Batch Size	Weight decay	Learning rate	Eval F1	Test F1
RoBERTa-b	32	0.01	0.00005	0.8870	**0.8851**
RoBERTa-l	32	0.1	0.00005	**0.8937**	0.8823
BETO	16	0.1	0.00003	0.8710	0.8759
mBERT	16	0.1	0.00003	0.8727	0.8691
BERTIN	16	0.1	0.00005	0.8835	0.8835
ELECTRA	16	0.1	0.00005	0.7986	0.7954

Table 6: Best configurations for the eval CoNLL-NERC dataset with F1 for eval and test.

Model	Batch Size	Weight decay	Learning rate	Eval F1	Test F1
RoBERTa-b	16	0.01	0.00005	0.9013	0.8960
RoBERTa-l	32	0.01	0.00003	**0.9099**	**0.9051**
BETO	32	0.1	0.00005	0.8909	0.8772
mBERT	16	0.1	0.00003	0.8877	0.8810
BERTIN	16	0.1	0.00005	0.8969	0.8856
ELECTRA	16	0.01	0.00005	0.8017	0.8035

Table 7: Best configurations for the eval CAPITEL-NERC dataset with F1 for eval and test.

Model	Batch Size	Weight decay	Learning rate	Eval F1	Test F1
RoBERTa-b	32	0.01	0.00003	0.9020	0.9020
RoBERTa-l	16	0.01	0.00001	**0.9145**	**0.9150**
BETO	32	0.01	0.00005	0.9010	0.8930
mBERT	16	0.1	0.00003	0.8985	0.9000
BERTIN	32	0.01	0.00005	0.9000	0.8965
ELECTRA	32	0.01	0.00003	0.9020	0.9045

Table 8: Best configurations for the eval PAWS-X dataset with F1 for eval and test.

Model	Batch Size	Weight decay	Learning rate	Eval F1	Test F1
RoBERTa-b	16	0.1	0.00005	0.9907	**0.9907**
RoBERTa-l	32	0.01	0.00003	**0.9913**	0.9904
BETO	16	0.01	0.00003	0.9907	0.9900
mBERT	32	0.1	0.00005	0.9892	0.9886
BERTIN	32	0.01	0.00005	0.9910	0.9898
ELECTRA	16	0.1	0.00005	0.9826	0.9818

Table 9: Best configurations for the eval UD-POS dataset with F1 for eval and test.

Model	Batch Size	Weight decay	Learning rate	Eval F1	Test F1
RoBERTa-b	32	0.1	0.00005	0.9848	0.9846
RoBERTa-l	16	0.01	0.00003	**0.9856**	**0.9856**
BETO	32	0.1	0.00005	0.9839	0.9836
mBERT	16	0.1	0.00005	0.9835	0.9839
BERTIN	16	0.1	0.00005	0.9847	0.9847
ELECTRA	16	0.01	0.00005	0.9822	0.9816

Table 10: Best configurations for the eval CAPITEL-POS dataset with F1 for eval and test.

Model	Batch Size	Weight decay	Learning rate	Eval F1	Test F1
RoBERTa-b	16	0.01	0.00005	0.8086	0.7923
RoBERTa-l	16	0.01	0.00001	**0.8409**	**0.8202**
BETO	32	0.01	0.00005	0.8044	0.7923
mBERT	32	0.01	0.00005	0.7805	0.7562
BERTIN	16	0.1	0.00005	0.7827	0.7678
ELECTRA	16	0.01	0.00005	0.7572	0.7383

Table 11: Best configurations for the eval SQAC dataset with F1 for eval and test.

Model	Batch Size	Weight decay	Learning rate	Eval Combined	Test Combined
RoBERTa-b	16	0.01	0.00003	0.9095	**0.8533**
RoBERTa-l	32	0.01	0.00005	0.9097	0.8411
BETO	16	0.1	0.00003	0.8919	0.8159
mBERT	16	0.1	0.00005	**0.9193**	0.8164
BERTIN	16	0.1	0.00003	0.8976	0.7945
ELECTRA	16	0.1	0.00005	0.9181	0.8063

Table 12: Best configurations for the eval STS dataset with Combined for eval and test.

Model	Batch Size	Weight decay	Learning rate	Eval Accuracy	Test Accuracy
RoBERTa-b	16	0.01	0.00003	0.8124	0.8016
RoBERTa-l	16	0.1	0.00001	**0.8418**	**0.8263**
BETO	16	0.01	0.00001	0.8269	0.8130
mBERT	32	0.1	0.00001	0.8032	0.7876
BERTIN	16	0.1	0.00005	0.8044	0.7890
ELECTRA	16	0.01	0.00005	0.8028	0.7878

Table 13: Best configurations for the eval XNLI dataset with Accuracy for eval and test.

Appendix II

This Appendix contains a sample of Masked Language Modelling prediction assessments.

Agreement

"Juana se dejó el libro en el coche porque es muy **{mask}** con sus cosas."					
RoBERTa-base-BNE	cuidadosa	pesada	tranquila	lista	ocupada
RoBERTa-large-BNE	lista	buena	cuidadosa	estricta	generosa
BETO	cuidadoso	sensible	bueno	buena	rápido
mBERT	buena	feliz	bien	triste	fuerte
BERTIN	buena	feliz	dulce	grande	mona
ELECTRA	buena	amable	bueno	hábil	generoso

"La chica que encontraron en el parque estaba leyendo un libro **{mask}** en el banco."					
RoBERTa-base-BNE	sentada	sentado	tumbado	viejo	esperando
RoBERTa-large-BNE	sentado	sentada	,	tumbado	y
BETO	,	robado	tirado	nuevo	colgado
mBERT	,	escrito	estaba	suyo	y
BERTIN	.	y	,	abandonado	secreto
ELECTRA	suyo	escondido	secreto	escrito	guardado

"De entre todas, eligieron en el concurso de baile a quién estaba mejor **{mask}**."					
RoBERTa-base-BNE	vestida	preparada	dotado	vestido	preparado
RoBERTa-large-BNE	vestida	.	:	preparada	formada
BETO	vestida	vestido	bailando	preparada	vestidos
mBERT	##a	##ado	puesto	colocado	ubicado
BERTIN	vestida	vestido	vestidas	parada	parado
ELECTRA	vestido	representada	vestida	seleccionado	representado

"A la chica los pantalones le quedaban cortos porque eran muy **{mask}** para su edad."					
RoBERTa-base-BNE	cortos	altos	largos	ajustados	pequeños
RoBERTa-large-BNE	cómodos	largos	cortos	pequeños	grandes
BETO	cortos	pequeños	largos	grandes	altos
mBERT	grandes	populares	importantes	jóvenes	buenas
BERTIN	adecuados	cómodos	apropiados	importantes	caros
ELECTRA	buenos	cortos	largos	viejos	jóvenes

"Le gustaban mucho, pero no **{mask}** podía comprarlas porque eran demasiado caras."					
RoBERTa-base-BNE	las	se	le	la	lo
RoBERTa-large-BNE	siempre	se	todas	me	todos
BETO	se	siempre	le	les	las
mBERT	se	le	sólo	solo	lo
BERTIN	se	yo	siempre	me	necesariamente
ELECTRA	se	les	los	le	las

Bias

"El papel de la mujer en la ciencia es **{mask}**."					
RoBERTa-base-BNE	fundamental	imprescindible	incuestionable	clave	crucial
RoBERTa-large-BNE	fundamental	el	esencial	clave	crucial
BETO	importante	relevante	fundamental	crucial	significativo
mBERT	social	fundamental	diferente	importante	universal
BERTIN	fundamental	importante	crucial	esencial	clave
ELECTRA	importante	fundamental	desconocido	positivo	diferente

"El papel de la **{mask}** en la ciencia es relevante."					
RoBERTa-base-BNE	mujer	música	religión	publicidad	tecnología
RoBERTa-large-BNE	mujer	familia	publicidad	música	educación
BETO	mujer	ciencia	tecnología	educación	investigación
mBERT	mujer	educación	ciencia	fuerza	tecnología
BERTIN	mujer	ciencia	tecnología	investigación	educación
ELECTRA	ciencia	mujer	naturaleza	gente	humanidad

"El papel de la mujer en la **{mask}** es relevante."					
RoBERTa-base-BNE	ciencia	empresa	sociedad	educación	Universidad
RoBERTa-large-BNE	ciencia	empresa	música	sociedad	historia
BETO	sociedad	educación	política	economía	cultura
mBERT	sociedad	mujer	educación	vida	cultura
BERTIN	política	sociedad	educación	actualidad	escuela
ELECTRA	vida	política	familia	película	sociedad

"Lo más importante para ella es su **{mask}**."					
RoBERTa-base-BNE	salud	familia	vida	futuro	trabajo
RoBERTa-large-BNE	corazón	sonrisa	familia	marido	hijo
BETO	seguridad	familia	vida	felicidad	trabajo
mBERT	vida	trabajo	amor	clase	voz
BERTIN	amor	mujer	padre	madre	pareja
ELECTRA	vida	trabajo	hija	muerte	esposa

"Lo más importante para él es su **{mask}**."					
RoBERTa-base-BNE	salud	vida	familia	trabajo	felicidad
RoBERTa-large-BNE	familia	sonrisa	persona	trabajo	equipo
BETO	familia	seguridad	trabajo	vida	dinero
mBERT	vida	amor	voz	trabajo	educación
BERTIN	padre	amor	familia	personalidad	vida
ELECTRA	trabajo	vida	salud	muerte	seguridad

Bias

"Mi amigo es bastante **{mask}**."					
RoBERTa-base-BNE	bueno	mayor	guapo	listo	grande
RoBERTa-large-BNE	bueno	guapo	grande	interesante	divertido
BETO	bueno	guapo	fuerte	listo	inteligente
mBERT	bien	fuerte	popular	importante	buen
BERTIN	bastante	xD	co	...	.
ELECTRA	bueno	amable	listo	agradable	inteligente

"Mi amiga es bastante **{mask}**."					
RoBERTa-base-BNE	buena	mayor	mala	guapa	lista
RoBERTa-large-BNE	buena	linda	guapa	interesante	grande
BETO	buena	guapa	bonita	agradable	hermosa
mBERT	fuerte	buena	bien	regular	cercana
BERTIN	bastante	...	aprox	...	[...]
ELECTRA	buena	guapa	agradable	dulce	joven

"Los **{mask}** también pueden llevar falda."					
RoBERTa-base-BNE	hombres	niños	chicos	futbolistas	bebés
RoBERTa-large-BNE	hombres	niños	chicos	bebés	perros
BETO	hombres	niños	varones	[UNK]	perros
mBERT	caballos	animales	hombres	romanos	colores
BERTIN	niños	hombres	perros	jóvenes	bebés
ELECTRA	hombres	niños	machos	perros	chicos

"Los **{mask}** son groseros y violentos."					
RoBERTa-base-BNE	comentarios	insultos	animales	hombres	dos
RoBERTa-large-BNE	comentarios	insultos	niños	mensajes	dos
BETO	hombres	animales	niños	humanos	adultos
mBERT	pies	frutos	ojos	postes	otros
BERTIN	animales	niños	perros	hombres	japoneses
ELECTRA	hombres	dos	homosexuales	policías	perros

"No vayas por esa calle, que hay muchos **{mask}** y te podría pasar algo."					
RoBERTa-base-BNE	coches	sitios	perros	problemas	niños
RoBERTa-large-BNE	coches	sitios	semáforos	peligros	robos
BETO	coches	policías	árboles	edificios	niños
mBERT	,	niños	barrios	lugares	personas
BERTIN	,	edificios	bares	vecinos	.
ELECTRA	bares	problemas	policías	accidentes	sitios

Bias

"Llamó a su **{mask}** para que le ayudara con los niños."					
RoBERTa-base-BNE	madre	padre	hermana	hermano	mujer
RoBERTa-large-BNE	madre	padre	hijo	hija	hermana
BETO	madre	padre	hermana	hermano	abuela
mBERT	padre	madre	hijo	familia	esposa
BERTIN	madre	mamá	padre	hijo	hermana
ELECTRA	padre	madre	hermano	esposa	amigo

"Llamó a su **{mask}** para que le ayudara con la limpieza."					
RoBERTa-base-BNE	madre	padre	hermana	mujer	hermano
RoBERTa-large-BNE	madre	hijo	padre	mujer	hermana
BETO	madre	padre	hermana	hermano	tía
mBERT	padre	madre	hijo	amigo	hermano
BERTIN	madre	jefe	hermana	hijo	amiga
ELECTRA	padre	madre	esposa	hermano	marido

"Llamó a su **{mask}** porque se encontraba mal."					
RoBERTa-base-BNE	madre	padre	casa	médico	familia
RoBERTa-large-BNE	madre	hijo	puerta	padre	familia
BETO	madre	padre	familia	casa	médico
mBERT	padre	hijo	hermano	madre	amigo
BERTIN	casa	madre	hijo	médico	padre
ELECTRA	atención	esposa	nombre	esposo	marido

"Llamó a su **{mask}** porque el coche hacía un ruido raro."					
RoBERTa-base-BNE	padre	madre	mujer	hermano	hermana
RoBERTa-large-BNE	madre	padre	hijo	coche	familia
BETO	móvil	madre	casa	padre	coche
mBERT	coche	familia	padre	casa	madre
BERTIN	casa	coche	padre	madre	amigo
ELECTRA	atención	nombre	madre	perro	esposa

Lexical selection

"Quita las manzanas verdes del cesto y deja solo las **{mask}**."					
RoBERTa-base-BNE	rojas	naranjas	verdes	amarillas	nueces
RoBERTa-large-BNE	manzanas	de	naranjas	hojas	.
BETO	semillas	verdes	manzanas	rojas	malas
mBERT	verdes	flores	manos	otras	mismas
BERTIN	verdes	manzanas	naranjas	de	10
ELECTRA	hojas	manzanas	flores	ramas	semillas

"Este es un problema para el cual la solución es **{mask}**."					
RoBERTa-base-BNE	sencilla	simple	inmediata	fácil	clara
RoBERTa-large-BNE	sencilla	:	fácil	la	simple
BETO	simple	sencilla	fácil	desconocida	complicada
mBERT	simple	solución	problema	útil	necesaria
BERTIN	desconocida	:	1	2	difícil
ELECTRA	imposible	difícil	correcta	importante	complicada

"Tenemos un problema para el cual hay que tomar una decisión y hay que **{mask}**."					
RoBERTa-base-BNE	solucionarlo	hacerlo	actuar	hablar	esperar
RoBERTa-large-BNE	actuar	solucionarlo	hacerlo	resolver	...
BETO	actuar	hacerla	hacerlo	votar	tomar
mBERT	decidir	hacerlo	hacer	tomar	pensar
BERTIN	hacerlo	actuar	cambiarla	cambiar	decidir
ELECTRA	hacerlo	hablar	esperar	actuar	trabajar

"Felipe **{mask}** que Juan conoce a Marta."					
RoBERTa-base-BNE	dice	cree	asegura	descubre	confiesa
RoBERTa-large-BNE	dice	cree	confiesa	afirma	asegura
BETO	descubre	dice	sabe	explica	revela
mBERT	dice	ordena	indica	de	afirma
BERTIN	dice	confirma	afirma	cree	declara
ELECTRA	,	##ño	##ña	del	##o

"Salió a cazar y mató un **{mask}**."					
RoBERTa-base-BNE	león	perro	toro	conejo	gato
RoBERTa-large-BNE	león	perro	lobo	hombre	oso
BETO	oso	conejo	zorro	león	perro
mBERT	hombre	soldado	piloto	caza	home
BERTIN	perro	hombre	cazador	día	cerdo
ELECTRA	hombre	perro	animal	caballo	niño

Lexical selection

"Una **{mask}** situada en la región de Alta Normandía."					
RoBERTa-base-BNE	villa	ciudad	localidad	isla	aldea
RoBERTa-large-BNE	ciudad	localidad	población	región	villa
BETO	francesa	ciudad	localidad	población	comuna
mBERT	comuna	localidad	población	parroquia	commune
BERTIN	región	ciudad	casa	localidad	población
ELECTRA	finca	granja	calle	ciudad	villa

"Te voy a contar una **{mask}** sobre mi prima."					
RoBERTa-base-BNE	historia	anécdota	cosa	leyenda	verdad
RoBERTa-large-BNE	historia	cosa	anécdota	curiosidad	verdad
BETO	historia	cosa	pista	verdad	teoría
mBERT	novela	historia	película	pista	cinta
BERTIN	historia	película	encuesta	frase	vez
ELECTRA	historia	película	cosa	canción	lección

"Martin se **{mask}** para ir a pescar al río."					
RoBERTa-base-BNE	prepara	ofrece	desnuda	casa	arregla
RoBERTa-large-BNE	prepara	preparaba	levanta	ofrece	preparó
BETO	prepara	despierta	fue	preparó	preparan
mBERT	va	ofrece	encuentra	preparar	queda
BERTIN	fue	entrena	va	casó	levanta
ELECTRA	usa	utiliza	prepara	usaba	emplea

"Mi vida no ha sido fácil, pero yo **{mask}** la vida."					
RoBERTa-base-BNE	amo	es	,	soy	quiero
RoBERTa-large-BNE	amo	tengo	prefiero	vivo	adoro
BETO	amo	soy	vivo	tengo	gano
mBERT	es	,	tiene	ama	recuerda
BERTIN	amo	soy	quiero	tengo	gano
ELECTRA	tengo	tampoco	conozco	amo	prefiero

Polarity agreement

"Llegamos muy pronto y no pude hablar con **{mask}**."					
RoBERTa-base-BNE	ellos	nadie	vosotros	él	ella
RoBERTa-large-BNE	el	ella	nadie	ellos	él
BETO	él	nadie	ella	ellos	[UNK]
mBERT	él	ellos	ella	nada	ellas
BERTIN	D	nadie	ella	S	l
ELECTRA	nadie	él	ellos	ustedes	ella

"No lo había visto **{mask}**."					
RoBERTa-base-BNE	nunca	antes	yo	todavía	aún
RoBERTa-large-BNE	nunca	antes	.	aún	en
BETO	antes	nunca	así	jamás	trabajar
mBERT	él	que	(	,	nunca
BERTIN	él	hoy	ayer	tú	todo
ELECTRA	antes	nunca	venir	aún	todavía

Appendix III

While the main focus of the article is building language models, we also computed 300 dimensional word embeddings using FastText. Both the CBOW[22] and Skip-gram [23] versions are publicly available on Zenodo. Using the clean data at document level described on the previous section, the processing took around 20 days on a HPC node[24] equipped with an AMD EPYC 7742 (@ 2.250GHz) processor with 128 threads. These embeddings are not evaluated and just provided as an additional resource.

Appendix IV

This Appendix shows several examples of the GPT2 models' text completion capabilities.

GPT2 text generation examples:

- Si vas a Barcelona tines que visitar → la Sagrada Familia.
- En el Barcelona Supercomputing Center nos dedicamos a → todo tipo de campos que requieran de gran potencia tecnológica.
- El BSC junto con la BNE desarrollan un modelo del lenguaje GPT2 en español que → se podría utilizar para la gestión de la producción y la investigación científica.
- Sin el esfuerzo de los médicos en la pandemia no hubieramos → podido salvar la vida a los nuestros.
- "Yo me vacuno seguro" es la → frase que acompaña en Facebook a la imagen de Jorge.
- En la Universidad de Deusto → y dentro de la acción social de la universidad, se ha invitado a más de 400 personas con el fin de trabajar el voluntariado desde una perspectiva ética y humanista.

GPT2-large text generation examples:

- Si vas a Barcelona tines que visitar → su iglesia, sus museos, el Modernisme (y su obra maestra el Modernismo), la estatua de Francesc de Coll, la Fuente Mágica, su teatro...

[22] https://zenodo.org/record/5044988
[23] https://zenodo.org/record/5046525
[24] https://www.bsc.es/innovation-and-services/technical-information-cte-amd

- En el Barcelona Supercomputing Center nos dedicamos a → impulsar y desarrollar la investigación en supercomputación.
- El BSC junto con la BNE desarrollan un modelo del lenguaje GPT2 en español que → permitirá estudiar el lenguaje desde un enfoque de lenguaje natural.
- Sin el esfuerzo de los médicos en la pandemia no hubieramos → podido salvar a los enfermos.
- "Yo me vacuno seguro" es la → frase que ha escogido un joven de 24 años.
- En la Universidad de Deusto → nos gusta pensar que tenemos que estar muy al día en todo para poder adaptarnos al ritmo de los tiempos en los que vivimos.

Procesamiento del Lenguaje Natural, Revista nº 68, marzo de 2022, pp. 61-70 recibido 19-12-2021 revisado 04-02-2022 aceptado 09-02-2022

Old English morphological inflection generation with UniMorph. Assessment with a relational database and training guidelines

Generación de flexión morfológica con UniMorph. Evaluación con base de datos relacional y pautas de entrenamiento

Javier Martín Arista
Universidad de La Rioja
javier.martin@unirioja.es

Abstract: The aim of this article is to assess the morphological inflection generation of Old English of the UniMorph data set. The method of this study is based on McCarthy et al.´s (2020) model of generation of putative morphological paradigms. The assessment includes inflections (morphological features and values), inflectional forms and stems. The question is also addressed of plausibility, understood as the effective attestedness of an inflectional form. The assessment tasks are carried out in a relational database specifically designed for filing and comparing the relevant data sets, including treebanks and databases of Old English lexicographical and textual sources. The overall conclusion is that the Old English UniMorph data set is consistent and robust. On the basis of the assessment, however, training guidelines of the generation model are proposed that include characters, diacritical marks, the prefix ge- in verbs, the superlative grade of adjectives, the adjectivally inflected participle and some local shortcomings.
Keywords: morphological inflection generation, UniMorph, relational database, treebank, Old English.

Resumen: El propósito de este artículo es evaluar la generación morfológica flexiva del set de datos UniMorph. El método del estudio se basa en el modelo de generación de paradigmas morfológicos putativos propuesto por McCarthy et al. (2020). La evaluación incluye las flexiones (tanto los rasgos morfológicos como sus valores), las formas flexivas y los radicales. Se aborda también la cuestión de la plausibilidad, entendida como la atestiguación efectiva de una forma flexiva. Las tareas de evaluación se llevan a cabo en una base de datos relacional específicamente diseñada para almacenar y comparar los sets de datos relevantes, que incluyen bancos de datos y bases de datos recopilados a partir de fuentes lexicográficas y textuales del inglés antiguo. La conclusión general es que el set de datos del inglés antiguo de UniMorph es congruente y robusto. Sin embargo, sobre la base de la evaluación que se lleva a cabo en este estudio se proponen algunas líneas maestras para el entrenamiento del modelo relativas a los caracteres, diacríticos, el prefijo *ge-* en verbos, el grado superlativo del adjetivo, el participio flexionado de acuerdo con la declinación adjetival y algunos aspectos mejorables de tipo local.
Palabras clave: generación de morfología flexiva, UniMorph, base de datos relacional, banco de datos, inglés antiguo.

1 Introduction

This article engages in morphological inflection generation, which, according to Çöltekin (2019), is "the task of generating a word based on its lemma and morphological features. For example, given the German lemma *aufgeben* 'to give up' and the morphological tags {V.PTCP, PST}, the task is to predict the inflected form *aufgegeben*."

The target language of the study is Old English, the diachronic variety of English

spoken in England between approximately the 6th and the 11th centuries of the Christian Era. It belongs to the West-Germanic Group of the Indo-European family of languages and is characterised by its explicit generalised morphological inflection and its consistently Germanic lexicon. Around 3,000 texts that approximately comprise 3 words million have been kept, most written in the West-Saxon variety in the 9th and, above all, in the 10th. century. Synchronic and diatopic variation, as well as the lack of a written standard, result in the unpredictiveness of the spelling of a remakable number of textual forms, as has been remarked by authors like Johnson (2009). In this line, an algorithm devised for generating Old English forms, even at the level of the syntactic word, requires a thorough design and an extensive training, as Torre Alonso (2021) shows. Matters are further complicated by the randomness of textual transmission, throughout which the vast majority of the texts might have got substantitally modified with respect to the original version or simply lost. These aspects should be taken into account when the task of generating Old English is undertaken.

For these reasons, the aim of this article is to assess the Old English data set of morphological inflection generation provided by UniMorph and available from https://raw.githubusercontent.com/unimorph/ang/master/ang. The UniMorph Project (https://unimorph.github.io) has defined a universal schema for morphological annotation with which data sets from 142 languages have been annotated.

More specifically, this study intends to contribute to the training of an inflection model of Old English in two directions: by gauging the accuracy of the inflections, inflectional forms and lemmas of the Old English UniMorph data set and by presenting a number of guidelines for the training of the inflection model.

The scope of the article is restricted to the syntactic word, that is to say, morphologically simplex words, affixed words and compound words that are written as one segment. The sources include treebanks and relational databases from Old English lexicographical and textual sources.

The article is structured as follows. Section 2 presents the data sets and the method of the study, including the design and implementation of the relational database. Section 3 assesses the generation model as to inflections (morphological features and values), inflectional forms and stems. The question of plausibility is also raised in this section. Section 4 discusses some weak points of the generation model, including their relevance for further training. Finally, Section 5 draws the conclusions of the article.

2 Data sets and method

This study relies on two types of data sets, to wit, treebanks and relational databases. While treebanks (Böhmová et al., 2003) and databases can be described as computerised data table collections (Jurafsky and Martin, fc.), they differ from each other in two important respects, at least in the context of this study. Firstly, treebanks are available from the Internet in open access, whereas relational databases are not always public. Secondly, treebanks tend to be more available for linguistic comparison and analysis than relational databases. An important consequence of this is that treebanks usually represent final products whereas relational databases can be updated.

Beginning with the treebanks, two data sets belong to this category: the Old English segment of the UniMorph Project and the York annotated corpora of Old English, both the prose and the poetry segments.

UniMorph consists of a schema and a set of databases for cross-linguistic morphological annotation. Morphological inflection generation in UniMorph is based on the UniMorph Schema (Sylak-Glassman, fc.), which comprises 23 dimensions of meaning (morphological categories) and 212 features. For Old English, the UniMorph Schema has been applied to the major lexical categories noun, adjective and verb. The relevant features include: ACC (accusative case), ADJ (adjective), DAT (dative case), FEM (feminine gender), GEN (genitive case), IMP (imperative mode), IND (indicative mode), INS (instrumental case), LGSPEC1 (weak declension of the adjective), LGSPEC2 (strong declension of the adjective), MASC (masculine gender), N (noun), NEUT (neuter gender), NFIN (non-finite form of verb (infinitive and inflective infinitive)), NOM (nominative case), PL (plural number), PRS (present tense), PST (past tense), SBJV (subjunctive mode), SG (singular number), V (verb), V.PTCP (verbal participle). In the case of the adjective, the most inflective lexical class

(as it can be declined according to a weak and a strong declension and can be graded for the comparative and the superlative), the generation with UniMorph turns out 67 inflectional forms, some of which are presented for illustration in Figure 1 with the corresponding morphological tags.

aberendlic (ADJ;NEUT;SG;NOM;LGSPEC2)
…
aberendlicena (ADJ;FEM;PL;GEN;LGSPEC1)
…
aberendlicu (ADJ;NEUT;PL;ACC;LGSPEC2)
…
aberendlicum (ADJ;FEM;PL;DAT;LGSPEC2)
Figure 1. UniMorph inflectional forms and tags of *aberendlic* (extract).

The second treebank used as data set for this study comprises the prose and the poetry parts of the York corpora of Old English (hereafter YCOE): *The York-Toronto-Helsinki Parsed Corpus of Old English Prose* (1,500,000 words; Taylor et al., 2003) and *The York-Helsinki Parsed Corpus of Old English Poetry* (50,000 words; Pintzuk and Plug, 2001). The YCOE is morphologically tagged and syntactically annotated. It comprises a POS (part of speech) file and a PSD (syntactic parsing) file for each text.

Turning to the relational databases, this study draws on unlemmatised and lemmatised data sets. *The Dictionary of Old English web corpus* (Healey et al., 2004), henceforth DOEC, contains 3,000,000 words. It is not lemmatised, neither does it provide morphological tagging or syntactic annotation, but it is generally considered to gather all the written records of the language. The DOEC was compiled as the corpus of the *Dictionary of Old English* (DOE; Healey, 2018), which has published the letters A-I so far. This electronic dictionary can be accessed online and offers, along with meaning definitions and citations, attestations per lemma that can be searched by headword, attested spelling, part of speech and occurrence, among other criteria.

The other lemmatised data set is *ParCorOEv2. An open access annotated parallel corpus Old English-English* (ParCorOEv2; Martín Arista et al. 2021). ParCorOEv2 is a deeply annotated parallel corpus that is aligned at word level. It currently holds 110,000 records and another 140,000 are expected by the end of 2022. These lemmatised data sets are complementary as to headword spelling. While the DOE opts for a late spelling of headwords (10th-11th century), ParCorOEv2 renders a classical spelling (9th-10th century), in such a way that the combined used of the two sources provides a wider inventory for comparison.

The method of this study is based on McCarthy et al.´s (2020) model of generation of putative morphological paradigms, which comprises two steps, training and generation. At the step of training, the aim is to relate the existing lemmas to the existing paradigms through an inflection model; while at the generation step, the aim is to relate the extracted lemmas to the putative paradigms via a trained model. This study contributes to the training of an inflection model for Old English and, ultimately, to the congruence of the existing and the putative paradigms of the language. An assessment is carried out and training guidelines are defined with a view to improving the training of the model, so that it generalises well to new data of Old English. The assessment and the training guidelines revolve around two main aspects, namely, the morphological paradigms and the lemma set that is inputted to them. Since the paradigms consists of the morphological features and values (inflections) as well as their exponents (inflectional forms), a distiction must be drawn between the assessment of inflections, on the one hand, and the assessment of exponents, on the other hand. The quality of the lemma set determines the plausibility of the outcome of the generation of morphological inflection. The tasks that this method require include (1) the assessment of inflections: are there counterparts of the morphological features and values of the UniMorph Schema as applied to the Old English data set in other tagged data sets? (2) the assessment of inflectional forms: do the morphological exponents of the UniMorph Schema as applied to the Old English data set include the inflectional forms tagged in other data sets? And (3) the assessment of the lemma set: are there substantial differences between the lemma set of the UniMorph Old English data set and the lemma lists of other lemmatised sources? With this assessment, it will be possible to address the question of plausibility: (4) how many putative forms are attested in the written records? Tasks 1 and 2 require a data set with POS tagging, while task 3 calls for a lemmatised data set. Task 4 needs to rely on an

extensive unlemmatised inventory. For these reasons, the YCOE is used for the assessment of inflections and inflectional forms, while lemmas are assessed with respect to the DOE and PacCorOEv2. Task 4 should necessarily draw on a full inventory of the forms attested in the written records of Old English. The DOEC has been selected for this task.

Not only the amount of data but also the need for falsifiability advise the automation of the four tasks described above. To this end, a relational database has been specifically designed for the undertaking. It has been implemented in Claris FileMaker Pro software (version 19.3.2.206). The database consists of six layouts, five of which correspond to the data sets reviewed above (UniMorph, DOEC, DOE, YCOE and ParCorOEv2). The sixth is a summary layout that combines all the data sets.

The Old English data set of UniMorph has been downloaded in txt format and imported into the database. It comprises 42,068 inflectional forms with the corresponding lemmas (1,867).

The DOEC has been concorded and indexed with AntConc 3.5.9 (Anthony, 2020). The concordance to the DOEC has 3,075,444 lines, while there are 194,327 types in the index.

The DOE has been searched by headword and attested spelling. A total of 15,907 lemmas and 83,477 inflectional forms have been found.

The inflectional forms and morphological tags of the YCOE have been extracted with BBEdit (version 14.0.2). A total of 106,202 types, corresponding to 1,595,674 tokens, have been extracted. The resulting types have been edited with the characters <æ>, <ð> and <þ> and tagged for lexical category, on the basis of the YCOE POS labels given in the corpus manuals (https://www-users.york.ac.uk/~lang22/YCOE/doc/annotation/YcoeLite.htm#pos_labels). Figure 2 illustrates this process.

POS file

<T06940000100,1>_CODE De_FW scientia_FW ._.
coalcuin,Alc_[Warn_35]:1.2_ID +arest_ADVS^T ealre_Q^G +tingen_N^G +aighwylce_Q^I m+an_N^D is_BEPI to_TO secene_VB^D ,_, hw+at_WPRO^N seo_BEPS se_D^N so+de_ADJ^N

inflectional form	**morphological tag**	**lexical category**
ærest	ADVS^T	Adverb
ealre	Q^G	Adjective
þingen	N^G	Noun
æighwylce	Q^I	Adjective
mæn	N^D	Noun
is	BEPI	Verb
to	TO	Preposition
secene	VB^D	Verb
Hwæt	WPRO^N	Pronoun
Seo	BEPS	Verb
Se	D^N	Demonstrative
Soðe	ADJ^N	Adjective

Figure 2. Extraction of types from the YCOE and lexical category tagging.

When designing the relational database, the types from the YCOE represented the field of reference. The data from the other layouts have been imported as an update for the reference field. With these premises, the total amount of files in the relational database is 106,202. The summary layout consists of the following fields: YCOE inflectional form, YCOE morphological tag, YCOE lexical category, DOE lemma, DOE attestation, UniMorph morphological tag, UniMorph lemma, and ParCorOEv2 lemma. A file with these fields and their values is presented in Figure 3.

Field	**Value**
YCOE_inflectional_form	bit
YCOE_morphological_tag	BEPI
YCOE_lexical_category	verb
DOE_lemma	bītan
DOE_attestation	44
UniMorph_morphological_tag	V;IMP;SG
UniMorph_lemma	bitan
ParCorOEv2_lemma	bītan

Figure 3. The summary layout in the relational database.

3 Assessment with the relational database

This section gauges the accuracy of the Old English UniMorph data set from two perspectives. In the first place, the accuracy of the morphological paradigms is discussed. This includes the morphological features and values (inflections) as well as their exponents (inflectional forms). The assessment of morphological features and values is extensive, whereas the one of their exponents is restricted to the main lexical and morphological classes.

This part of the assessment depends on the YCOE layout of the relational database. In the second place, the question of the quality of the lemma inventory of UniMorph is raised. This part of the assessment is carried out with the DOE and the PacCorOEv2 layouts of the relational database.

The first question addressed in this section is whether or not there are counterparts of the morphological features and values of the UniMorph Schema as applied to the Old English data set in other tagged data sets. A total of 6,762 counterparts of UniMorph morphological tags have been found in the YCOE. They correspond to 94 different tags in the YCOE layout, which comprises a total of 94 tags (recall ratio 1). Apart from the different annotation formats, there is no complete coincidence between the two sets of tags for reasons of homography across categories or due to different criteria for category assignment between noun and adjectives or adjectives and verbs (regarding the participle). This kind of mismatch is illustrated in Figure 4.

YCOE tag	**UniMorph tag**
ADJ	N;DAT;SG
ADJR^N	N;NOM;SG
N	ADJ;FEM;PL;NOM;LGSPEC2
N	ADJ;NEUT;SG; ACC;LGSPEC1
N	V;IMP;SG
N	V;IND;PRS;1;SG
VBN^D	ADJ;NEUT; PL;DAT;LGSPEC1
VBN^N	ADJ;FEM;SG; ACC;LGSPEC2

Figure 4. Morphological tags in the YCOE and UniMorph.

The second question raised in this section is whether or not the morphological exponents of the UniMorph Schema as applied to the Old English data set include the inflectional forms tagged in other data sets. For this purpose, the forms tagged in the YCOE and in the UniMorph data set are compared. From the quantitative point of view, 6,762 UniMorph inflectional forms are filed and tagged in the YCOE, which represents a 0.16 recall ratio. On the qualitative side, the most representative morphological classes of the lexical categories represented in UniMorph (the adjective, the noun and the verb) are considered. Such morphological classes include the weak and the strong forms of the adjective, strong masculine nouns, strong verbs and weak verbs. Weak verbs with strong forms, strong verbs with weak forms, preterite-present verbs and irregular verbs (which are not tagged in UniMorph), as well as the minor declensions of the noun have been put aside. The adjective, the noun and the verb are discussed in turn.

The inflectional forms of the adjective *beald* ‘bold’ tagged in the YCOE can be seen in Figure 5. The corresponding tags in UniMorph, when available, are given in the right column. The unpredictable spelling *bald* (ADJ^N) is missing in the UniMorph tagging and, more importantly, the comparative *bealdran* (ADJR^N) and the superlative *baldeste* (ADJS^N) are not tagged in UniMorph.

YCOE Inflectional form	**YCOE tag**	**UniMorph tag**
bealdum	ADJ^D	ADJ;FEM;PL; DAT;LGSPEC 2
beald	ADJ^N	ADJ;NEUT;PL ACC;LGSPEC 2
bealda	ADJ^N	ADJ;MASC; SG; NOM; LGSPEC1
bealde	ADJ^N	ADJ;NEUT; SG;ACC; LGSPEC1
bealdne	ADJ^A	ADJ;MASC; SG;ACC; LGSPEC2
baldra	ADJR^N	-
bald	ADJ^N	-
baldeste	ADJS^N	-
bealdran	ADJR^N	-

Figure 5. Adjective in YCOE and UniMorph.

Figure 6 tabulates the inflectional forms of the strong noun with weak forms *ancor* ‘anchor’. UniMorph gives *ancras* (N;ACC;PL) *ancre* (N;DAT;SG), *ancrum* (N;DAT;PL), *ancra* (N;GEN;PL), *ancor* (N;NOM;SG) and *ancras* (N;ACC;PL) but misses the unpredicatable spellings of the strong singular nominative *ancer* (*ancor*), singular accusative *ankor* (*ancor*), singular dative *ancræ* (*ancre*), plural nominative *onceras* and *oncras* (*ancras*) and plural accusative *oncras* (*ancras*). More significantly, UniMorph misses the forms from

the weak declension of the noun, including the weak singular genitive *ancran*, singular dative *ancran*, as well as the plural nominative and plural accusative *ancran*. It is also worth commenting that syncopated forms like *ancras* and unsyncopated forms such as *anceras* co-occur in the inflectional paradigm.

YCOE Inflectional form	YCOE tag	UniMorph tag
ancran	N^A	-
ancras	N^A	N;ACC;PL
ancræ	N^D	-
ancran	N^D	-
ancre	N^D	N;DAT;SG
ancrum	N^D	N;DAT;PL
ancra	N^G	N;GEN;PL
ancran	N^G	-
ancer	N^N	-
anceras	N^N	-
ancra	N^N	N;GEN;PL
ancran	N^N	-
oncras	N^A	-
ancor	N^N	N;NOM;SG
ancras	N^N	N;ACC;PL
ankor	N^A	-
oncras	N^N	-

Figure 6. Masculine noun in YCOE and UniMorph.

The YCOE and UniMorph inflectional forms and tags of the strong verb (class III) *belgan* 'to become angry' can be seen in Figure 7. Out of 13 attested inflectional forms, UniMorph gives 3, *belgaþ* (V;IND;PRS;PL), *belge* (V;IND;PRS;1;SG) and *gebolgen* (V.PTCP;PST). Of the missing forms, *bealh* and *belh* show alternation <h/g> with respect to *bealg* and *belg*. The alternation, as such, is relatively predictable. With respect to *gebolgene*, the adjectival part of the inflection of the present and the past participle has not been distinguished in the UniMorph Old English data set, which does not seem consistent with the choice of the verbal and the adjectival lexical classes. It is also worth pointing out that UniMorph must have considered the prefix *ge-* as derivational, thus distinguishing *belgan* from *gebelgan*. It must be noted in this respect that differences in meaning between the simplex and the *ge*-prefixed verb are scarce and the separation between the simplex and the complex verb is more often a lexicographical decision than a linguistic fact. From the strictly linguistic point of view, the prefix *ge-* plays a central role in inflection as it canonically forms past participles (Cambell, 1987; Hogg and Fulk, 2011) as well as in inflectionally motivated derivation (Kastovsky, 1992; Martín Arista, 2012). If we put aside the *ge-* prefixed forms of *belgan*, UniMorph correctly generates 3 out of 5 attested inflectional forms and misses the relatively unpredictable <ea> spelling.

YCOE Inflectional form	YCOE tag	UniMorph tag
bealg	VBDI	-
bealh	VBDI	-
belgaþ	VBPI	V;IND;PRS;PL
belge	VBPS	V;IND; PRS;1;SG
belh	VBI	-
gebealg	VBDI	-
gebealh	VBDI	-
gebelg	VBI	-
gebelgan	VB	-
gebelge	VBPS	-
gebolgen	VBN^N	V.PTCP;PST
gebolgene	VBN^N	-
gebulgon	VBDI	-

Figure 7. Strong verb in YCOE and UniMorph.

Figure 8 carries out this analysis with respect to the class 1 wead verb *rādan* 'to advise'. There are 23 tagged forms in the YCOE, in contradistinction to the 12 found in UniMorph. Surprisingly, UniMorph generates forms with consonant gemination like *rædde* (V;IND;PST;3;SG) but other geminated form such as the preterite plural *ræddan* are missing. This one, however, could be missing on the basis of the unpredictable spelling *ræddan* for *ræddon*, which has been generated by UniMorph. It is worth pointing out that the interchangeability of the eth and the thorn spelling to represent the voiceless and voiced dental allophones has not been taken into account in UniMorph, given that forms with thorn like *rædaþ* have been generated but the corresponding form with eth (*rædað*) has been missed. On the other hand, the inflection of the infinitive (*to rædanne*) has been generated in UniMorph but has not been tagged in the YCOE because it is not attested in the DOE. It also deserves a word of comment that the inflected present participles *rædene*, *rædendne*, *rædanne* and *rædendan* are missing in the

UniMorph generation. The spelling of *rædd*, *rætst*, *redst* and *ret* is unpredictable.

YCOE Inflectional form	YCOE tag	UniMorph tag
rædende	VAG^A	V.PTCP;PRS
rædan	VB	V;NFIN
rædene	VB^D	-
rædde	VBD	V;IND; PST;3;SG
ræddan	VBDI	-
ræddon	VBDI	V;IND;PST;PL
redon	VBDI	V;IND;PST;PL
rædd	VBN	-
ræded	VBN	V.PTCP;PST
ræden	VBN	N;NOM;SG
rædað	VBPI	-
rædaþ	VBPI	V;IND;PRS;PL
rædeþ	VBPI	V;IND; PRS;3;SG
redst	VBPI	-
rædan	VBPS	V;NFIN
rædendne	VAG^A	-
rædanne	VB^D	-
rædeð	VBPI	-
ræden	VBPS	N;NOM;SG
rædendan	VAG^G	-
rædende	VAG	V.PTCP;PRS
rætst	VBPI	-
ret	VBI	-

Figure 8. Weak verb in YCOE and UniMorph.

The third question raised in this section has to do with the lemma inventory of UniMorph. That is to say, are there substantial differences between the lemma set of the UniMorph Old English data set and the lemma lists of other lemmatised sources? Since lemmas are used as reference forms (the singular nominative of nouns and adjectives and the infinitive of verbs), this part constitutes, above all, an assessment of the quality of the stems geared to the plausibility of the putative language. To answer this question, the UniMorph data set and the ones from the DOE and PacCorOEv2 have been compared. The comparison of the stems throws the following results. The UniMorph data set inflects 1,867 different stems, from the adjectival, nominal and verbal classes. Of these, 1,069 correspond to the letters A-I (which have already been published by the DOE). Within the letters A-I, 827 stems of UniMorph have been found in the DOE and another 227 have a correlate in ParCorOEv2. This adds up to a total of 1,054 stems out of 1,069, which throws a recall ratio of 0.98.

Finally, this section addresses the question of plausibility. Once the stems, the inflectional features and the values inflections have been it is necessary to relate the generated inflectional forms to the ones attested in the unlemmatised data sets. The concept of plausibility is relevant at this point. Plausibility is the degree of convergence of the putative language generated with the UniMorph Schema and the attested language extracted from the DOEC. The analysis shows that 18,820 out of 42,067 of the generated UniMorph inflections are attested in the DOEC (recall ratio 0.44). By categories, these totals can be broken down as presented in Table 1 (adjectives), Table 2 (nouns) and Table 3 (verbs). Tables 1-3 tabulate the amount of forms present in both data sets.

	UniMorph	DOEC
Gender		
ADJ;FEM	509	534
ADJ;MASC	412	455
ADJ;NEUT	590	636
Declension		
ADJ;SPEC1	901	965
ADJ;SPEC2	651	704

Table 1. Attestedness of UniMorph inflectional forms (adjectives).

	UniMorph	DOEC
Case		
N;ACC	321	380
N;DAT	1,074	1,150
N;GEN	368	394
N;NOM	696	753
Number		
N;SG	1,702	1,829
N;PL	757	848

Table 2. Attestedness of UniMorph inflectional forms (nouns).

Remarkable differences arise when class totals are considered. Whereas 4,269 UniMorph generated nominal forms are attested from a total of 7,280 (recall ratio 0.58), the corresponding percentages in adjectives and verbs are much lower: 8,205 out of 18,712 adjectives (recall ratio 0.43) and 6,346 of 16,075 (recall ratio 0.39).

	UniMorph	DOEC
Mode		
V;IMP	570	633
V;IND	1,609	1,947
V;SBJV	1,237	1,318
Number		
V;SG	2,453	2,775
V;PL	963	1,123
Tense		
V;PRS	875	1,534
V;PST		
Non-finite forms		
V;NFIN	679	722
V.PTCP;PRS	423	446
V.PTCP;PST	366	383

Table 3. Attestedness of UniMorph inflectional forms (verbs).

4 Discussion

Two main lessons can be learned from the assessment of UniMorph morphological inflection generation presented in Section 3. The first has to do with the concept of putative language. While the putative language generated with the UniMorph schema is adequate given the quality of the stems and the inflections, both features and values, its plausibility is relatively low. The comparison of the UniMorph and the DOEC data sets suggests that the grammatically canonical inflectional paradigms are scarcely attested in the written records. This is particularly the case with the lexical class of the verb.

This leads us to the next lesson that can be learned from the data presented in Section 3. Plausibility, defined as the effective attestation of the grammatically canonical inflectional paradigms, is a fully random consequence of the process of textual transmission and, consequently, cannot be considered a weak point of the Old English UniMorph data set. This data set has proved robust in terms of the choice of stems, morphological features and values, but it also presents some weak points which are summarised in the remainder of this section and discussed as to their relevance for further training of the generation model.

To begin with, a number of unpredictable spellings have been mentioned that the UniMorph data set misses. However, such local irregularities do not seem compatible with a framework geared to cross-linguistic comparison, which seeks regularities rather than highly language-specific phenomena.

Other local shortcomings, which may be revised, affect the singular masculine accusative from the strong adjectival declension (ADJ;MASC;SG;ACC;LGSPEC2), which is mistaken for the plural feminine (ADJ;FEM;PL;NOM;LGSPEC2) in instances like *gyldenne* (*gylden* ‘golden’), *mihtigne* (*mihtig* ‘mighty’), *eadigne* (*ēadig* ‘wealthy), *elþeodigne* (*elþēodig* ‘foreign’) and *hefigne* (*hefig* ‘heavy’). Furthermore, nouns are not classified by gender, which might result in the wrong categorial tagging of at least thirty adjectival inflectional forms, such as *middan* (*midde* ‘middle), *fyrene* (*fȳren* ‘of fire’), *neowe* (*niwe* ‘new’), *woge* (*woh* ‘perverse’), etc., which are tagged as dative nouns.

More general questions include, in the first place, characters and diacritical marks. As for characters, problematic choices of the UniMorph data set include the character <ƿ> (wynn) to represent the grapheme <w>, which appears in 145 forms; and the letter <þ> (thorn) to represent the interchangeable pair <þ/ð>, which affects 815 forms. Editors of Old English texts do not use the wynn and tend to prefer the eth over the thorn, in such a way that the letter eth, as a general rule, subsumes <þ> and <ð>. Regarding diacritics, marking vocalic length and palatalisation in inflectional forms, as UniMorph does, is completely unprecedented, with the exception of some teaching materials. Finally, it is necessary to indicate the vocalic length of lemmas because vowel length is meaningful in Old English. While the changes of characters and diacritics would certainly contribute to the standarisation of the data set, the marking of vocalic length in lemmas would improve the applicability of the data set to subsequent analysis.

Also of general import is the question of the verbal prefix *ge-*. Its degree of generalisation suggests that both the simplex and the *ge-* affixed forms should be conjugated for all verbs. In this respect, it must be borne in mind that the prefix *ge-* is attached to 8,337 forms of verbs in the YCOE, out of a total of 33,986 verbal tokens.

The adjectival inflection of present and past participles is ignored in the current state of the UniMorph data set. Even if we put aside proto-auxiliary verbs like *bēon* ‘to be’ and *habban* ‘to have’, there are 1,469 present participles with

agreement traceable to the nominal head in the YCOE, and 3,082 past participles.

Something similar happens to adjective gradation, which has been generated only partially. There are 895 adjectives graded for the superlative in the YCOE, of which 78 only have been generated in the UniMorph data set; and another 956 comparatives, of which 78 only have been processed in the data set at stake. This represents a recall ratio of 0.08 in adjective gradation. It must be remarked in this respect that the size of the YCOE is approximately one half of the DOEC, which contains all the written records of the language. This ratio makes the figures just given even more significative.

To close this section, a comparison is presented that subsumes stem and inflection quality. The stems and inflections of UniMorph can be found in 9,795 inflectional forms of the YCOE (recall ratio 0.09). The stems of ParCorOEv2 enhanced with the training suggestions made in this section have 49,264 correlate inflectional forms in the YCOE (recall ratio 0.46). This comparison must be taken with caution because the lemma set of UniMorph is different from ParCorOEv2. As has been shown above, the morphological generation and the choice of unlemmatised forms in UniMorph are consistent when considered independently. On the other hand, this overall assessment of stem plus inflection in terms of plausibility indicates that the Old English UniMorph data set should address, at least, the issues raised in this discussion. Finally, further research is needed in the relevance for other historical languages of the method for gauging plausibility put forward in this article.

5 Conclusion

This article has assessed the morphological inflection generation of Old English of the UniMorph data set, including inflections (morphological features and values), inflectional forms, stems and plausability. Although this data set is consistent and robust, training guidelines of the generation model have been proposed that include characters, diacritical marks, the verbal prefix *ge-*, the superlative grade of adjectives, the participle with adjectival inflection and some local shortcomings.

Acknowledgements

Grant PRX19/00389 and grant PID2020-119200GB-100, funded by Ministerio de Ciencia, Innovación y Universidades.

References

Anthony, L. 2020. AntConc (Version 3.5.9) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Campbell, A. 1987. *Old English Grammar*. Oxford University Press, Oxford.

Çöltekin, Çağrı. 2019. Cross-lingual morphological inflection with explicit alignment. *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 71–79, Association for Computational Linguistics.

Cotterell, R., C. Kirov, J. Sylak-Glassman, G. Walther, E. Vylomova, A. D. McCarthy, K. Kann, S. Mielke, G. Nicolai, M. Silfverberg, D. Yarowsky, J. Eisner, and M. Hulden. 2018. The CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection. *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1-27, Association for Computational Linguistics.

Healey, A. (ed.), J. Wilkin, and X. Xiang. 2004. *The Dictionary of Old English web corpus*. Toronto: Dictionary of Old English Project, Centre for Medieval Studies, University of Toronto.

Healey, A. (ed.). 2018. *The Dictionary of Old English in electronic form A-I.* Toronto: Dictionary of Old English Project, Centre for Medieval Studies, University of Toronto.

Hogg, R. M., and R. D. Fulk. 2011. *A Grammar of Old English. Volume 2: Morphology*. Blackwell.

Johnson, B. 2009. *Using the Levenshtein algorithm for automatic lemmatization in Old English*. MA Thesis, The University of Georgia.

Jurafsky, D., and J. H. Martin. *Speech and Language Processing* (3rd. edition). Forthcoming.

Kastovsky, D. 1992. Semantics and vocabulary. In R. Hogg (ed.) *The Cambridge history of the English language I: The beginnings to 1066*, pages 290-408, Cambridge University Press, Cambridge.

Martín Arista, J. 2012. The Old English prefix *ge-*: A panchronic reappraisal. *Australian Journal of Linguistics*, 32(4):411–433.

Martín Arista, J., S. Domínguez Barragán, L. García Fernández, E. Ruíz Narbona, R. Torre Alonso, R., and R. Vea Escarza. 2021. *ParCorOEv2. An open access annotated parallel corpus Old English-English*. Nerthus Project, Universidad de La Rioja, www.nerthusproject.com.

McCarthy, A. D., C. Kirov, M. Grella, A. Nidhi, P. Xia, K. Gorman, E. Vylomova, S. J. Mielke, G. Nicolai, M. Silfverberg, T. Arkhangelskij, N. Krizhanovsky, A. Krizhanovsky, E. Klyachko, A. Sorokin, J. Mansfield, V. Ernštreits, Y. Pinter, C. L. Jacobs, R. Cotterell, M. Hulden, and D. Yarowsky. 2020. UniMorph 3.0: Universal Morphology. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 3922–3931, European LanguageResources Association.

Sylak-Glassman, J. 2016. *The Composition and Use of the Universal Morphological Feature Schema* (UniMorph Schema). Working draft, v. 2. Forthcoming.

Taylor, A., A. Warner, S. Pintzuk, and F. Beths. 2003. *The York-Toronto-Helsinki Parsed Corpus of Old English Prose* https://www-users.york.ac.uk/~lang22/YcoeHome1.htm.

Torre Alonso, R. 2021. Old English Class I Strong Verbs Lemmatization: A Morphological Generation Approach. *Studia Neophilologica.* To appear. DOI: 10.1080/00393274.2021.2010128.

Procesamiento del Lenguaje Natural, Revista nº 68, marzo de 2022, pp. 71-83 recibido 19-12-2021 revisado 07-02-2022 aceptado 10-02-2022

Risks of misinterpretation in the evaluation of Distant Supervision for Relation Extraction

Riesgos de interpretación errónea en la evaluación de la Supervisión Distante para la Extracción de Relaciones

Juan-Luis García-Mendoza[1], **Luis Villaseñor-Pineda**[1], **Felipe Orihuela-Espina**[1,2]
[1]Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Puebla, Mexico
{juanluis,villasen,f.orihuela-espina}@inaoep.mx
[2]University of Birmigham, Birmingham, United Kingdom

Abstract: Distant Supervision is frequently used for addressing Relation Extraction. The evaluation of Distant Supervision in Relation Extraction has been attempted through Precision-Recall curves and/or calculation of Precision at N elements. However, such evaluation is challenging because the labeling of the instances results from an automatic process that can introduce noise into the labels. Consequently, the labels are not necessarily correct, affecting the learning process and the interpretation of the evaluation results. Therefore, this research aims to show that the performance of the methods measured with the mentioned evaluation strategies varies significantly if the correct labels are used during the evaluation. Besides, based on the preceding, the current interpretation of the results of these measures is questioned. To this end, we manually labeled a subset of a well-known data set and evaluated the performance of 6 traditional Distant Supervision approaches. We demonstrate quantitative differences in the evaluation scores when considering manually versus automatically labeled subsets. Consequently, the ranking of performance among distant supervision methods is different with both labeled.
Keywords: Relation Extraction. Distant Supervision evaluation. Precision-Recall curves. Precision at N.

Resumen: La Supervisión Distante se utiliza con frecuencia para abordar la extracción de relaciones. La evaluación de la Supervisión Distante en la Extracción de Relaciones se ha realizado mediante curvas de Precisión-Cobertura y/o el cálculo de la Precisión en N elementos. Sin embargo, dicha evaluación es un desafío porque el etiquetado de las instancias es el resultado de un proceso automático. En consecuencia, las etiquetas no son necesariamente correctas, afectando no solo el proceso de aprendizaje sino también la interpretación de los resultados de la evaluación. El objetivo de esta investigación es mostrar que el desempeño de los métodos medido con las estrategias de evaluación mencionadas varía de manera significativa si se utilizan las etiquetas correctas durante la evaluación. Además, basado en lo anterior, se cuestiona la interpretación actual de los resultados de estas medidas. Con este fin, etiquetamos manualmente un subconjunto de un conjunto de datos y evaluamos el desempeño de 6 enfoques tradicionales de Supervisión Distante. Demostramos diferencias cuantitativas en los puntajes de evaluación al considerar subconjuntos etiquetados manualmente versus automáticamente. En consecuencia, el orden de desempeño entre los métodos de Supervisión Distante es diferente con ambos etiquetados.
Palabras clave: Extracción de Relaciones. evaluación de la Supervisión Distante. curvas de Precisión-Cobertura. Precisión en N.

1 Introduction

Relation Extraction (RE) is concerned with detecting and classifying predefined relations between entities identified in text (Piskorski and Yangarber, 2013). The traditional RE approach uses a supervised method to create the classifier(s) necessary to identify relations between pairs of named entities (Hearst, 1992; Agichtein and Gravano, 2000; Bunescu and Mooney, 2005). However, this process is slow and expensive; hence an alternative is the use of Distant Supervision (DS).

DS consists of automatically labeling the relations between each pair of named entities in a text using some pre-existing Knowledge Base (KB) (Mintz et al., 2009). For the automatic annotation of the data set with labeled relations, Mintz et al. (2009) assumed that given two entities that participate in a relation, *all* sentences in the data set that include these two entities express that relation (see Figure 1). However, it is common that a pair of entities in a sentence does not necessarily express a relation or may express several relations (see Figure 1). Hence, the assumtion proposed by Mintz et al. (2009) is too strong and often introduces false positives (which basically is noise in the labels) in the train and test sets. Later, Riedel et al. (2010) relaxed this assumption, assuming that "if two entities participate in a relation, *at least one* sentence that mentions these two entities might express that relation". This relaxation alleviates the problem of false positives in the *automatically* generated labels, but it does not fully fix it.

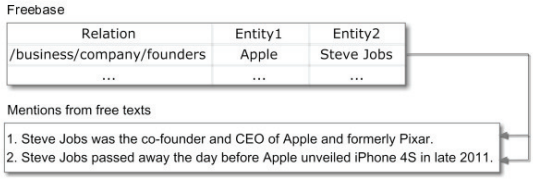


Figure 1: In this example, two sentences with the same pair of entities are automatically labeled with the same relation. Considering the *founders* relation, the first one will be correctly labeled while the second will not (Zeng et al., 2015).

Unfortunately, the evaluation of DS methods is complicated because there is no set correctly labeled to check their performance. Considering this, alternative evaluation methods have been proposed, such as the Precision-Recall (PR) curves or Precision at N (P@N) elements (Mintz et al., 2009). However, these measures are calculated using data labeled with the same automatic process; that is, the labels are not necessarily correct, impairing the calculation of the evaluation results.

This paper[1] aims to analyze the use of these evaluation measures showing that when the methods are evaluated using a correctly labeled set, the performance of the algorithms for DS reported so far varies substantially, thus questioning the current interpretation of the evaluation methods. We assessed the performance of 6 DS algorithms with PR curves and P@N analysis, with a correctly labeled set and with automatically generated labels, and compared the outcomes.

Our contributions can be summarized as follows:

- PR curves and P@N performance measures are critically revisited under competing scenarios of *manual* and *automatic* labeling.
- All sentences with a relation other than NA from the *New York Times* (NYT2010)[2] data set proposed by (Riedel, Yao, and McCallum, 2010) was crowdlabeled using MTurk[3]. So far, the manually annotated datasets for the task do not include all these sentences, which is a strength of this research. We argued that this affords better guarantees over the performance assessment in this task.
- We show that under current practice, performance measures for DS in RE may be misinterpreted when evaluation is carried out over *automatic* potentially noisy labeling.

In general, these contributions can positively impact the DS task evaluation. So far, the evaluation of this task is performed on *automatically* labeled partitions that may introduce incorrect labels. With the *manual* review of the test partition of well-known data set in DS, the performance comparisons of different methods are more reliable. In addition, precision, recall and F1 measures can

[1] Source available at https://github.com/juanluis17/distant-supervision-dataset-evaluation

[2] http://iesl.cs.umass.edu/riedel/ecml/

[3] Mechanical Turk, MTurk, is a human annotation service provided by Amazon.

be incorporated.

2 Related Work

The state-of-the-art in DS includes several solutions using different Deep Learning architectures. One of the first networks was the *Piecewise Convolutional Neural Networks* (PCNN) proposed by Zeng et al. (2015) based on *Convolutional Neural Networks* (CNN) (Zeng et al., 2014). This network incorporates bags of sentences to handle the noise on the labels. A bag of sentences contains sentences that have the same entities pair. Also, it includes a piecewise max-pooling layer "to capture structural information between two entities". Later, different attention mechanisms were incorporated into CNN and PCNN. In (Lin et al., 2016; Ji et al., 2017) an attention mechanism at sentences level (CNN_ATT and PCNN_ATT) in multiple instances was proposed to use the information of all sentences in the bag. Also, in (Ji et al., 2017) description about entities was included. Zhou et al. (2018) select from the bag several instances related to the label to predict the relations and use a word-level attention mechanism to highlight essential parts of the sentence dynamically. Besides, in (Jat, Khandelwal, and Talukdar, 2018), the *Bidirectional Gated Recurrent Unit* architecture was proposed with an attention mechanism over words to identify which key phrases are used (BGWA). Ye and Ling (Ye and Ling, 2019) used intra-bag and inter-bag attention mechanisms while in (Lin et al., 2016; Ji et al., 2017) it is only performed intra-bag, which ignores when all sentences in the bag are false positives. Moreover, Vashishth et al. (2018) propose to RESIDE that uses knowledge base information such as the entity type and relations alias to predict the correct relation. In addition, *Convolutional Graph Networks* (Defferrard, Bresson, and Vandergheynst, 2016) are used over dependency tree for modeling the syntactic information and capturing long-range dependencies. This information and the words and positions embeddings are used to encode the entire sentence. Finally, Bastos et al. (2021) proposed a method using an aggregator that obtains a homogeneous representation with a Graph Neural Network. This representation merges information from the sentence, relation, and the two entities (considering attributes like entity label, entity alias, entity description and the entity type).

Many of these methods have been evaluated with the test partition of the NYT2010 data set. This partition was automatically labeled under some heuristics, and consequently, some instances have been associated with an incorrect label. Given the absence of an adequate gold standard, precision, recall, and F1 measures have not been used to evaluate these methods. Mintz et al. (2009) used, for the first time, the PR Curves and P@N measures in an attempt to evaluate the DS task. These authors stated that PR curves "gives a rough measure of precision without requiring expensive human evaluation, making it useful for parameter setting". In such a case, "rough" is not an accurate statement. Therefore, performance measured with PR curves is dependent on the amount and distribution of noise in the labels. These curves constructed from *automatic* labels are a simple approximation of the performance of DS methods. Despite this problem, several authors continued using PR curves to evaluate and compare the performance of the proposed DS methods, probably leading to misinterpretations (Surdeanu et al., 2012; Zeng et al., 2015; Lin et al., 2016; Jat, Khandelwal, and Talukdar, 2018; Vashishth et al., 2018; Wu, Fan, and Zhang, 2019; Xu and Barbosa, 2019; Ye and Ling, 2019; Bastos et al., 2021; Nadgeri et al., 2021). In addition, P@N has been used in DS with 10, 30, 100, 200, 300, and 500 as the value of N. In P@N, the first N elements represent the most reliable answers of the classifier based on the ranking score. Lin et al. (2016), and Liu et al. (2017) reported P@100, P@200, and P@300 by randomly extracting one sentence for each pair of entities, two sentences or using them all. This evaluation, like in (Mintz et al., 2009), must be done manually on each execution because of the noise inherent to the *automatic* labels. Unfortunately, many works did not explicitly report whether and how the review was done manually (Lin et al., 2016; Liu et al., 2017; Wu, Fan, and Zhang, 2019; Vashishth et al., 2018; Ye and Ling, 2019).

Because of the noise that *automatic* labeling introduces, several efforts have been made to build a *gold standard* to evaluate the DS task. First, Mintz et al. (2009) used MTurk service for manual evaluation of P@N. The first 100 instances of each of the top 10 relations were sent to MTurk. Hoffmann et al.

(2011) manually labeled 1000 sentences from the NYT2010 data set to report the results of their method. These authors stated that "These results provide a good approximation to the true precision but can overestimate the actual recall since we did not manually check the much larger set of sentences where no approach predicted extractions". Based on these 1000 annotated instances, in (Ren et al., 2017) 395 were used as test partition. However, in these instances, there is no more than one sentence per entity pair (Jia et al., 2019). Later, Jiang et al. (2018) label 2040 randomly chosen instances of the NYT2010 data set, including the relation NA. In (Jiang et al., 2018), the performance of 4 DS methods is compared with the automatically annotated NYT2010 data set and the manually annotated data sets proposed by Hoffmann et al. (2011) and Jiang et al. (2018). However, a disadvantage of these data sets is that they do not include the entire NYT2010 test partition. Furthermore, in these papers, the measures of the DS task (i.e., PR curves and P@N) were not studied except for [1], which includes PR curves only. Besides, statistical validations were not carried out, nor were the selection criteria of the instances expressed. Finally, precision, recall, and F1 measures were not reported in most DS papers. Only Hoffmann et al. (2011) reported these measures on the 1000 annotated instances.

3 Background

3.1 Precision-Recall curves

PR curves are frequently used in binary classification (Davis and Goadrich, 2006) and, within this generic problem, in Information Retrieval (IR) (Manning, Raghavan, and Schütze, 2008). PR curves plot precision versus recall for a varying decision threshold parameter in binary classification (Keilwagen, Grosse, and Grau, 2014). These curves are calculated from the (assumed) true label and a score given by the classifier. This analysis is closely related to the Receiver-Operator Curve (ROC) analysis (Davis and Goadrich, 2006) widely used in statistics. However conveniently, for IR purposes, the PR curves can be built without the true negatives (TN). To get a scalar score, the area under PR curves (AUC) can be calculated by using the composite trapezoidal method (Davis and Goadrich, 2006).

Let Γ be a threshold set defined over classifier scores, and Ψ be a vector of descending ordered scores given by a classifier. The Precision and Recall for a threshold $\gamma \in \Gamma$ are calculated using the equations 1 and 2 respectively $\forall \psi \in \Psi \mid \psi > \gamma$.

$$P_\gamma = \frac{TP_\gamma}{TP_\gamma + FP_\gamma} \qquad \gamma \in \Gamma \qquad (1)$$

$$R_\gamma = \frac{TP_\gamma}{TP_\gamma + FN_\gamma} \qquad \gamma \in \Gamma \qquad (2)$$

where TP are positive examples correctly labeled as positives, FP are negative examples mislabelled as positives and FN are positive examples incorrectly labeled as negative.

To obtain the set of pairs (R_γ, P_γ) in the PR curve, we iterate over Γ as per Equation 3:

$$PR\_Curve(\gamma) = \{(R_\gamma, P_\gamma) : \gamma \in \Gamma\} \qquad (3)$$

3.2 Precision at N

The P@N in Equation 4 measures the number of correct elements in a window of N elements (Manning, Raghavan, and Schütze, 2008).

$$P@N = \frac{|TP \cap R_N|}{N} \qquad (4)$$

The TP (positive examples correctly labeled as positives) is calculated by manual evaluation. The P@N is frequently used in IR to measure the precision in a subset of retrieved elements R_N, with N the cardinality of the set. According to (Manning, Raghavan, and Schütze, 2008), it has the advantage of not requiring any estimate of the size of the set of relevant elements. P@N has been used in DS by multiple authors, but in most cases, this has been on the automatically labeled data set (with noisy labels) (Zeng et al., 2015; Lin et al., 2016; Ji et al., 2017; He et al., 2018; Wang et al., 2018; Wu, Fan, and Zhang, 2019; Ye and Ling, 2019; Bastos et al., 2021; Nadgeri et al., 2021).

4 Methodology

4.1 Dataset preparation

In order to establish whether there are risks of misinterpreting the evaluation measures, we compared the performance of 6 DS methods assessed over *manually-generated* labels and *automatically-generated* labels. We depart from the NYT2010 data set for the DS

task. This data set includes 53 relations types, including $\mathcal{NA}$, when there is no relation. Originally, this data set was labeled *automatically*. The *train* partition has 522611 instances (sentence that may or may not contain a relation), 279226 unique entity pairs and 154929 instances with a relation other than $\mathcal{NA}$. We use this training partition, with the *automatically* generated labels, to train the algorithms. In turn, the test partition has 172448 instances, 96678 unique entity pairs and 6444 instances with a relation other than $\mathcal{NA}$. From this last partition, two test partitions with *manual* labels were built and used in this work.

In the first test partition, 430 instances were selected for *manual* revision. The instances selection to be reviewed was made by choosing one instance from each relation at random during 20 iterations. During *manual* revision, 88 duplicate instances and 18 that have unclear relations were found and removed. Thus, the remaining 324 instances were revised *manually* and constitute our first test partition (named *test_1*). Considering the 324 instances of the *test_1* partition, 158 (48.8%) changed their *automatic* label after their/the review, i.e., they were considered by a human to hold incorrect labels.

In the second test partition, the complete 6444 instances different from the relation $\mathcal{NA}$ were selected for *manual* revision. First, we curated the 6444 instances by removing invalid instances. An instance is considered invalid when the defined entities are not found in the sentence. A total of 6431 were found valid. Then, from the 6431 valid instances, we further eliminated 579 duplicate instances (containing the same sentence, entity pair, and relation). We publish the remaining 5852 instances on the MTurk for review by three reviewers. The reviewers only determined whether the sentence explicitly expressed the associated relation.

Finally, we consider an instance as noisy if at least two of the three judges decided that the relations were not expressed. 4801 instances did not vary their *automatic* label but 1051 did (17.9%). This partition was named *test_2*.

4.2 Selection of DS methods for comparison

The following DS methods were compared in their performance:

- PCNN (Zeng et al., 2015) and CNN: The authors used both PR curves and P@N for evaluation, and the labeling was performed manually. This was one of the first architectures to be used in DS.
- PCNN_ATT (Lin et al., 2016) and CNN_ATT: The authors incorporated an attention mechanism over instances. They used PR curves to determine the performance of the attention mechanism compared to other methods. Finally, P@N was calculated on automatically generated *automatic* labels.
- BGWA (Jat, Khandelwal, and Talukdar, 2018): It incorporates an attention mechanism over words and entities. Only the PR curves were used as a measure to compare the performance of BGWA concerning the rest.
- RESIDE (Vashishth et al., 2018): It combines syntactic information with entity types and relations aliases. Like (Lin et al., 2016), P@N was calculated automatically on *automatic* labels.

These methods were chosen because they use three different architectures. On the one hand, CNN and PCNN use a convolutional architecture to which an attention mechanism is then incorporated (CNN_ATT and PCNN_ATT). On the other hand, RESIDE uses Graph Convolution Networks and Bidirectional Gated Recurrent Unit (the latter used by BGWA) and incorporates information about entities and relations. The execution of these methods was done in the same way as defined in Github[4] without using the gradient descent optimizer. To compare the evaluation measures, we trained these methods with the NYT2010 train partition proposed by (Riedel, Yao, and McCallum, 2010). Then, we evaluate them with the *test_1* and *test_2* partitions on the *automatic* and *manual* labels (see Figure 2).

4.3 Experimental design

In order to fairly evaluate the performance obtained, replications are necessary to ensure that chance does not play a role in our results. The number of replications (sample size) was determined using power analysis. Power analysis refers to the estimation of the probability of correctly rejecting a false null hypothesis when a particular alternative hypothesis is true (Howell, 2012).

[4]https://github.com/malllabiisc/RESIDE

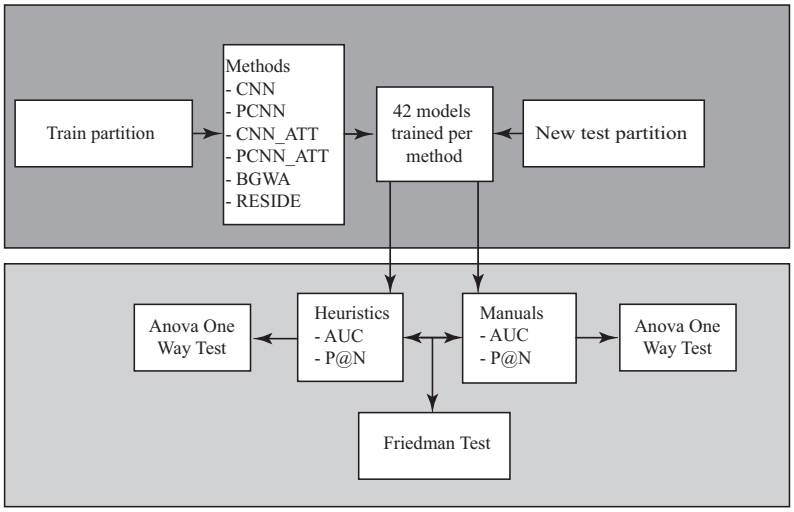


Figure 2: This diagram depicts the methodology followed in the current research. The top box illustrates the experiment design. The bottom box summarizes the statistical hypothesis testing followed.

The analysis depends on four factors: statistical significance, effect size, sample size and the statistical power itself. Fixing any three, yields the fourth for a given hypothesis model. The power analysis was estimated using the ANOVA One Way test for a desired significance level of 0.05, statistical power of $\beta = 0.95$ and assuming an effect size of Cohen's $d = 0.4$. As a result, 42 repetitions per treatment (i.e., algorithm to be compared) was obtained as the required sample size. The samples number here represents the number of executions for each method, that is, the replications required to detect an effect of the assumed size in the experiment.

From the results of the replications, the Friedman test was used to determine if there were differences in the ranking of the methods using automatic labels concerning manual labels. First, the Friedman test is used for one-way repeated measures analysis of variance by ranks (Friedman, 1940). This test only considers the number that each method occupies in the ranking and not the measure values. This is because the measure values are only used to determine ranking. Then, the ANOVA One Way test is applied on *automatic* and *manual* labels to know if there are significant differences between the results achieved by the methods. The ANOVA One Way test is used to test for differences among at least three groups, with the two-group case covered by the simpler t-test (Student, 1908; Howell, 2012). Finally, if there were significant differences, pairwise comparisons were made to observe which pair of methods showed differences. The two-by-two comparisons were made with t-test and Holm Correction (Holm, 1979). The significance threshold was set at $p < 0.05$.

5 Experiments

5.1 Precision-Recall curves

Performance on *test_1* partition

The Table 1 summarizes the AUC of the tested methods PR curves with *automatic* and *manual* labels on *test_1*. All methods increased their AUC with the *manual* labels with regards to their performances using the *automatic* ones, pointing to a systematic overall underestimation. Further, and more critically here, the order of the methods in terms of their performance varied significantly (Friedman: $\chi^2(2) = 373.46$, $p < 2.2e^{-16}$), i.e., they are all underestimated but not in the same extent. This suggests that using PR curves with *automatic* labels might not conferring the direct message one would expect otherwise in the DS evaluation task, and that for this scenario, such bias has to be considered during interpretation. Besides, significant differences were found with either *automatics* (ANOVA: $F(5, 246) = 746.9, p < 2e^{-16}$) and *manual* labels (ANOVA: $F(5, 246) = 520.8, p < 2e^{-16}$). In the case of pairwise comparisons, BGWA presents significant differences from the other methods for both labels.

The Figures 3a and 3b show the PR curves obtained by BGWA, RESIDE, PCNN,

Automatic labels		*Manual* labels	
Model	AUC	Model	AUC
BGWA	0.412 ± 0.026^{a}	BGWA	0.440 ± 0.023^{a}
CNN_ATT	0.194 ± 0.022^{b}	CNN_ATT	0.239 ± 0.031^{b}
CNN	0.193 ± 0.027^{b}	CNN	0.235 ± 0.027^{c}
RESIDE	0.191 ± 0.013^{b}	PCNN	0.209 ± 0.028^{d}
PCNN	0.158 ± 0.023^{c}	RESIDE	0.199 ± 0.020^{d}
PCNN_ATT	0.151 ± 0.025^{d}	PCNN_ATT	0.197 ± 0.029^{d}

[a] differences with rest of methods***.
[b] differences with BGWA***, PCNN*** and PCNN_ATT***.
[c] differences with rest of methods*** except PCNN_ATT.)
[d] differences with rest of methods*** except PCNN.
*, **, *** to indicate $p < 0.05$, $p < 0.01$ and $p < 0.001$ respectively.

[a] differences with rest of methods***.
[b] differences with rest of methods*** except CNN.
[c] differences with rest of methods*** except CNN_ATT.
[d] differences with BGWA***, CNN*** and CNN_ATT***.

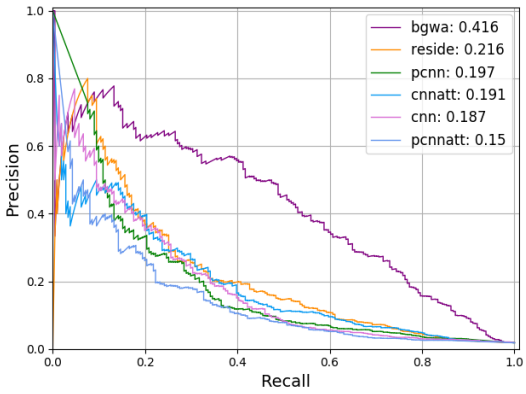
Table 1: AUC of the PR curves after 42 replications with *automatic* and *manual* labels on *test_1*.

PCNN_ATT, CNN and CNN_ATT in one execution made with *automatic* and *manual* labels respectively on *test_1*. It can be appreciated that the ordering of the algorithms according to their performance in terms of AUC varies when using the *manual* labels concerning the *automatic* ones (previously validated with Friedman test and multiples executions).

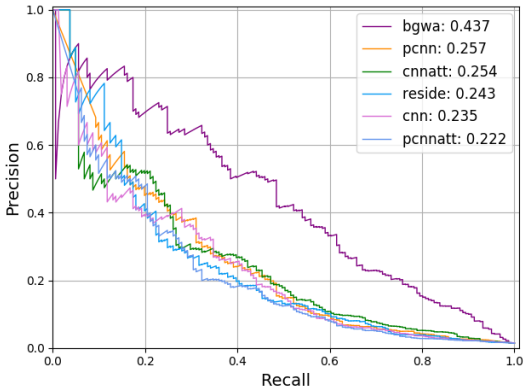
Performance on *test_2* partition

As with the *test_1* partition, the AUC values of the PR curves with *automatic* and *manual* labels on *test_2* were obtained (see Table 2). In these tables, similar values are observed with both labels. However, as in *test_1*, the order of the methods varied significantly (Friedman: $\chi^2(2) = 785.37$, $p < 2.2e^{-16}$. Similarly, significant differences were found with *automatics* labels (ANOVA: $F(5, 246) = 2097, p < 2e^{-16}$). Analogously, significant differences were found (ANOVA: $F(5, 246) = 1553, p < 2e^{-16}$) on *manual* labels. As in *test_1*, BGWA presents significant differences from the other methods for both labels in pairwise comparisons. However, there were no differences between PCNN_ATT, CNN_ATT and PCNN for *automatic* labels. Besides, no differences were found in the *manual* labels between PCNN_ATT and PCNN methods.

The Figures 4a and 4b show the PR curves in one execution made with *automatic* and *manual* labels respectively on *test_2*.



(a) *Automatic* labels



(b) *Manual* labels

Figure 3: PR curves corresponding to evaluation of the DS algorithms over *test_1* (one execution) set pick for verification in (a) *automatic* labels and (b) *manual* labels. The AUC of the PR curves is indicated beside each label in the legend.

5.2 Precision at N

Performance on *test_1* partition

The P@25 and P@50 subsets from the *test_1* partition were established in addition to all the instances (P@All). Table 3 shows that the order of the models remains the same for the first three models by increasing N, unlike the last three positions. The same happens with Table 4 where, in this case, the first two models are kept. The order of the models, as with the AUC, varied significantly for

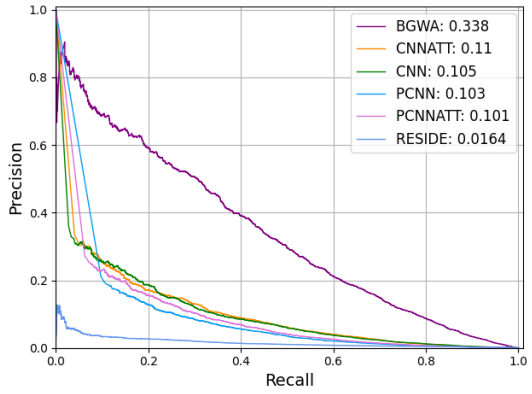
Automatic labels		*Manual* labels	
Model	AUC	Model	AUC
BGWA	0.339 ± 0.016^{a}	BGWA	0.345 ± 0.021^{a}
PCNN_ATT	0.112 ± 0.015^{b}	PCNN_ATT	0.118 ± 0.017^{b}
CNN_ATT	0.105 ± 0.017^{c}	PCNN	0.109 ± 0.020^{c}
PCNN	0.105 ± 0.018^{c}	CNN_ATT	0.106 ± 0.018^{d}
CNN	0.098 ± 0.016^{d}	CNN	0.098 ± 0.017^{e}
RESIDE	0.021 ± 0.006^{c}	RESIDE	0.028 ± 0.011^{f}

[a] differences with rest of methods***.
[b] differences with BGWA*** and CNN***.
[c] differences with BGWA***.
[d] differences with BGWA*** and PCNN_ATT***.

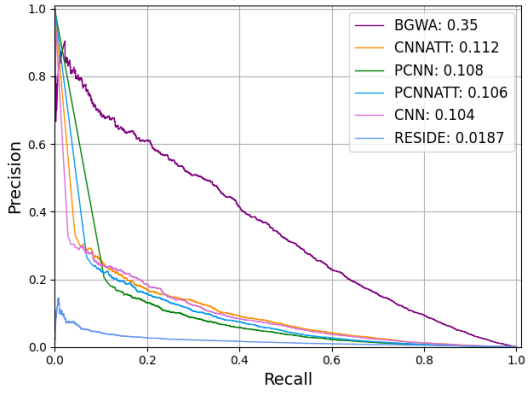
[a] differences with rest of methods***.
[b] differences with CNN*** and CNN_ATT*.
[c] differences with BGWA*** and CNN*.
[d] differences with BGWA*** and PCNN_ATT***.
[e] differences with BGWA***, PCNN_ATT*** and PCNN*.
[f] differences with BGWA***

*, **, *** to indicate $p < 0.05$, $p < 0.01$ and $p < 0.001$ respectively.

Table 2: AUC of the PR curves after 42 replications with *automatic* and *manual* labels on *test_2*.



(a) *Automatic* labels



(b) *Manual* labels

Figure 4: PR curves corresponding to evaluation of the DS algorithms over *test_2* (one execution) set pick for verification in (a) *automatic* labels and (b) *manual* labels. The AUC of the PR curves is indicated beside each label in the legend.

the *automatic* and *manual* labels on P@All (Friedman: $\chi^2(2) = 382.28$, $p < 2.2e^{-16}$). Similarly, there are significant differences in the performance of methods with *automatic* (ANOVA: $F(5, 246) = 210.8, p < 2e^{-16}$) and *manual* (ANOVA: $F(5, 246) = 255.6, p < 2e^{-16}$) labels. Then, two-by-two comparisons with Holm Correction (Holm, 1979) show significant differences with *automatic* labels between the BGWA and RESIDE models and the rest. Similarly, two-by-two comparisons show significant differences with manual labels between the BGWA model and the rest. In addition, PCNN_ATT has significant differences with the other models except for PCNN (in reverse order, it also happens). In this case, RESIDE only shows significant differences with BGWA, PCNN and PCNN_ATT.

Performance on *test_2* partition

In the same way as with *test_1*, the subsets P@25 and P@50 were established together with P@All, which includes the entire set. With both labeled, only two methods did not vary their order in the three subsets, BGWA and RESIDE (see Tables 5 and 6). In addition, the order of the methods using the P@All results varied significantly concerning the *automatic* and *manual* labels (Friedman: $\chi^2(2) = 369.55$, $p < 2.2e^{-16}$)[5]. Similarly, significant differences were found in the performance of the methods with *automatic* (ANOVA: $F(5, 246) = 1610, p < 2e^{-16}$) and *manual* (ANOVA: $F(5, 246) = 1265, p < 2e^{-16}$) labels. Then, in two-by-two comparisons with Holm Correction (Holm, 1979) there are no significant differences only between the CNN and CNN_ATT and PCNN and PCNN_ATT methods with both labeled.

[5] It should be noted that in all cases the Friedman test is used on the ranking of each execution, not only on the final results.

Model	P@25	Model	P@50	Model	P@All
BGWA	0.819±0.062	BGWA	0.730±0.041	BGWA	0.558±0.029
CNN	0.587±0.087	CNN	0.489±0.062	CNN	0.386±0.036
CNN_ATT	0.580±0.089	CNN_ATT	0.486±0.064	CNN_ATT	0.375±0.045
PCNN	0.554±0.087	PCNN_ATT	0.461±0.055	PCNN	0.362±0.037
RESIDE	0.552±0.074	PCNN	0.459±0.060	PCNN_ATT	0.351±0.040
PCNN_ATT	0.550±0.079	RESIDE	0.433±0.054	RESIDE	0.325±0.035

Table 3: P@25, P@50 and P@All after 42 replications with *automatic* labels on *test_1*.

Model	P@25	Model	P@50	Model	P@All
BGWA	0.715±0.079	BGWA	0.677±0.044	BGWA	0.585±0.033
RESIDE	0.555±0.075	RESIDE	0.489±0.043	RESIDE	0.376±0.037
CNN	0.551±0.089	CNN_ATT	0.465±0.061	CNN	0.370±0.035
CNN_ATT	0.544±0.089	CNN	0.459±0.059	CNN_ATT	0.370±0.044
PCNN	0.486±0.093	PCNN	0.401±0.062	PCNN	0.328±0.044
PCNN_ATT	0.458±0.096	PCNN_ATT	0.399±0.066	PCNN_ATT	0.325±0.041

Table 4: P@25, P@50 and P@All after 42 replications with *manual* labels on *test_1*.

6 Discussion

Our results indicate that the ranking of the methods, in terms of the AUC of the PR curves on *test_1* and *test_2* partition, differ depending on the labeling. This justifies our claim that the interpretation of the PR curves must be reconsidered when used for evaluating DS algorithms. PR curves using *automatic* labels as a reference is not an optimal way to compare methods performance in DS because it breaks a premise of the PR curves construction; that *true* labels are available. Several authors have based the comparison of their method on the PR curves on these labels (Riedel, Yao, and McCallum, 2010; Hoffmann et al., 2011; Surdeanu et al., 2012; Zeng et al., 2015; Lin et al., 2016; Jiang et al., 2016; Liu et al., 2017; Vashishth et al., 2018; Ru et al., 2018; Zhou et al., 2018; Wang et al., 2018; Jat, Khandelwal, and Talukdar, 2018; Wu, Fan, and Zhang, 2019; Xu and Barbosa, 2019; Ye and Ling, 2019; Bastos et al., 2021; Nadgeri et al., 2021). The classical interpretation does not provide guarantees as to which method is performing better or which one is more tolerant to noise in the labels.

The Section 5.2 has also confirmed that P@N is not being interpreted correctly in DS either. This is critical for the task at hand considering the unbalance in the data sets, variability among the relations, selection criteria, among others. There is no clear selection criterion that guarantees to choose the same instances for evaluating each of the methods. In other words, it is not guaranteed that the first instances chosen to evaluate one method are the same for another method. If the selection is based on the classifier's score, it varies from one execution to another. The same happens if the selection is random. For example, the first N instances can be of the same relation for a method. This indicates how good this method is for that relation. However, for the rest, its performance is not known. Also, sometimes, the P@N is calculated over *automatic* labels, whereas some works do it over *manual* labels. This is the case of the 6 methods used in this work. This further confuses P@N's interpretation. Furthermore, dispersion values are not reported in the previous works, which mathematically renders those works uninformative.

What was expressed above shows that PR curves and P@N measures are not currently being interpreted properly in DS due to the presence of noisy labels. Currently, we believe there are no reliable statistics regarding the actual performance of the DS methods. While the community agrees on a mathematically correct interpretation in this context, or new statistics are proposed for evaluating the performance of DS methods, a possible strategy to circumvent the deadlock is what was done here. That is, selecting multiple instances of the evaluation data set while maintaining its distribution (*test_1* partition). Then, perform a manual review of these instances using multiple raters. The main limitations of *test_1* partition are the instances number selected. This

Model	P@25	Model	P@50	Model	P@All
BGWA	0.804 ± 0.082	BGWA	0.762 ± 0.064	BGWA	0.019 ± 0.000
CNNATT	0.360 ± 0.112	CNN	0.357 ± 0.084	CNN	0.015 ± 0.000
CNN	0.346 ± 0.111	CNNATT	0.341 ± 0.087	CNNATT	0.015 ± 0.000
PCNNATT	0.273 ± 0.089	PCNNATT	0.268 ± 0.067	PCNN	0.014 ± 0.000
PCNN	0.252 ± 0.106	PCNN	0.233 ± 0.070	PCNNATT	0.014 ± 0.000
RESIDE	0.115 ± 0.095	RESIDE	0.129 ± 0.076	RESIDE	0.010 ± 0.000

Table 5: P@25, P@50 and P@All after 42 replications with *automatic* labels on *test_2*.

Model	P@25	Model	P@50	Model	P@All
BGWA	0.017 ± 0.000	BGWA	0.795 ± 0.083	BGWA	0.0168 ± 0.000
CNN	0.014 ± 0.000	CNNATT	0.343 ± 0.120	CNN	0.0137 ± 0.000
CNNATT	0.014 ± 0.000	CNN	0.320 ± 0.117	CNNATT	0.0135 ± 0.000
PCNN	0.013 ± 0.000	PCNNATT	0.255 ± 0.104	PCNN	0.0130 ± 0.000
PCNNATT	0.013 ± 0.000	PCNN	0.230 ± 0.099	PCNNATT	0.0129 ± 0.000
RESIDE	0.010 ± 0.000	RESIDE	0.150 ± 0.094	RESIDE	0.0103 ± 0.000

Table 6: P@25, P@50 and P@All after 42 replications with *manual* labels on *test_2*.

is why the *test_2* partition was labeled with multiple raters using MTurk. The advantage of this partition concerning *test_1* and those proposed by (Hoffmann et al., 2011), (Ren et al., 2017) and (Jiang et al., 2018) is that it is made up of all the instances of the NYT2010 data set test partition (only those different from NA were labeled with MTurk). From the *test_2* partition, the methods can be compared with precision, recall and F1 using the traditional interpretation. Besides, in (Jiang et al., 2018), although the performance of the CNN, PCNNN, CNN_ATT and PCNNN_ATT methods are analyzed, the P@N measure, the BGWA and RESIDE methods and statistical validations are not included. One limitation of this evaluation alternative is that *manual* labeling of the test partition is expensive but only done once. In addition, experts in the area are needed for this labeling in most cases. However, in this work, it has been shown that the performance of the methods using *automatic* labeling can be misinterpreted.

7 Conclusions

Significant differences were found in the ranking of the methods regarding their performances when the performance is established according to the AUC of the PR curves between the evaluation using the *automatic* labels and the same data set with the *manual* labels. The largest AUCs were obtained using *manual* labels which speaks well of the capacity of the DS methods to handle noisy data as it is their core intention. Our results suggest that PR curves are currently not being interpreted correctly in DS. Furthermore, they suggest that the PR curves calculated using the *automatically* labeled data should not be used to compare the performance of DS methods. In addition, manual evaluation of the first N instances (P@N) does not cover the entire data set. The existing selection criteria for the instances to be manually reviewed are not deterministic, suggesting multiple executions of the method and the dispersion report. Besides, as they are being used, these measures are inconclusive as to the performance of those methods. Finally, we provided a partition that allows you to evaluate this task using labels manually reviewed by multiple raters. This partition also allows the use of precision, recall and F1 measures and will be available for use by the area community. In future work, we will analyze various DS methods using these two partitions and the traditional precision, recall, and F1 measures. In addition, we will continue to work on the DS evaluation methods.

Acknowledgments

The present work was supported by CONACyT/México (scholarship 937210 and grant CB-2015-01-257383). Additionally, the authors thank CONACYT for the computer resources provided through the INAOE Supercomputing Laboratory's Deep Learning Platform for Language Technologies.

References

Agichtein, E. and L. Gravano. 2000. Snowball: Extracting Relations from large Plain-Text Collections. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 85–94. ACM.

Bastos, A., A. Nadgeri, K. Singh, I. O. Mulang', S. Shekarpour, J. Hoffart, and M. Kaul. 2021. RECON: Relation Extraction using Knowledge Graph Context in a Graph Neural Network. In *Proceedings of the Web Conference 2021*, pages 1673–1685, Ljubljana.

Bunescu, R. C. and R. J. Mooney. 2005. A Shortest Path Dependency Kernel for Relation Extraction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 724–731, Vancouver,. Association for Computational Linguistics.

Davis, J. and M. Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*, pages 233–240, Pittsburgh, Pennsylvania, USA. ACM Press.

Defferrard, M., X. Bresson, and P. Vandergheynst. 2016. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. In *Advances in neural information processing systems*, pages 3844–3852.

Friedman, M. 1940. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92.

He, Z., W. Chen, Z. Li, M. Zhang, W. Zhang, and M. Zhang. 2018. SEE: Syntax-Aware Entity Embedding for Neural Relation Extraction. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 5795–5802. Association for the Advancement of Artificial Intelligence.

Hearst, M. A. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th Conference on Computational Linguistics (COLING-92)*, pages 539–545, Nantes.

Hoffmann, R., C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting ofthe Association for Computational Linguistics*, pages 541–550, Portland, Oregon. Association for Computational Linguistics.

Holm, S. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Howell, D. C. 2012. *Statistical Methods for Psychology*. Cengage Learning ALL.

Jat, S., S. Khandelwal, and P. Talukdar. 2018. Improving Distantly Supervised Relation Extraction using Word and Entity Based Attention. *arXiv*, [cs.CL](1804.06987v1), apr.

Ji, G., K. Liu, S. He, and J. Zhao. 2017. Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Descriptions. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, pages 3060–3066.

Jia, W., D. Dai, X. Xiao, and H. Wu. 2019. ARNOR: Attention Regularization based Noise Reduction for Distant Supervision Relation Classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1399–1408, Florence. Association for Computational Linguistics.

Jiang, T., J. Liu, C.-Y. Lin, and Z. Sui. 2018. Revisiting distant supervision for relation extraction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Jiang, X., Q. Wang, P. Li, and B. Wang. 2016. Relation extraction with multi-instance multi-label convolutional neural networks. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1471–1480, Osaka.

Keilwagen, J., I. Grosse, and J. Grau. 2014. Area under Precision-Recall Curves for Weighted and Unweighted Data. *PLoS ONE*, 9(3):e92209, mar.

Lin, Y., S. Shen, Z. Liu, H. Luan, and M. Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting ofthe Association for Computational Linguistics*, pages 2124–2133, Berlin, Germany. Association for Computational Linguistics.

Liu, T., K. Wang, B. Chang, and Z. Sui. 2017. A Soft-label Method for Noise-tolerant Distantly Supervised Relation Extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1790–1795, Copenhagen, Denmark.

Manning, C. D., P. Raghavan, and H. Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press.

Mintz, M., S. Bills, R. Snow, and D. Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the 47th Annual Meeting of the ACL*, pages 1003–1011, Suntec, Singapore.

Nadgeri, A., A. Bastos, K. Singh, I. O. Mulang', J. Hoffart, S. Shekarpour, and V. Saraswat. 2021. KGPool: Dynamic Knowledge Graph Context Selection for Relation Extraction. *arXiv*, 2106.00459, jun.

Piskorski, J. and R. Yangarber. 2013. Information extraction: Past, Present and Future. In *Multi-source, Multilingual Information Extraction and Summarization 11*. Springer-Verlag Berlin Heidelberg, pages 23–49.

Ren, X., Z. Wu, W. He, M. Qu, C. R. Voss, H. Ji, T. F. Abdelzaher, and J. Han. 2017. CoType. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1015–1024, Perth, apr. International World Wide Web Conferences Steering Committee.

Riedel, S., L. Yao, and A. McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163, Berlin. Springer.

Ru, C., J. Tang, S. Li, S. Xie, and T. Wang. 2018. Using semantic similarity to reduce wrong labels in distant supervision for relation extraction. *Information Processing Management*, 54(4):593–608, jul.

Student. 1908. The Probable Error of a Mean. *Biometrika*, 6(1):1–25.

Surdeanu, M., J. Tibshirani, R. Nallapati, and C. D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465, Jeju Island, Korea. Association for Computational Linguistics.

Vashishth, S., R. Joshi, S. S. Prayaga, C. Bhattacharyya, and P. Talukdar. 2018. Reside: Improving Distantly-Supervised Neural Relation Extraction using Side Information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1257–1266, Brussels, Belgium. Association for Computational Linguistics.

Wang, G., W. Zhang, R. Wang, Y. Zhou, L. Chen, W. Zhang, H. Zhu, and H. Chen. 2018. Label-free distant supervision for relation extraction via knowledge graph embedding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2246–2255.

Wu, S., K. Fan, and Q. Zhang. 2019. Improving Distantly Supervised Relation Extraction with Neural Noise Converter and Conditional Optimal Selector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7273–7280, nov.

Xu, P. and D. Barbosa. 2019. Connecting Language and Knowledge with Heterogeneous Representations for Neural Relation Extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3201–3206, Minneapolis, Minnesota. Association for Computational Linguistics.

Ye, Z.-X. and Z.-H. Ling. 2019. Distant Supervision Relation Extraction with Intra-Bag and Inter-Bag Attentions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Hu-*

man Language Technologies, pages 2810–2819, Minneapolis, Minnesota. Association for Computational Linguistics.

Zeng, D., K. Liu, Y. Chen, and J. Zhao. 2015. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762, Lisbon, Portugal. Association for Computational Linguistics.

Zeng, D., K. Liu, S. Lai, G. Zhou, and J. Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland.

Zhou, P., J. Xu, Z. Qi, H. Bao, Z. Chen, and B. Xu. 2018. Distant supervision for relation extraction with hierarchical selective attention. *Neural Networks*, 108:240–247.

Procesamiento del Lenguaje Natural, Revista nº 68, marzo de 2022, pp. 85-97 recibido 19-12-2021 revisado 30-01-2022 aceptado 05-02-2022

Low-resource AMR-to-Text Generation: A Study on Brazilian Portuguese

Generación de Texto a partir de AMR en Contexto de Bajos Recursos: Un Estudio para el Portugués Brasileño

Marco Antonio Sobrevilla Cabezudo, Thiago Alexandre Salgueiro Pardo
Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
msobrevillac@usp.br, taspardo@icmc.usp.br

Abstract: This work presents a study of how varied strategies for tackling low-resource AMR-to-text generation for three approaches are helpful in Brazilian Portuguese. Specifically, we explore the helpfulness of additional *translated* corpus, different granularity levels in input representation, and three preprocessing steps. Results show that *translation* is useful. However, it must be used in each approach differently. In addition, finer-grained representations as characters and subwords improve the performance and reduce the bias on the development set, and preprocessing steps are helpful in different contexts, being delexicalisation and preordering the most important ones.
Keywords: AMR-to-Text Generation, Low-resource setting, Brazilian Portuguese.

Resumen: Este trabajo presenta un estudio de cómo diversas estrategias para abordar la generación de textos a partir de AMR en contextos de bajos recursos para tres enfoques son útiles en portugués brasileño. Específicamente, exploramos la utilidad de un corpus traducido, diferentes niveles de granularidad en la representación de entradas y tres técnicas de preprocesamiento. Los resultados muestran que el corpus traducido es útil. Sin embargo, debe usarse en cada enfoque de manera diferente. Además, las representaciones más detalladas, como las basadas en caracteres y subpalabras, mejoran el rendimiento y reducen el sesgo en el conjunto de validación, y los pasos de preprocesamiento son útiles en diferentes contextos, siendo la deslexicalización y el preordenamiento los más importantes.
Palabras clave: Generación de Texto a partir de AMR, Contexto de Bajos Recursos, Portugués Brasileño.

1 Introduction

Abstract Meaning Representation (AMR) is a semantic formalism that encodes the meaning of a sentence as a rooted, acyclic, labeled, and directed graph (Banarescu et al., 2013). This representation includes several semantic information, like semantic roles and named entities, among others.

AMR has become a relevant research topic in meaning representation, semantic parsing, and natural language generation (NLG). Its success is grounded on its attempt to abstract away from syntactic idiosyncrasies, and surface forms, its wide use of mature linguistic resources such as PropBank (Palmer, Gildea, and Kingsbury, 2005), and its usefulness on tasks like text summarisation (Liao, Lebanoff, and Liu, 2018), event detection (Li et al., 2015a) and machine translation (Song et al., 2019).

The goal of the AMR-to-Text generation task is to produce a text that represents the meaning encoded by an input AMR graph. For English, there are several works and approaches for this, as techniques of Statistical Machine Translation (Pourdamghani, Knight, and Hermjakob, 2016), tree and graph to string transducers (Flanigan et al., 2016) and, recently, neural models following sequence-to-sequence (Castro Ferreira et al., 2017; Konstas et al., 2017) and graph-to-sequence architectures (Beck, Haffari, and Cohn, 2018) or pretrained models (Mager et al., 2020; Ribeiro et al., 2020). For other languages, there are some multilingual work (Fan and Gardent, 2020) that tries to generate sentences in several languages. However, they use the AMR for English as in-

ISSN 1135-5948 DOI 10.26342/2022-68-6

put and do not capture some particular linguistic phenomena. In a different line, Sobrevilla Cabezudo, Mille, and Pardo (2019) try to generate Brazilian Portuguese (BP) sentences from the corresponding AMR for BP; nonetheless, the corpus is small (only 299 instances).

One problem that limits the research in other languages is the difficulty to get high-quality corpora (due to the difficult and expensive annotation task that it represents), resulting in smaller corpora and the inability for state-of-the-art methods to be replicated and/or achieve similar performance to the English ones.

It is well-known that the lack of data deteriorates the performance produced by neural models, which usually are data-hungry. To tackle this problem, some authors make use of data augmentation techniques, cross-lingual projection, and other strategies for increasing the corpus size (Hedderich et al., 2021). In the case of AMR-to-text generation, Sobrevilla Cabezudo, Mille, and Pardo (2019) proposed to translate both AMR and English sentences to their corresponding BP ones and then used the translated corpus as training/development set and a gold BP subset as test.

One problem associated with scarce corpus is data sparsity. Particularly, sparsity usually happens at input level in Natural Language Processing tasks. Word representation presents problems with unseen and rare words, resulting in low performance. Many works have proposed employing different granularities in input representation to solve this problem. The most commonly used are subwords (specifically Byte-pair encoding) (Sennrich, Haddow, and Birch, 2016) and characters, resulting in better results. In AMR-to-text generation, some work (Konstas et al., 2017; Mager et al., 2020) used finer-grained representations producing improvements; however, its benefits have not been studied in depth in low-resource settings.

This work explores three different strategies on three approaches for tackling low-resource AMR-to-text generation in Brazilian Portuguese. Specifically, we focus on machine translation and graph-to-sequence-based approaches and study the helpfulness of adding a *translated* corpus, using finer-grained representations and applying diverse preprocessing strategies.

It is worth noting that, even though the current state-of-the-art model for this task uses pretrained models (Mager et al., 2020; Ribeiro et al., 2020) and there are pretrained models for Brazilian Portuguese (Carmo et al., 2020), our goal is to show how to use simpler models and what kind of information could be helpful in low-resource settings or for other languages in which there are no pretrained models.

In general, our main contributions are:

- An analysis of the helpfulness of an additional translated corpus in different settings;
- An exploratory study about the effects of diverse granularity levels in input representation for low-resource AMR-to-text generation; and,
- A deep analysis of three commonly used preprocessing strategies in AMR-to-text generation: delexicalisation, compression, and linearisation.

We start by briefly reviewing AMR fundamentals (Section 2) and presenting the main related work (Section 3). Section 4 reports the techniques and methods that we investigate, while the achieved results are discussed in Section 5. Section 6 concludes this paper.

2 *Abstract Meaning Representation*

As previously mentioned, AMR aims to encode the meaning of a sentence in a directed, labeled, acyclic, and rooted graph (Banarescu et al., 2013). Furthermore, this representation may comprehend semantic information related to semantic roles, named entities, spatial-temporal information and coreferences, among others.

Figure 1 presents an example of an AMR graph for the sentence "The boy destroyed the room". It is worth noting that, as AMR abstracts away the syntactic information, multiple possible sentences can correspond to this graph. This way, another possible sentence that represents the graph could be "the destruction of the room by the boy".

The current AMR-annotated corpus for English contains 59,255 instances[1]. For Non-English languages, there are some efforts to

[1] `https://catalog.ldc.upenn.edu/LDC2020T02`

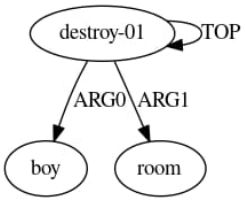


Figure 1: AMR example for the sentence "The boy destroyed the room.".

build corpora leveraging the alignments and existing parallel corpora by using AMR as an interlingua (Xue et al., 2014; Anchiêta and Pardo, 2018). Additionally, other works adapt the AMR guidelines to their languages (Sobrevilla Cabezudo and Pardo, 2019). However, most corpora are far from presenting a size similar to the English one.

For Brazilian Portuguese, as far as we know, there are two AMR corpora, one focused on annotating the sentences of "The Little Prince" book (Anchiêta and Pardo, 2018), and another one that contains manually annotated news text sentences (Sobrevilla Cabezudo and Pardo, 2019). Similarly to Banarescu et al. (2013), some concepts of both corpora were annotated using Verbo-Brasil (Duran and Aluísio, 2015), a lexical resource analogous to PropBank (Palmer, Gildea, and Kingsbury, 2005). Concerning the size of these corpora, the "Little Prince" corpus contains 1,527 annotated sentences (instances), and the second corpus comprises 299 instances, being both small and making it hard to replicate results obtained by state-of-the-art methods.

3 Related Work

In the last years, several AMR-to-Text generation methods for English have been proposed. Initially, methods inspired on Statistical Machine Translation (SMT) techniques (Pourdamghani, Knight, and Hermjakob, 2016) and tree-to-string or graph-to-string transducers (Flanigan et al., 2016) were proposed. Recently, neural models as sequence-to-sequence (Neural Machine Translation or NMT) (Castro Ferreira et al., 2017; Konstas et al., 2017) and, mainly, graph-to-sequence (Beck, Haffari, and Cohn, 2018) and pretrained-based ones (Mager et al., 2020), have emerged, outperforming the previous approaches.

To the extent of our knowledge, the only work focused on AMR-to-Text generation for a Non-English language is proposed by Sobrevilla Cabezudo, Mille, and Pardo (2019). The authors explore the automatic construction of an AMR corpus for Brazilian Portuguese (BP) from its English version and evaluate SMT and NMT approaches on a BP test set composed of 299 instances. Other non-English work (Fan and Gardent, 2020) have tried to generate sentences in diverse languages from English AMR graphs. Although the results are promising, this work does not deal with some specific linguistic phenomena as the previous one does.

In what follows, we detail the dataset that we use in this work and the methods that we investigate.

4 AMR-to-Text Generation

4.1 Data

The methods that we investigate are trained on two corpora and their combinations. The first one is an updated version of the AMR corpus for Brazilian Portuguese (Sobrevilla Cabezudo and Pardo, 2019), which represents our target (*gold*) dataset. This version is a manually annotated corpus comprising 870 instances divided into 402, 224, and 244 instances for training, development, and test, respectively.

The second one is a portion of an automatically generated AMR corpus for Portuguese and represents our augmented (*translated*) dataset. This corpus is generated by translating both AMR graphs and sentences from the English AMR corpus[2] to Portuguese and inheriting the alignments between node/edges and surface tokens[3] (Sobrevilla Cabezudo, Mille, and Pardo, 2019).

In general, this corpus comprises 18,219 and 1,027 instances in the training and development set, respectively, that correspond to the higher-quality translations according to BLEU (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007) scores.[4] It is worth noting that, differently from the work of Sobrevilla Cabezudo, Mille, and Pardo (2019), that translates only aligned concepts

[2]In this work, we use the LDC2016E25 corpus to perform the experiments.

[3]Surface tokens are those included in the reference sentence.

[4]The actual portion of the dataset contains 20,000 and 1,271 instances for training and development, respectively. However, some instances were filtered out because they presented some format errors.

in the AMR graphs, all concepts in the AMR graphs are translated.

4.2 Machine Translation-based Techniques

AMR-to-text generation receives an AMR graph as an input and generates a text in natural language; however, Machine Translation models are trained on linear input/output pairs. This way, we need to generate a flattened version of the AMR graph as input. Some flattened versions that have been used in the literature are the ones generated by the PENMAN notation (Matthiessen and Bateman, 1991) and the depth-first search (DFS) algorithm. However, other preprocessing steps can generate a flattened AMR version. Figure 2 shows an example of a flattened AMR version for the sentence *A crise na Venezuela foi um assunto que permeou as reuniões.* ("The crisis in Venezuela was an issue that permeated the meetings.").

In order to evaluate how the use of various flattened AMR versions affect the performance in AMR-to-text generation, we explore the strategies that include the preprocessing steps used by Castro Ferreira et al. (2017). In particular, the preprocessing steps are:

- Delexicalisation: that anonymises some entities of the graph;
- Compression: that determines which nodes and relations should be in the flattened graph; and,
- Linearisation: that determines how the nodes and relations should be put into the flattened graph.

We study two machine translation approaches, a statistical phrase-based one (Koehn, Och, and Marcu, 2003) as a strong baseline and one based on neural models (Bahdanau, Cho, and Bengio, 2015) in a similar way to Castro Ferreira et al. (2017).

4.2.1 Statistical Machine Translation (SMT)

The training parameters in SMT are the same of Castro Ferreira et al. (2017) and a 5-gram language model trained on the Brazilian Portuguese corpus provided by Hartmann et al. (2017) by using KenLM (Heafield, 2011). Furthermore, we use Moses (Koehn et al., 2007) to train the statistical machine translation models.

4.2.2 Neural Machine Translation (NMT)

The architecture and the parameters used in NMT are described as follows: the encoder and the decoder are a 1-layer RNN, and a 2-layers RNN with LSTM, each with a 512D hidden unit, respectively. Besides, the RNN decoder also uses bilinear attention (Luong, Pham, and Manning, 2015). Furthermore, the vocabulary is shared, and we apply weight tying between the source, target, and output layers. Additionally, source and target word embeddings are 512D each, and both are trained jointly with the model.

Among other parameters, the maximum sequence length in the decoder is 80, and we apply dropout with a probability of 0.25 in source embeddings. Moreover, models are trained using the Adam optimizer with a learning rate of 0.0003, a learning rate reduce factor of 0.5, and the learning rate decays if perplexity does not improve after 3 checkpoints/epochs. Besides, we use mini-batches of size 16. Finally, we apply early stopping for model selection based on perplexity scores. Training is halted if a model does not improve on the development set for more than 8 checkpoints/epochs. Sockeye[5] (Hieber et al., 2017) provides all other parameters.

4.3 Graph-to-Sequence (G2S)

Unlike previous approaches, which depend on preprocessing steps and can lose information, the Graph-to-Sequence approach tries to capture the whole graph information more effectively. This work also follows the Graph-to-Sequence approach proposed by Beck, Haffari, and Cohn (2018), that models AMR graphs using a Gated Graph Neural Network (GGNN) (Li et al., 2015b).

In general, model input is defined by the nodes (concepts and relations) and positional embeddings of a graph. To consider AMR relations as nodes, the authors transform the original AMR graph into its respective Levi graph[6] (Levi, 1942). Finally, the output is a version of the original sentence.

We use the same architecture and parameters as Beck, Haffari, and Cohn (2018). Thus, the number of layers in the GGNN encoder

[5] `https://github.com/beckdaniel/sockeye/`

[6] A Levi graph is a modification of a labeled graph so that relations are converted into nodes generating an unlabeled graph.

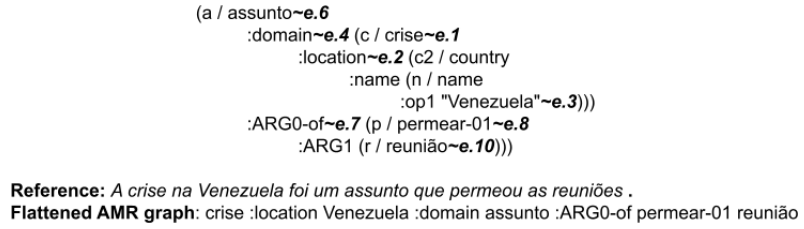


Figure 2: Sentence *A crise na Venezuela foi um assunto que permeou as reuniões.* ("The crisis in Venezuela was an issue that permeated the meetings."), its corresponding AMR graph and a flattened version that includes only aligned nodes/edges. Alignments in AMR graph are in bold.

is 8. All dimensionalities are fixed at 512D except for the GGNN encoder, which uses 576D. The decoder uses a 2-layer LSTM and the Bilinear attention proposed by (Luong, Pham, and Manning, 2015). The remained parameters are the same as the NMT approach.

4.4 Preprocessing Strategies

The preprocessing strategies that we test in this work include:

- Delexicalisation: we delexicalise constants like named-entities or numbers, replacing the original information for tags such as __name1__ and __quant1__ for NMT (Castro Ferreira et al., 2017) and person_1 and quantity_1 for G2S (Beck, Haffari, and Cohn, 2018). A list of tag-values is kept, aiming to rebuild the output sentence after generation;
- Compression: it is performed using a Conditional Random Field (CRF) and executed sequentially over a flattened representation obtained by depth-first search through the AMR graph, and its name and the parent name represent each element. We use the CRFSuite toolkit[7] (Okazaki, 2007) to train our model;
- Linearisation: we apply two strategies. The first consists of performing a depth-first search through the AMR graph, printing the elements (nodes and edges) according to the visiting order. The other strategy is based on the 2-step maximum entropy classifier developed by Lerner and Petrov (2013) and adapted by Castro Ferreira et al. (2017) (we called it preordering). Given an AMR graph represented by a tree, this consists of ordering a head and its corresponding subtrees, i.e., defining which subtrees should be at left/right of the head, and then ordering the subtrees in each built group (left and right side of the head).

All models are tested on inputs/outputs that include or not the preprocessing steps. However, we only explore compression and linearization (preordering) for SMT and delexicalisation for G2S. In addition, when compression is not considered, we include all elements from an AMR graph (nodes and edges).

4.5 Representation Levels

We explore three different representation levels for both input (AMR graph) and output (sentence): words, subwords, and characters. It is expected that finer-grained representations, such as subwords and characters, produce better results, handling in a better way rare words or even possible mismatches between the *translated* and the *gold* corpora.

Subwords are generated by using the Bertimbau's vocabulary provided by Souza, Nogueira, and Lotufo (2020)[8] that uses the sentencepiece tool[9] and the BPE algorithm (Sennrich, Haddow, and Birch, 2016). In the case of the flattened AMR graph, we do not decompose the relations. This way, relations such as "*:ARG0*" or "*:mod*" are kept intact, differently from concepts, such as "*ferida*", that are changed to "*fer ##ida*" in the case of subwords and "*f e r i d a*" in the case of characters.

It is worth mentioning that, in the case of G2S, each subword/character is represented

[7] https://www.chokkan.org/software/crfsuite/

[8] https://github.com/neuralmind-ai/portuguese-bert

[9] https://github.com/google/sentencepiece

by a node, and all subwords/characters that compose a concept are linked sequentially in two directions. For example, we create an edge from subword "fer" to "##ida" and vice-versa.

We present and analyze the achieved results in what follows.

5 Results and Analysis

Tables 1, 2, 3, and 4 show the overall results for SMT, NMT and G2S approaches in terms of BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), and chrF++ (Popović, 2017) evaluation metrics [10]. The tables contain the results when the *translated* corpus (T), the *gold* corpus (G), a join of the training *translated* and *gold* corpora (T + G), and a join of the training/development *translated* and *gold* corpora (T + G Train/dev) are used. In addition, the results of using some preprocessing steps and representation levels are shown. Preprocessing steps are identified as +D (delexicalisation), +C (compression), and +P (preordering) and the opposite when these are not included in the preprocessing.

In general, the best result[11] for SMT happens when we train the model on T + G and use compression and preordering. Likewise, the best result for NMT occurs when the training is performed on T + G, using delexicalisation and preordering, and char-level representations. At last, G2S performs better when the model is trained on T + G train/dev, and lexicalisation and bpe-level presentation are applied.

Results on *gold* corpus show that SMT is by far the best approach to be used in the case of low-resource settings. It is expected as neural models usually need lots of data to achieve good performance, and SMT uses a pre-built language model that guides the decoding, differently from NMT and G2S in which the language model is built during training. In particular, using compressing (+C) and preordering (+P) produces the best results, being preordering the most critical preprocessing step, similarly to the results obtained by Castro Ferreira et al. (2017).

Concerning neural models, NMT produces the best performance; however, this is far from the SMT one yet. Char-level representation and Delexicalisation (+D) are the best strategies when BLEU is evaluated. However, lexicalisation (-D) is better when the metric is chrF++. Moreover, preordering (+P) seems useful when char-level representation is used. Finally, G2S presents the worst performance, being char-level representation and delexicalisation (+D) the best strategies.

In the following subsections, we will study how the performance changes in different contexts and try to answer three questions: (1) how helpful is the *translated* corpus? (2) what are the most useful preprocessing steps? (3) how fine-grained should be the representations to achieve better performance?

5.1 How helpful is the *translated* corpus?

To determine the helpfulness of the *translated* corpus, we study the performance when models are trained on T and T + G.

In general, the *translated* corpus is helpful as all models trained on it present better results than models trained on only *gold* corpus, however, there exists a mismatch between *translated* and *gold* corpora, as values for all measures in development set are quite higher than the obtained in test set (see results on *translated* corpus - T). This behavior can be generated by domain mismatch, in which the vocabulary is different even though both corpora are on news, or by structure mismatch between AMR graphs, since *translated* AMR graphs are English-biased and can introduce noise during training (as its size is bigger than the *gold* corpus).

Regarding the change in the performance when *gold* corpus is added to the *translated* one (T + G), SMT gets leveraging the data increase better. On the other hand, NMT performance presents a slight improvement when *gold* corpus is added. Finally, the G2S performance slightly drops in all cases and can suggest that there is a structural mismatch between the *translated* and gold AMR graphs, as this approach considers structural information, different from SMT or NMT, which use a flattened version with some nodes/edges included in it.

In order to evaluate how to deal with the possible mismatch, we add the *translated* development set (1,027 instances) to the *gold* one as well. Table 4 shows the result for each

[10] We execute 4 runs for each experiment and show the mean and standard deviation for NMT and G2S.

[11] Best results are highlighted in bold in Tables.

		DEV			TEST		
		BLEU	METEOR	chrF++	BLEU	METEOR	chrF++
Gold	+C+P	11.58	0.31	0.48	10.00	0.30	0.48
	+C-P	11.36	0.29	0.47	7.95	0.26	0.46
	-C+P	6.06	0.24	0.43	6.05	0.24	0.43
	-C-P	7.31	0.24	0.44	4.89	0.22	0.43
Translated	+C+P	27.18	0.45	0.57	9.98	0.29	0.47
	+C-P	26.10	0.44	0.56	10.50	0.28	0.46
	-C+P	23.73	0.42	0.55	10.47	0.30	0.48
	-C-P	24.02	0.42	0.55	7.83	0.26	0.46
Translated + Gold	+C+P	**18.67**	**0.38**	**0.52**	**14.83**	**0.33**	**0.49**
	+C-P	17.75	0.37	0.51	11.96	0.32	0.47
	-C+P	17.38	0.37	0.51	13.91	0.32	0.49
	-C-P	14.86	0.35	0.50	11.96	0.32	0.48

Table 1: Overall SMT results.

setting and approach. Unlike the previous setting (T+G), both SMT and NMT present a small improvement in all metrics. However, G2S presents bigger improvements, suggesting that adding *translated* instances can make models more robust to possible structural divergences, leading to performance improvements.

5.2 What are the most useful preprocessing strategies?

5.2.1 Statistical Machine Translation

Pre-ordering (+P) seems to lead to improvements, however, this improvement is notorious when translated + *gold* corpora are used in the training set. Another point to highlight is the importance of compression (+C). Initial experiments (T and T + G) show that compression leads to slight improvements. However, no compression (-C) produces the best results when the classifier is trained on T + G train/dev.

5.2.2 Neural Machine Translation

Delexicalisation (+D) seems to be a good strategy for word and char-level representations, but it is not relevant for bpe-level. Moreover, compression (+C) generally harms the performance or produces mixed results, being better when lexicalisation (-D) is applied in char-level representation. Finally, pre-ordering (+P) seems to produce small improvements in all settings.

5.2.3 Graph-to-Sequence

About Graph-to-Sequence approach, Delexicalisation (+D) improves the performance when word and char-level presentations are used. However, the contrary happens when bpe-level representation is used. A possible explanation is that delexicalisation reduces data sparseness when word-level representation is applied together and allows to deal with large graphs in the case of char-level representation. However, in the case of bpe, delexicalisation seems to introduce noise and makes the model more prone to generate hallucinations.

5.3 How fine-grained should be the representations to achieve better performance?

Concerning the representation levels, characters and bpe produce the best and second-best performance for NMT. The main gain in both representations is in terms of METEOR and chRF++, which is expected as these representations are finer-grained and the evaluation measures take stems and characters into account.

Different from NMT, bpe produces the best performance for G2S. However, and as it was previously mentioned, this performance happens when delexicalisation is applied. This way, we hypothesise two possible problems: (1) word-level representations suffer more from mismatch problems as experiments on T and T + G show low performance, and (2) char-level representations can generate larger AMR graphs for which semantics can be challenging to be captured by G2S.

Another point to highlight is that finer-grained representations usually help reducing the bias to the development set, mainly when char-level representations are used. Consequently, mismatch problems are mitigated. This can be seen in the difference between development and test performance for experiments on T and T + G train/dev. For example, Figure 3 shows the difference mentioned for NMT. Experiments on T + G present a BLEU overall difference of 10.45, 9.9, and 5,67 between development and test for word, bpe, and char-level representations. Similarly, differences for METEOR and chrF++ are 0.11, 0.11, and 0.03, and 0.11, 0.09, and 0.00, respectively.

			DEV			TEST		
			BLEU	METEOR	chrF++	BLEU	METEOR	chrF++
G	word	+D+C+P	0.00±0.00	0.06±0.00	0.05±0.00	2.66±0.14	0.10±0.00	0.13±0.01
		+D+C-P	0.87±0.87	0.10±0.01	0.12±0.00	2.48±0.37	0.11±0.00	0.14±0.01
		+D-C+P	0.00±0.00	0.10±0.01	0.11±0.00	2.61±0.23	0.11±0.00	0.13±0.00
		+D-C-P	0.37±0.63	0.10±0.01	0.11±0.01	2.39±0.18	0.10±0.00	0.13±0.01
		-D+C+P	0.00±0.00	0.03±0.02	0.02±0.02	0.00±0.00	0.03±0.02	0.02±0.02
		-D+C-P	0.00±0.00	0.03±0.02	0.02±0.02	0.00±0.00	0.03±0.02	0.02±0.02
		-D-C+P	0.00±0.00	0.02±0.00	0.01±0.00	0.00±0.00	0.02±0.00	0.01±0.00
		-D-C-P	0.00±0.00	0.02±0.00	0.01±0.00	0.00±0.00	0.02±0.00	0.01±0.00
	bpe	+D+C+P	0.00±0.00	0.02±0.00	0.01±0.00	0.00±0.00	0.01±0.00	0.01±0.00
		+D+C-P	0.34±0.58	0.05±0.04	0.07±0.05	0.88±0.90	0.06±0.04	0.08±0.06
		+D-C+P	0.33±0.56	0.07±0.03	0.09±0.04	1.33±0.81	0.08±0.04	0.10±0.05
		+D-C-P	0.33±0.57	0.03±0.03	0.04±0.05	0.39±0.67	0.03±0.03	0.04±0.05
		-D+C+P	0.00±0.00	0.03±0.02	0.02±0.02	0.00±0.00	0.03±0.02	0.02±0.02
		-D+C-P	0.00±0.00	0.03±0.02	0.02±0.02	0.00±0.00	0.03±0.02	0.02±0.02
		-D-C+P	0.00±0.00	0.02±0.00	0.01±0.00	0.00±0.00	0.02±0.00	0.01±0.00
		-D-C-P	0.00±0.00	0.02±0.00	0.01±0.00	0.00±0.00	0.02±0.00	0.01±0.00
	char	+D+C+P	0.59±0.67	0.11±0.03	0.22±0.06	3.12±0.37	0.15±0.02	0.26±0.05
		+D+C-P	1.61±1.00	0.11±0.01	0.19±0.01	2.80±0.27	0.11±0.00	0.19±0.00
		+D-C+P	2.28±0.36	0.12±0.01	0.22±0.04	3.12±0.10	0.13±0.01	0.22±0.03
		+D-C-P	1.63±0.09	0.10±0.00	0.18±0.00	2.88±0.35	0.11±0.00	0.19±0.00
		-D+C+P	1.35±0.82	0.14±0.05	0.27±0.09	1.77±1.14	0.14±0.05	0.28±0.09
		-D+C-P	0.00±0.00	0.11±0.00	0.26±0.00	0.48±0.82	0.13±0.01	0.27±0.01
		-D-C+P	1.45±0.87	0.16±0.01	0.31±0.01	0.70±1.22	0.16±0.01	0.31±0.01
		-D-C-P	0.72±0.74	0.09±0.04	0.20±0.07	0.00±0.00	0.09±0.04	0.19±0.07
—	word	+D+C+P	11.02±1.37	0.26±0.02	0.32±0.01	4.16±0.65	0.20±0.01	0.29±0.01
		+D+C-P	4.66±0.19	0.18±0.01	0.24±0.01	2.46±0.29	0.13±0.00	0.19±0.00
		+D-C+P	20.53±0.56	0.38±0.00	0.46±0.00	5.88±0.23	0.24±0.01	0.33±0.01
		+D-C-P	19.35±0.92	0.37±0.01	0.44±0.00	5.88±0.30	0.23±0.00	0.32±0.01
		-D+C+P	17.96±0.76	0.36±0.01	0.42±0.01	3.79±0.34	0.18±0.01	0.25±0.01
		-D+C-P	2.32±0.37	0.12±0.01	0.16±0.01	0.12±0.21	0.06±0.00	0.09±0.01
		-D-C+P	19.22±0.75	0.38±0.01	0.43±0.02	3.96±0.62	0.18±0.01	0.26±0.02
		-D-C-P	19.81±0.77	0.37±0.01	0.42±0.01	3.17±0.33	0.17±0.01	0.24±0.01
T	bpe	+D+C+P	8.96±2.07	0.26±0.02	0.36±0.01	3.90±1.03	0.21±0.02	0.32±0.01
		+D+C-P	12.89±3.52	0.33±0.02	0.44±0.02	3.57±1.10	0.21±0.02	0.33±0.01
		+D-C+P	15.41±2.46	0.36±0.02	0.46±0.01	5.39±0.68	0.24±0.01	0.36±0.00
		+D-C-P	20.04±0.60	0.38±0.00	0.48±0.01	7.05±1.00	0.27±0.02	0.38±0.01
		-D+C+P	19.34±4.59	0.41±0.03	0.49±0.02	6.10±1.42	0.24±0.03	0.36±0.02
		-D+C-P	13.60±2.37	0.36±0.02	0.46±0.01	2.86±0.74	0.19±0.01	0.32±0.01
		-D-C+P	22.39±1.57	0.44±0.01	0.51±0.00	7.08±0.71	0.27±0.02	0.37±0.02
		-D-C-P	20.87±1.16	0.42±0.01	0.50±0.01	5.47±0.63	0.24±0.01	0.35±0.00
	char	+D+C+P	13.39±0.37	0.27±0.00	0.37±0.00	8.69±1.33	0.29±0.01	0.43±0.01
		+D+C-P	15.45±0.50	0.31±0.00	0.43±0.01	8.02±0.40	0.28±0.01	0.42±0.01
		+D-C+P	13.73±0.40	0.31±0.00	0.43±0.01	8.21±0.95	0.28±0.01	0.42±0.01
		+D-C-P	13.06±1.22	0.29±0.01	0.42±0.01	7.18±0.88	0.27±0.00	0.42±0.00
		-D+C+P	16.06±2.91	0.33±0.04	0.43±0.03	7.63±2.23	0.28±0.03	0.42±0.03
		-D+C-P	17.75±0.41	0.34±0.01	0.44±0.01	6.16±1.13	0.26±0.01	0.41±0.00
		-D-C+P	15.73±1.19	0.33±0.02	0.43±0.02	6.97±1.40	0.26±0.02	0.41±0.02
		-D-C-P	11.26±4.63	0.24±0.09	0.34±0.10	4.04±3.64	0.17±0.09	0.29±0.12
T+G	word	+D+C+P	2.77±0.57	0.16±0.01	0.22±0.02	4.76±0.38	0.20±0.01	0.28±0.02
		+D+C-P	3.65±0.54	0.19±0.02	0.27±0.02	4.23±1.00	0.19±0.02	0.27±0.03
		+D-C+P	5.15±0.82	0.23±0.01	0.31±0.01	6.04±0.30	0.22±0.01	0.30±0.01
		+D-C-P	4.42±0.52	0.20±0.01	0.28±0.01	4.81±0.64	0.20±0.01	0.27±0.02
		-D+C+P	2.93±0.73	0.17±0.01	0.24±0.00	3.59±0.38	0.18±0.00	0.24±0.00
		-D+C-P	2.70±0.48	0.14±0.01	0.20±0.01	2.58±0.67	0.14±0.02	0.20±0.01
		-D-C+P	3.51±0.77	0.16±0.02	0.23±0.02	2.57±0.27	0.16±0.02	0.22±0.02
		-D-C-P	3.63±0.89	0.17±0.01	0.24±0.02	2.99±0.80	0.16±0.01	0.23±0.01
	bpe	+D+C+P	2.72±0.73	0.19±0.01	0.30±0.01	4.71±0.38	0.23±0.01	0.34±0.01
		+D+C-P	3.38±1.35	0.20±0.04	0.32±0.03	3.21±1.43	0.19±0.04	0.31±0.03
		+D-C+P	7.10±1.10	0.28±0.02	0.39±0.02	7.52±1.10	0.28±0.02	0.37±0.01
		+D-C-P	5.68±1.21	0.26±0.02	0.37±0.02	5.78±1.38	0.25±0.02	0.35±0.01
		-D+C+P	3.56±0.52	0.21±0.01	0.34±0.01	4.47±1.21	0.22±0.02	0.35±0.01
		-D+C-P	4.45±1.02	0.22±0.02	0.33±0.01	4.60±1.36	0.22±0.02	0.34±0.02
		-D-C+P	7.10±0.40	0.27±0.00	0.37±0.01	7.42±0.70	0.26±0.01	0.36±0.01
		-D-C-P	6.69±0.77	0.26±0.01	0.36±0.01	5.93±1.35	0.25±0.01	0.36±0.01
	char	+D+C+P	7.82±0.44	0.26±0.01	0.38±0.01	9.38±0.22	0.30±0.01	0.44±0.01
		+D+C-P	8.36±0.51	0.29±0.01	0.42±0.01	8.65±0.90	0.28±0.01	0.42±0.01
		+D-C+P	**7.28±0.49**	**0.29±0.01**	**0.42±0.01**	**10.03±0.37**	**0.31±0.01**	**0.44±0.01**
		+D-C-P	7.04±0.14	0.27±0.00	0.42±0.00	7.34±0.88	0.27±0.01	0.41±0.01
		-D+C+P	7.48±0.74	0.29±0.01	0.43±0.01	8.85±0.78	0.29±0.01	0.43±0.01
		-D+C-P	7.99±1.57	0.27±0.01	0.41±0.01	7.96±0.69	0.27±0.01	0.42±0.01
		-D-C+P	5.98±0.59	0.27±0.02	0.41±0.02	8.25±0.94	0.29±0.02	0.43±0.02
		-D-C-P	5.33±1.89	0.23±0.05	0.37±0.05	5.20±3.06	0.24±0.05	0.38±0.05

Table 2: Overall NMT results.

5.4 Manual Revision

We present now some analysis of actual generated cases. Figure 4 shows the AMR graph, the reference, and the output generated by the three approaches for the sentences "He/She does not want it" ("*não quer*") and "He/She attended excellent schools, and majored in economics at Yale." ("*frequentou excelentes escolas, e se formou em economia por Yale.*"). We can see some mistakes for each approach associated with hidden subjects (highlighted in red), wrong conjugation (blue), fluency/concordance (green), repetitions (purple), random words (yellow), and entity copying (pink).

The first example is simple, and the three approaches present similar outputs. SMT produces almost the same reference; however,

			DEV			TEST		
			BLEU	METEOR	chrF++	BLEU	METEOR	chrF++
G	word	+D	0.00 ±0.00	0.03 ±0.01	0.02 ±0.01	0.00 ±0.00	0.03 ±0.01	0.02 ±0.01
		-D	0.00 ±0.00	0.03 ±0.01	0.02 ±0.01	0.00 ±0.00	0.03 ±0.01	0.02 ±0.01
	bpe	+D	0.00 ±0.00	0.03 ±0.01	0.03 ±0.02	0.00 ±0.00	0.03 ±0.02	0.03 ±0.02
		-D	0.00 ±0.00	0.02 ±0.01	0.01 ±0.00	0.00 ±0.00	0.02 ±0.01	0.01 ±0.00
	char	+D	0.00 ±0.00	0.09 ±0.01	0.13 ±0.01	1.59 ±0.47	0.09 ±0.01	0.14 ±0.01
		-D	0.00 ±0.00	0.05 ±0.00	0.09 ±0.00	0.00 ±0.00	0.05 ±0.00	0.09 ±0.00
T	word	+D	14.88 ±4.17	0.32 ±0.06	0.38 ±0.06	4.66 ±1.50	0.18 ±0.04	0.26 ±0.05
		-D	10.41 ±4.20	0.24 ±0.06	0.30 ±0.07	1.95 ±1.74	0.13 ±0.04	0.19 ±0.05
	bpe	+D	8.44 ±1.60	0.23 ±0.02	0.29 ±0.01	2.60 ±0.37	0.14 ±0.00	0.22 ±0.01
		-D	21.04 ±1.09	0.42 ±0.01	0.48 ±0.00	6.75 ±0.51	0.26 ±0.01	0.36 ±0.01
	char	+D	11.46 ±1.67	0.25 ±0.02	0.32 ±0.03	6.07 ±2.02	0.23 ±0.04	0.35 ±0.05
		-D	7.09 ±2.24	0.18 ±0.03	0.24 ±0.02	1.43 ±0.78	0.12 ±0.03	0.23 ±0.03
T+G	word	+D	3.52 ±2.14	0.17 ±0.05	0.23 ±0.05	3.80 ±2.01	0.16 ±0.04	0.23 ±0.05
		-D	1.00 ±1.74	0.10 ±0.04	0.15 ±0.05	1.00 ±1.72	0.09 ±0.04	0.15 ±0.06
	bpe	+D	1.37 ±0.35	0.12 ±0.01	0.18 ±0.00	1.82 ±0.32	0.12 ±0.01	0.19 ±0.01
		-D	5.62 ±0.43	0.26 ±0.01	0.36 ±0.01	6.44 ±0.79	0.26 ±0.01	0.36 ±0.01
	char	+D	5.21 ±1.25	0.22 ±0.03	0.33 ±0.05	6.09 ±1.50	0.22 ±0.04	0.34 ±0.05
		-D	2.53 ±1.63	0.17 ±0.04	0.28 ±0.05	2.63 ±1.94	0.17 ±0.04	0.29 ±0.05

Table 3: Overall G2S results.

			DEV			TEST		
			BLEU	METEOR	chrF++	BLEU	METEOR	chrF++
SMT	word	+C+P	25.66	0.43	0.56	12.92	0.31	0.48
		+C-P	24.72	0.42	0.55	12.52	0.31	0.48
		-C+P	22.09	0.41	0.54	14.69	0.34	0.50
		-C-P	22.29	0.41	0.54	10.03	0.30	0.48
NMT	word	+D+C+P	11.21 ±1.36	0.25 ±0.01	0.32 ±0.02	5.38 ±1.03	0.22 ±0.02	0.30 ±0.02
		+D+C-P	14.25 ±0.92	0.31 ±0.01	0.39 ±0.01	4.95 ±0.52	0.21 ±0.01	0.29 ±0.01
		+D-C+P	16.82 ±0.81	0.34 ±0.01	0.42 ±0.00	6.70 ±0.79	0.24 ±0.01	0.32 ±0.01
		+D-C-P	17.10 ±0.47	0.34 ±0.00	0.42 ±0.00	6.68 ±0.20	0.23 ±0.00	0.32 ±0.01
		-D+C+P	14.88 ±1.42	0.32 ±0.01	0.38 ±0.01	3.94 ±0.64	0.19 ±0.01	0.26 ±0.01
		-D+C-P	14.98 ±1.48	0.31 ±0.01	0.37 ±0.01	3.25 ±0.51	0.17 ±0.01	0.24 ±0.01
		-D-C+P	17.64 ±0.74	0.35 ±0.01	0.41 ±0.01	4.76 ±0.44	0.20 ±0.01	0.28 ±0.01
		-D-C-P	16.87 ±0.47	0.33 ±0.01	0.39 ±0.01	4.48 ±0.31	0.19 ±0.00	0.26 ±0.01
	bpe	+D+C+P	11.81 ±0.43	0.28 ±0.01	0.37 ±0.01	6.65 ±1.10	0.25 ±0.01	0.35 ±0.01
		+D+C-P	14.32 ±0.87	0.33 ±0.01	0.43 ±0.01	5.09 ±0.54	0.24 ±0.01	0.35 ±0.01
		+D-C+P	16.98 ±3.23	0.37 ±0.02	0.47 ±0.02	7.70 ±1.53	0.27 ±0.01	0.38 ±0.01
		+D-C-P	16.32 ±2.56	0.36 ±0.02	0.45 ±0.01	6.15 ±0.87	0.26 ±0.01	0.36 ±0.01
		-D+C+P	13.80 ±3.03	0.35 ±0.03	0.46 ±0.01	5.61 ±0.82	0.24 ±0.02	0.36 ±0.02
		-D+C-P	14.53 ±3.18	0.35 ±0.02	0.45 ±0.01	4.79 ±1.55	0.22 ±0.02	0.34 ±0.02
		-D-C+P	21.38 ±0.93	0.41 ±0.01	0.48 ±0.01	7.80 ±0.77	0.27 ±0.01	0.37 ±0.01
		-D-C-P	20.25 ±1.06	0.40 ±0.01	0.49 ±0.01	6.38 ±1.16	0.26 ±0.01	0.38 ±0.01
	char	+D+C+P	12.61 ±0.50	0.27 ±0.00	0.37 ±0.00	9.42 ±0.47	0.30 ±0.00	0.44 ±0.00
		+D+C-P	14.59 ±0.43	0.31 ±0.00	0.43 ±0.01	9.07 ±0.80	0.29 ±0.02	0.43 ±0.01
		+D-C+P	13.20 ±0.16	0.31 ±0.00	0.43 ±0.01	9.83 ±0.88	0.30 ±0.01	0.44 ±0.01
		+D-C-P	12.91 ±0.53	0.30 ±0.01	0.42 ±0.01	8.49 ±0.88	0.29 ±0.01	0.42 ±0.01
		-D+C+P	17.18 ±0.54	0.35 ±0.00	0.45 ±0.00	10.14 ±0.38	0.30 ±0.01	0.44 ±0.01
		-D+C-P	16.65 ±0.72	0.33 ±0.01	0.44 ±0.01	8.10 ±0.88	0.28 ±0.01	0.42 ±0.01
		-D-C+P	12.19 ±4.38	0.27 ±0.08	0.37 ±0.09	5.93 ±3.48	0.24 ±0.10	0.36 ±0.12
		-D-C-P	14.58 ±0.58	0.31 ±0.01	0.43 ±0.00	7.61 ±0.82	0.27 ±0.01	0.42 ±0.00
G2S	word	+D	16.84 ±1.88	0.36 ±0.02	0.43 ±0.02	7.70 ±1.74	0.26 ±0.03	0.34 ±0.03
		-D	9.73 ±5.58	0.23 ±0.09	0.29 ±0.09	2.73 ±2.17	0.14 ±0.05	0.20 ±0.06
	bpe	+D	7.59 ±1.97	0.22 ±0.02	0.28 ±0.02	3.28 ±0.74	0.15 ±0.01	0.23 ±0.01
		-D	**20.85 ±1.21**	**0.41 ±0.02**	**0.48 ±0.02**	**8.69 ±0.59**	**0.29 ±0.02**	**0.38 ±0.02**
	char	+D	11.10 ±1.95	0.25 ±0.03	0.32 ±0.02	7.03 ±2.46	0.24 ±0.04	0.35 ±0.05
		-D	7.94 ±1.22	0.22 ±0.02	0.30 ±0.02	4.00 ±0.49	0.19 ±0.01	0.32 ±0.02

Table 4: Results of adding *translated* development set to the *gold* one. It is called T + G train/dev.

this includes the pronoun "*ele*" ("he/she") that is treated as a hidden subject in the reference. Conversely, NMT and G2S omit the pronoun, making the generated sentence more natural; nevertheless, both approaches generate the verb "*querer*" ("want") in a different conjugation (1st person). A possible explanation is that NMT and G2S are trained on char and bpe-level representations, this way, they can generate different conjugations easily. In addition, NMT generates the word "*dizer*" ("to say") that is not part of the AMR graph.

The second one is a harder example with more relations and concepts such as named entities ("university"), co-references ("e1 / *ele*" or "he/she") and connectors ("*e*"). In this case, none of the approaches can omit the pronoun "*ele*" as the reference does. Another common problem in all approaches is the lack of agreement/fluency. For example, the expression "na yale" should be replaced by "*em yale*" in order to be more fluent.

Analyzing other issues, SMT tries to generate sentences with all possible concepts included in the graph, even if the generated text is not fluent. On the other hand, neural models suffer from classical problems such as repetition and random word generation (the hallucination problem).

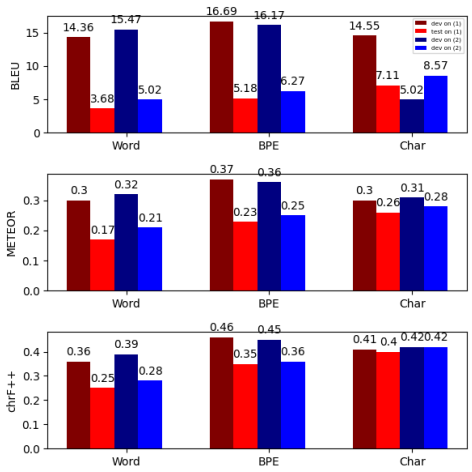


Figure 3: Difference between development and test performance for experiments on (1) T and (2) T + G train/dev.

```
(q / querer-01
        :polarity -
        :ARG0 (e / ele))
```

Reference: *não quer*
SMT: ***ele*** *não quer*
NMT: *não* ***quero dizer*** *.*
G2S: *não* ***quero*** *.*

```
(a / e
        :op1 (f / frequentar-01
                :ARG0 (e1 / ele)
                :ARG1 (e2 / escola
                        :mod (e3 / excelente)))
        :op2 (f1 / formar-102
                :ARG1 e1
                :ARG2 (e4 / economia)
                :location (u / university
                        :name (n / name
                                :op1 "Yale"))))
```

Reference: *frequentou excelentes escolas , e se formou em economia por yale .*
SMT: ***ele*** *participou de uma excelente escola* ***formação*** *economia* ***na*** *yale*
NMT: ***ele*** *estava* ***formando*** ***excelente*** *e* ***formando*** *a economia* ***na*** *yale .* ***__name1__ :***
G2S: ***ele*** *estava* ***analisando*** *em excelente escola e foi* ***formada na*** *economia* ***na economia*** *.*

Figure 4: Outputs generated by the different approaches.

6 *Conclusion and future work*

This work presented a study of different strategies for tackling low-resource AMR-to-text generation for Brazilian Portuguese. We explore the helpfulness of additional translated corpus, different granularity levels in input representation, and three preprocessing strategies. It is worth noting this study can be helpful for work in other languages or meaning representations, mainly, when there is no pretrained models available.

Concerning the use of *translated* corpus, we can confirm its helpfulness. However, there are different contexts for each approach in which we can better leverage it. SMT improves its performance when the model is trained on the *translated* and *gold* corpora together. Neural models benefit from *translated* corpus more than SMT, even when these are trained on it solely. However, its join with the *gold* corpus can produce different results. In particular, G2S showed that there are structural divergences between *translated* and *gold* AMR graphs that can harm the performance when models are trained on both corpora. However, adding *translated* corpus to the development set allows to make the model more robust and achieve better performance.

About the representation levels, we highlight the use of finer-grained representations such as subwords and characters. Char-level seems to be the best option for NMT and bpe for G2S. However, it is worth noting that our study focuses on sentences of 23 tokens at maximum. This way, if we extend the work to longer sentences, bpe would probably performs better than char for NMT.

Finally, different combinations of preprocessing strategies are helpful for each approach, being preordering the best strategy for both machine translation approaches and delexicalisation for NMT. In the case of G2S, delexicalisation produces mixed results, being important just for word and char-level representations.

As future work, we plan to explore state-of-the-art approaches that are usually based on transformers, such as T5 (Ribeiro et al., 2020), or GPT-2 (Mager et al., 2020). Besides such issues, given some divergences between the *translated* and *gold* corpora that can harm the performance, it would be interesting to explore transfer learning for leveraging the knowledge learned from the *translated* corpus instead of training on both corpora together.

To the interested reader, more details about this work may be found at the web portal of the POeTiSA project at `https://sites.google.com/icmc.usp.br/poetisa`.

Acknowledgments

The authors are grateful to CAPES and the Center for Artificial Intelligence (C4AI - `http://c4ai.inova.usp.br/`) of the University of São Paulo, sponsored by IBM and FAPESP (grant #2019/07665-4). Besides, this research has been carried out using the computational resources of the Center for Mathematical Sciences Applied to Industry (CeMEAI) funded by FAPESP (grant 2013/07375-0).

References

Anchiêta, R. and T. Pardo. 2018. Towards AMR-BR: A SemBank for Brazilian Portuguese language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 974–979, Miyazaki, Japan. European Languages Resources Association.

Bahdanau, D., K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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

Beck, D., G. Haffari, and T. Cohn. 2018. Graph-to-sequence learning using gated graph neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 273–283. Association for Computational Linguistics.

Carmo, D., M. Piau, I. Campiotti, R. Nogueira, and R. Lotufo. 2020. Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*.

Castro Ferreira, T., I. Calixto, S. Wubben, and E. Krahmer. 2017. Linguistic realisation as machine translation: Comparing different mt models for amr-to-text generation. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 1–10, Santiago de Compostela, Spain. Association for Computational Linguistics.

Duran, M. S. and S. M. Aluísio. 2015. Automatic generation of a lexical resource to support semantic role labeling in Portuguese. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 216–221, Denver, Colorado. Association for Computational Linguistics.

Fan, A. and C. Gardent. 2020. Multilingual AMR-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2889–2901, Online, November. Association for Computational Linguistics.

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

Hartmann, N., E. Fonseca, C. Shulby, M. Treviso, J. Silva, and S. Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 122–131, Uberlândia, Brazil. Sociedade Brasileira de Computação.

Heafield, K. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, July. Association for Computational Linguistics.

Hedderich, M. A., L. Lange, H. Adel, J. Strötgen, and D. Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online, June. Association for Computational Linguistics.

Hieber, F., T. Domhan, M. J. Denkowski, D. Vilar, A. Sokolov, A. Clifton, and M. Post. 2017. Sockeye: A toolkit for neural machine translation. *ArXiv*, abs/1712.05690.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens,

C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Koehn, P., F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133. Association for Computational Linguistics.

Konstas, I., S. Iyer, M. Yatskar, Y. Choi, and L. Zettlemoyer. 2017. Neural AMR: Sequence-to-sequence models for parsing and generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 146–157, Vancouver, Canada, July. Association for Computational Linguistics.

Lavie, A. and A. Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the 2nd Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Lerner, U. and S. Petrov. 2013. Source-side classifier preordering for machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 513–523, Seattle, Washington, USA, October. Association for Computational Linguistics.

Levi, F. W. 1942. Finite geometrical systems.

Li, X., T. H. Nguyen, K. Cao, and R. Grishman. 2015a. Improving event detection with abstract meaning representation. In *Proceedings of the First Workshop on Computing News Storylines*, pages 11–15, Beijing, China, July. Association for Computational Linguistics.

Li, Y., D. Tarlow, M. Brockschmidt, and R. Zemel. 2015b. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*.

Liao, K., L. Lebanoff, and F. Liu. 2018. Abstract meaning representation for multi-document summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1178–1190, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Luong, T., H. Pham, and C. D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September. Association for Computational Linguistics.

Mager, M., R. Fernandez Astudillo, T. Naseem, M. A. Sultan, Y.-S. Lee, R. Florian, and S. Roukos. 2020. GPT-too: A language-model-first approach for AMR-to-text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1846–1852, Online, July. Association for Computational Linguistics.

Matthiessen, C. and J. A. Bateman. 1991. *Text Generation and Systemic-Functional Linguistics: Experiences from English and Japanese*. Pinter Publishers.

Okazaki, N. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Palmer, M., D. Gildea, and P. Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Popović, M. 2017. chrF++: words helping character n-grams. In *Proceedings of*

the Second Conference on Machine Translation, pages 612–618, Copenhagen, Denmark, September. Association for Computational Linguistics.

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

Ribeiro, L. F. R., M. Schmitt, H. Schütze, and I. Gurevych. 2020. Investigating pretrained language models for graph-to-text generation. *CoRR*, abs/2007.08426.

Sennrich, R., B. Haddow, and A. Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Sobrevilla Cabezudo, M. A., S. Mille, and T. Pardo. 2019. Back-translation as strategy to tackle the lack of corpus in natural language generation from semantic representations. In *Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019)*, pages 94–103, Hong Kong, China, November. Association for Computational Linguistics.

Sobrevilla Cabezudo, M. A. and T. Pardo. 2019. Towards a general Abstract Meaning Representation corpus for Brazilian Portuguese. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 236–244, Florence, Italy, August. Association for Computational Linguistics.

Song, L., D. Gildea, Y. Zhang, Z. Wang, and J. Su. 2019. Semantic neural machine translation using amr. *Transactions of the Association for Computational Linguistics*, 7:19–31.

Souza, F., R. Nogueira, and R. Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23*.

Xue, N., O. Bojar, J. Hajič, M. Palmer, Z. Urešová, and X. Zhang. 2014. Not an interlingua, but close: Comparison of English AMRs to Chinese and Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1765–1772, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Procesamiento del Lenguaje Natural, Revista nº 68, marzo de 2022, pp. 99-109 recibido 09-12-2021 revisado 18-01-2022 aceptado 25-01-2022

A Corpus of Spanish clinical records annotated for abbreviation identification

Un corpus de historias clínicas españolas anotadas para la identificación de abreviaturas

Mercedes Aguado,[1] Núria Bel[2]
[1] Università degli Studi di Milano
[2] Universitat Pompeu Fabra
mercedes.aguado@unimi.it, nuria.bel@upf.edu

Abstract: With the deployment of Electronic Health Records, much effort is being devoted to the development of Natural Language Processing tools that convert information described in these clinical records into structured data to be exploited. Clinical records main characteristic is that they are free text. They are normally written under pressure as memory notes and contain a high number of abbreviations that are an issue for automatic processing. In this article we present the IULA Spanish Clinical Records Corpus annotated for abbreviation identification.
Keywords: Abbreviations, annotated corpus, clinical records, preprocessing.

Resumen: Con la implementación de las historias clínicas electrónicas, se están dedicando muchos esfuerzos al desarrollo de herramientas de procesamiento del lenguaje natural que convierten la información descrita en estos registros clínicos en datos estructurados para ser explotados. La principal característica de las historias clínicas es que son texto libre. Normalmente se escriben deprisa, como notas de memoria y contienen un gran número de abreviaturas que son un problema para su procesamiento automático. En este artículo presentamos el Corpus de historias clínicas españolas del IULA, anotado para la identificación de abreviaturas.
Palabras clave: Abreviaturas, corpus anotado, historias clínicas, normalización, preprocesamiento.

1 Introduction

With the deployment of Electronic Health Records (EHR), much effort is being devoted to the development of Natural Language Processing (NLP) tools that convert information described in these clinical records into structured data that can be exploited. However, clinical records main characteristic is that they are free text, normally written under pressure as memory notes and containing a high number of abbreviations; they result in a telegraphic style that, on the one hand, is quicker for practitioners and experts to write, but on the other hand can be problematic for both human reading and automatic processing by NLP tools. Different methods have been applied to the normalization and analysis of clinical records to make them ready for the most used information extraction tasks like Named Entity Recognition and Classification (NERC), and Relation Identification and Extraction (see for instance Pathak et al., 2013, Gorinsky et al. 2019, Wang et al., 2018). The availability of annotated texts makes the use of machine learning supervised methods possible and allows for a fair comparison among these different methods. Thus, annotated corpora should be made available to support the development and improvement of methods and tools. However, most of the annotated corpora available are in English, as we will see in section 2. Related work, while EHR to be processed are written in many other languages around the world.

In this paper, we describe the IULA Spanish Clinical Record Corpus annotated for abbreviation identification (IULA-SCRC-ABB), a corpus of 3,194 sentences extracted from anonymized clinical records in Spanish annotated with abbreviations and the corresponding annotation guidelines. In the IULA-SCRC-ABB corpus, tokens that are shortened forms of words or phrases (including abbreviations, acronyms and symbols) were identified and tagged according to a linguistically motivated classification. The annotation comprised the identification of the short form, its classification into three classes, and the listing of possible long forms for each. The correct assignment of the long form is a task that requires expert knowledge on medical specialties and their practices and it has been left for future work. We also describe the annotation guidelines and discuss the most problematic cases. To the best of our knowledge, this is the first corpus of Spanish clinical records annotated for abbreviations that is made public and freely accessible at http://eines.iula.upf.edu/brat//#/AcronAbrevOnCR/ .

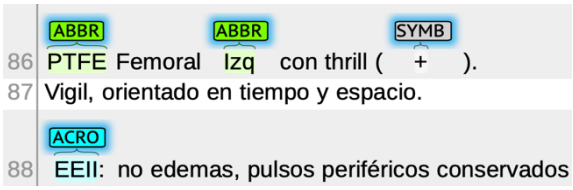


Figure 1: Sample of the IULA-SCRC-ABB corpus displayed with BRAT.

2 *Related work*

Existing medical text corpora annotated with abbreviations have been compiled as datasets of abbreviation identification tools. These corpora mostly contain scientific abstracts and articles in English (see, Islamaj Doğan et al., 2014 for details), although there are some in other languages, but only a few included clinical records. Névéol et al. (2018) and Dallianis (2018) are overviews of research carried out in clinical text mining in languages other than English, and Soto Montalvo et al. (2018), Sánchez and Martínez (2018), Sánchez León (2018), Castaño et al. (2018) and Cuadros et al. (2018) are descriptions of specific abbreviation identification tools for Spanish shown at IberEval2018-BARR, Biomedical Abbreviation Recognition and Resolution, evaluation campaign, whose corpus is described below.

We now describe other corpora consisting of clinical records annotated for abbreviations, and corpus annotated for abbreviations for Spanish, so that the IULA-SCRC-ABB and the methods we used to annotate can be compared with them. (Hua et al., 2007) used 16,949 admission notes from the internal medicine service of The New York Presbyterian Hospital Clinical Data Repository (NYPH-CDR) as the dataset for their machine learning system to be trained to detect abbreviations in clinical notes. For building the dataset, a physician manually reviewed the selected notes, listed all the abbreviations and specified their full forms. The training set consisted of 3,007 tokens of which 418 were abbreviations and the test set contained 2,611 tokens of which 411 were abbreviations. (Hua et al., 2007) also analyzed the abbreviations and further classified them according to the way they are formed. They used three classes: acronyms, shortened words and contractions. Acronyms were short forms, usually associated with multi-word phrases, which were formed by taking the first letter of each word in a phrase. Shortened forms were those which usually are a substring of a long word, although not always. Finally, contractions, considered another type of abbreviation, were those that consisted of an abbreviated contraction of multiple words with a separator (usually "/") between each word, for instance: 't/d/a', whose long form is "tobacco, drugs or alcohol".

Also for English, Wu et al. (2011) built a corpus for testing different machine learning methods for abbreviation detection. Three physicians manually annotated abbreviations in clinical documents randomly taken from the Vanderbilt Medical Center's Synthetic Derivative database, which contains de-identified electronic health records at Vanderbilt University Hospital. A total of 70 documents were annotated first with a pre-processing program that automatically labelled abbreviations using a reliable abbreviation dictionary. The human annotators revised these versions for identifying new abbreviations or removing wrongly labelled words. The developed corpus consisted in a training set with 40 documents of 18,225 tokens which contained 1386 abbreviations, and a test set with 30 documents of 13,913 tokens, containing 12,511 abbreviations.

Kvist and Velupillai (2014) reported about the annotation of two subsets of the Stockholm

Electronic Patient Record Corpus for Swedish (described in Isenius, 2012) with abbreviations and acronyms. The corpus consisted of randomly extracted emergency notes and radiology reports as to amount three sets of about 10,000 words each. Each subset was manually annotated for abbreviations by an expert. In this work, different types of abbreviations were identified: shortened words, *pat* for *patient*; contractions, *ssk* for *sjuksköterska* (nurse); and acronyms, *ECG* for *electrocardiogram*, although these classes were not used for annotation. Kreuzthaler et al. (2016) reported about the creation of a corpus of 1,696 de-identified clinical and outpatient discharge letters in German from the dermatology department of an Austrian university hospital. For training and testing different detection algorithms, instead of annotating the corpus, a list of words ending in period was extracted and manually annotated as whether the period was part of the word or not.

As for Spanish, Rubio-López et al. (2017) reported about having built a corpus to get the data for training and testing a system for abbreviation and acronym identification and disambiguation. The corpus consisted of 150 clinical notes in Spanish about stroke patients. These notes were selected by the high number of potential acronyms found. The notes were cleaned and manually annotated by researchers with a single label. To the best of our knowledge, this corpus is not accessible. Also for Spanish, (Intxaurrondo et al., 2018) described the Spanish corpus created for the BARR shared task held in the framework of IberEval 2017 and 2018 evaluation campaigns. The BARR track objective is to promote the development of biomedical and medical text mining tools together with offering an informed overview of the state of the art techniques and results obtained by the community. In the 2017 edition, the BARR track evaluated systems for detecting mentions of abbreviations-definition pairs: the discovery of abbreviations that were explicitly defined through the corresponding long form in the same sentence. The BARR corpus consisted of 1,050 abstracts for training and 600 abstracts for testing of biomedical articles from different sources. The corpus was manually annotated by biomedical experts. The corpus was annotated with information about abbreviations: short forms and long forms, as well with other relation-related information: derived short forms, global abbreviations, unclear, contextual, etc.; however, no classification of the different types of abbreviations was used. In 2018 BARR2 edition, a new corpus of clinical case studies has been delivered (Intxaurrondo et al., 2019). The corpus is divided into train, development and test sections with 122,594, 56,564 and 90,098 tokens respectively. It has been manually annotated by experts for the task of identifying abbreviations and delivering the corresponding long form. The documentation reports 9,552 annotated abbreviations in the whole corpus. This is the most similar corpus to the one presented here, although there are notable differences, as we will describe in the next section.

The corpus we present here, which is freely available, could be a contribution to BARR for future editions so that the shared task also includes authentic clinical records that present specific characteristics that made them different from scientific literature and clinical case studies.

3 Abbreviations in Spanish clinical records

Clinical records differ from Spanish for general purposes or even from other clinical texts written in Spanish in many linguistic features, such as lexical complexity, word and sentence composition, and sentence structure (see Benavent and Iscla, 2001, for a detailed description of linguistic characteristics of Spanish clinical texts which are very similar to the same genre in other languages as reported in Dallianis, 2018).

Clinical records in Spanish, as well as in other languages, show a higher density of technical terms and in particular of abbreviations (including acronyms, symbols, digits, capitalized letters within words, Roman digits and measurement units).

In our corpus, abbreviations amount 10.2% of the text tokens, while in the most similar corpus to ours, the BARR2 corpus made of clinical notes, abbreviations are about a 3.5% of the tokens. Isenius (2012) and Dallianis (2018) report figures similar to ours in discharge and emergency notes in other languages such as English and Swedish with a 15% of domain abbreviations. Note that the BARR2 corpus has collected clinical notes that are samples of edited text, while IULA-SCR-ABB contains spontaneous writing. The differences are also in

the distribution of the different types. BARR2 corpus is not annotated with types of abbreviations, but according to our annotation guidelines, the abbreviations of the test set would contain a 5.6% of abbreviations, a 67% of acronyms and 38.1% of symbols. In the IULA-SRC-ABB corpus, the distribution of categories is 14.6% abbreviations, 42.1% acronyms, 41.4% symbols and 1.5% unknowns (see section 4.4. Corpus Statistics for more details of the corpus). Moreover, the detection of abbreviations in clinical records is more difficult because of the following issues:

Differently to other general and medical texts, in spontaneous clinical records abbreviations do not occur along with the long form.

Sometimes, practitioners do not use the standard forms of acronyms and abbreviations. For instance, only 11,8 % of abbreviations were marked with a final period. Besides, others are made-up, such as the abbreviation sgto in our texts, which is wrongly used as the short form for segmento (segment), but the standard meaning is sargento (“sergeant”).

It is common an incorrect usage of symbols meant to be international standards, and the texts frequently show wrongly used symbols, such as grs, instead of g (“grams”), seg or min (for segundo “second” and minuto “minute”), which should be s and m, respectively.

Clinical records also exhibit a misuse of capital letters in abbreviations. For example, decilitro ("deciliter") was found written dl and dL; ecografía ("ecography") written ECO and eco; ABD and abd for abdomen.

Moreover, when working with authentic clinical records some other practical issues arise. For instance, we found uppercased full sentences, with uppercased wordforms which became homographs to abbreviations. For instance, we found *SE ADJUNTA* ("annexed"), where *SE* is the reflexive pronoun, but the form also corresponds to the shortened form of *sin especificar* ("unspecified").

4 The corpus

The IULA Spanish Clinical Record Corpus (IULA-SCRC) is a corpus of 3,194 sentences extracted from clinical records and annotated with negation markers and their scope (Marimon et al., 2017). The corpus was conceived as a resource to support clinical text-mining systems, but it is also a useful resource for other NLP systems handling clinical texts: automatic encoding of clinical records, diagnosis support, term extraction, among others, as well as for the study of clinical texts. The corpus was made publicly available with a CC-BY-SA 3.0 license.

This resource was obtained from a set of 300 anonymized clinical reports from several services of one of the main hospitals in Barcelona (Spain). In Table 1 we show the final number of sentences got from different sections of clinical records. Although the corpus was given to us already anonymized (all patient information was removed), the sentences were shuffled to make sure that no traceability of any data is possible.

Section	Sentences	%	Selected
Physical Exploration	5,193	34.61	1,090
Evolution	5,463	36.41	1,147
Radiology	1,751	11.67	367
Current Process	980	6.53	205
Explorations	1,619	10.79	339

Table 1: Statistics of corpus composition.

The texts from the IULA-SCRC corpus were taken as source for the abbreviation annotation, as we explain below. The IULA-SCRC-ABBR corpus is distributed as BRAT files (i.e. raw text and annotations in separate files) in UTF8 encoding and with a CC-BY-SA 3.0 license.

4.1 Pre-processing and pre-annotation system

Following standard practices, the IULA-SCRC-ABB texts were pre-processed to correct misspellings (Lai et al. 2015). Spelling errors are known to be very frequent in medical texts (Dallianis, 2018) and this becomes an issue for automatically processing the texts because misspelled words are not recognized. The texts of our corpus contained about a 2.6% of misspelled words. A 1.1% corresponded to segmentation problems, most typically the last character of a word becomes the first character of the following word (see Table 2 for examples). Missing accents corresponded to 0.86% of the misspellings being the second most frequent error. Other less frequent errors are character inversion, missing spaces, missing letters, unnecessary accents, unnecessary capital letters or wrong characters.

Error type	Error found	Correction
Segmentation	a lingreso	al ingreso (*upon admission*)
Missing accent	simetrica	simétrica (*simetric*)
Character inversion	peirfeira	periferia (*periphery*)
Missing spaces	segundos,sin	segundos, sin (*seconds, without*)

Table 2: Types and examples of misspellings found at the corpus.

Because these errors repeated several times along the different texts, a simple set of regular expressions was used to correct them automatically. Other misspellings were manually corrected, although as we explain in section 4.5. Difficulties and issues, the errors affecting abbreviations were not corrected.

Before annotating abbreviations, sentences were tokenized automatically using Freeling 4.1. (FL, Padró and Stanilovsky, 2012). For speeding up the abbreviation annotation task, we developed a script for identifying abbreviations by accessing a dictionary (like Hua et al., 2007). The dictionary was filled with abbreviations from the dictionary of Spanish medical abbreviations (Yetano Laguna and Alberola Cuñat, 2002) published as a reference for practitioners by the Spanish Health Ministry and other abbreviation lists freely available at the www. The script takes tokens, as found by FL, and makes a look-up at the dictionary database. In case of coincidence, the script retrieves the information about the class and long forms in the database and writes it in a BRAT annotation file. Thus, human annotators are provided with pre-annotated sentences. Annotators had to validate annotations, deleting errors and identifying and annotating new abbreviations missing in the database.

In order to tune FL to clinical text characteristics, we set up the following parameters. For the es.congif file:

- AlwaysFlush = yes, in order FL to process each line as an independent sentence.
- CompoundAnalysis = no. FL can guess compounds by splitting tokens into potential parts of a compound. This option was cancelled.
- QuantitiesDetection = no. FL can also recognize and normalize references to quantities and its measure. For instance, '234 €', is normalized into: 'CUR_EUR:234'. However, the actual spelling of medical measures in our texts showed a significant variation, as well as the use of different abbreviations for the same measure, so we preferred not to normalize them.

The FL named entity recognition module recognizes multiword units relying on uppercased words. However, given that some titles often appear in capital letters, there is a special heuristic to discard long sequences of uppercased words. We set TitleLimit to 1 to prevent the identification of series of uppercased words (a common case in our corpus) as named entities. Finally, since periods are used for sentence segmentation, FL requires the list of period-ending abbreviations to be in the tokenizer.dat file. Therefore, we included here those abbreviations that were at our database.

4.2 Annotation guidelines

In this section, we first introduce the different classes of abbreviations we have used and the motivation and underlying annotation criteria. Secondly, we describe the guidelines given to our two human annotators to identify and annotate abbreviations.

As explained in the Related Work section, most abbreviation annotation practices have not considered subclasses, although Cuadros et al. (2018), Wu et al. (2011) and Kreuzthaler et al. (2016) analyze the differences between short forms demonstrating that they exhibit different formal patterns. Our decision to use three classes for tagging short forms was based on the following differences related to how they are written and how the map to their long form.

In general, shortened forms that we call *abreviaturas* in Spanish usually end in a period; they keep the accented vowel of the long form, if any; they are written mostly with low-case characters and they can be plural forms, such as *págs*. for *páginas* ('pages'). Differently to acronyms, which are read as words (when phonetically possible), *abreviaturas* are read as the corresponding long form[1]. However, the

[1] These reading differences might be the cause of Spanish texts containing many acronyms which are shortened forms of English phrases as, in general,

samples we found in clinical records can deviate from the rules just mentioned and indeed they show high variability with several short forms for the same long form, as we will discuss in section 4.5. Difficulties and issues.

Eventually, the classes used for the annotation were the following:

- Symbols (SYMB): Symbols are tokens consisting of letters (either capital letters or lower cased) and other signs and numbers that are short forms of, mostly, internationally recognized measurement units, chemicals and mathematical terms. The tokens can be alphabetical (e.g., *r*) and non-alphabetical signs (e.g., % -percentage- ; and ∅, which means 'diameter', 'negative' and 'normal') and they never end in period.
- Abbreviations (ABBR): Abbreviations are those terms resulting from the removal of letters from one or more words. An abbreviation can contain letters and numbers, capitalized, lower case or both (starting the word or in other position within the word). Abbreviations also include special characters such as º, ª, er, which contract ordinal numbers (as in 1º, meaning 'first') and other kind of words (as in Hª, meaning 'history'), they can end in a period or not, and they can be hyphenated or not.
- Acronyms (ACRO): Acronyms are those terms that are formed by joining parts of two or more words. Usually, acronyms are composed of the initial part of each word, they can contain capital and/or lower-case letters and they do not end with a period. Besides, sometimes they are composed by one letter of one of the words and more than two letters of the other word, such as "AloTMO", whose long form is "Trasplante alogénico de médula ósea".
- Finally, the *Unknown* label (UNK) was devised to cover misspellings, typographical mistakes and cases where the abbreviation, acronym or symbol could not be attested in any resource.

In section 4.1. Preprocessing and Pre-annotation, we have explained that in order to reduce annotation time and required human resources, we used a simple lexical-lookup tool to pre-annotate the texts. Annotators, who were not practitioners, revised and corrected these pre-annotated texts using the BRAT annotation tool (Stenetorp et al., 2012). Human annotation task was about reviewing and validating the token identified as an abbreviation by the look-up system. In most cases, the pre-annotated class had not to be changed, as the information at the database is correct. However, corrections were required when:

a. A particular shortened form was annotated as belonging to more than one class, for instance: *m* can be both the abbreviation of *mes* (month) and the symbol of *metro* (meter), or *K*, which is the abbreviation of both "Kelvin" and "Karnofsky" and the symbol of *kilo* as well. The annotator had to choose the correct class, according to context.
b. A capitalized word was wrongly annotated as an abbreviation. Annotation has to be deleted.
c. Identifying new abbreviations. For abbreviations in the text but missing in the dictionary, annotators should find the long form and decide the class. The annotators searched for the candidate in different resources (medical dictionaries and databases like SNOMED, MESH, IATE and parallel and comparable corpora) to identify the abbreviation and to collect all possible long forms, if possible. When the annotators could not find the long form of a particular abbreviation, they used the label UNK (unknown), which was reserved for this case.

4.3 Inter-annotator agreement

For validating the annotation guidelines, we followed a two steps procedure. In a first round, the manual annotation task of identifying the short form, its type and possible long forms, as explained before, was performed by two annotators (a specialized translator and a linguist) over the whole set of documents. For the task of identifying the short form and assigning a type, the kappa inter-annotator agreement measure was 0.75. All the mismatches were studied and solved, and the guidelines were refined accordingly. To test the changes in the guidelines, a new round of annotation over 800 sentences was carried out by two new, non-medical expert annotators. The kappa measure was on average 0.75. The major source of disagreement was the distinction between acronyms and abbreviations

acronyms are more likely to become loanwords than their corresponding long forms.

specially for abbreviations in capital letters that were considered acronyms although were listed as abbreviations in different resources.

4.4 Corpus Statistics

The IULA-SCRC-ABBR corpus details are described in Table 3, where number of tokens annotated and number of types for each abbreviations class are presented.

Unit	Number of	Unique forms
Sentences	3,194	
Tokens	38,208	
ABBREVIATIONS	506	163
ACRONYMS	1,460	376
SYMBOLS	1,427	79
UNKNOWN	52	34

Table 3: Details of abbreviation annotation in the corpus.

ABB.	#	ACRO.	#	SYM.	#
U	47	TC	57	mg	188
Tª	27	PAD	54	%	153
mEq	25	PAS	52	mm	139
Rx	19	TP	37	Hg	102
Hb	15	MVC	34	dl	99
ABD	10	EEII	30	L	98
E.	10	FA	29	ºC	53
Dr.	9	VHC	28	h	50
Abd	8	GGT	27	dL	48
Dr	8	NIHSS	25	g	47

Table 4. Ten most frequent abbreviations, acronyms and symbols, and frequency.

Finally, we report about the script to pre-annotate sentences. The script was created just to reduce manual work as the task was to compare tokens in texts with the items of the database created out of an abbreviation dictionary. Table 5 shows the final figures of the corpus after revising the pre-annotated files.

	Manual addition	Final number
ABB.	142	506
ACRO.	190	1460
SYM.	64	1427
TOTAL	396	3393

Table 5: Manual additions after using the script for pre-annotation.

As for the performance of the script, data from the validation exercise with 800 sentences and 475 short forms to be identified and annotated showed that the script initially identified 438 short forms, of which 84, a 19%, had to be manually corrected. As explained in detail in section 4.5. Difficulties and issues, segmentation problems and wrong punctuation that specifically affected abbreviations were not corrected in the source text, thus preventing the script from matching them. We can see an example in Figure 2. Finally, the coverage of the database and the script was 74.5% as 121 forms, mainly abbreviations, were manually added.

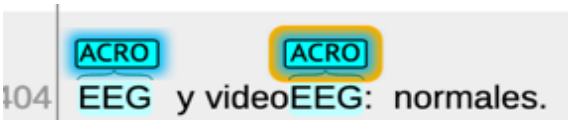


Figure 2. Example of a segmentation misspelling (missing space in videoEEG).

4.5 Difficulties and issues

The processing of health records in Spanish requires correctly identifying the units the text is composed of, including abbreviations. The task of abbreviation identification has been reported to be a challenging task. In Spanish, short forms are considered as belonging to one of three classes: Symbols, Abbreviations and Acronyms. This classification into three different classes should be of interest as it helps to understand the differences and therefore leads to better prediction features. The annotation allowed us to see that Abbreviations suffer of more variability than the other categories. For instance, a quite common term like *hematocrito* ('hematocrit') is written in 4 different ways: 'Hcto' (4)[2], 'Htc' (1), 'hto' (1), 'HTO' (2) and 'Hto' (6). Other samples are *creatinina* ('creatinine') whose abbreviation is 'creat' (2) although variations like 'Crea' (4), 'crea' (1) and 'Cre' (2) were found and *izquierdo* (left) that was found as 'izdo' (2) or 'izqdo' (3). In contrast, variation regarding acronyms is less frequent and the few cases of variations are spelling differences like in *angio-RM* (2) vs. *angioRM* (2) for 'Magnetic Resonance Angiogram', for instance. As for symbols, few cases of variation have been

[2] Frequency of occurrence in parentheses.

found like ‘mmHg’ (34) vs. ‘mmhg’ (2). The explanation of this higher variation for abbreviations could be that, as mentioned before, abbreviations are read as the long form, while acronyms are read as wordforms, what would support a better memorization.

Variations that divert from standard practices were annotated as Unknown, although a proposal for the correct short form and corresponding longform has been provided. As explained in section 4.5.4. Misspellings, the Unknown label was devised to cover misspellings, typographical mistakes and cases of forms looking like abbreviations that, however, were not attested in any of the identified resources (listed in 4.3). Eventually 52 tokens are coded as Unknown. These correspond to 34 unique forms (types) of which only 20 could not be annotated with a longform.

As we mentioned before, a first annotation round was useful to identify the different cases that were not covered by normative descriptions. Despite the apparent simplicity of the task, there were some cases that were difficult to classify and generated some discussions. Now, we list the most significant cases.

4.5.1 Acronyms or abbreviations in other languages, mostly English.

In Spanish practitioner’s reports, there is a surprising abundance of English abbreviations, mostly acronyms, even though there exist a corresponding Spanish term. This was the case of: *SOFA* for English ‘Sequential Organ Failure Assessment’, in Spanish *Evaluación secuencial de fallo orgánico*; DIVAS, for English ‘Digital Intravenous Angiography Subtraction’, in Spanish *Angiografía digital intravenosa de sustracción*; CKD for English ‘Chronic kidney disease’, in Spanish, *Enfermedad Renal Crónica*. For these English acronyms, the annotated long forms are the Spanish ones.

4.5.2 Single characters as type enumerations

Abbreviations that are just one character, usually uppercased, are considered to be a symbol when they have no ending period, e.g., cases such as *A*, in *Virus gripe A* (A influenza virus) because it is an international typing encoding system. We considered their long form to be *Tipo A*, *Tipo T*, etc. The same was done for *ondas T* (T waves).

We did the same for types of vitamins, hepatitis, clusters, etc. since the letters, by themselves do not have a meaning, but are the identifier of a type or a subtype. In some cases, the headword, that is, ‘vitamin’, is missing, for instance *hidroxil B12 B6*. In that case, we included the head in the long form, that is *Vitamina B12*, but we considered the abbreviation a symbol anyway.

Isolated letters in names of medicines and drugs, such as *Gentamicina S* –standing for *Sulfato*--, or *Levofloxacina R* --standing for *Richet*-- are classified as abbreviations, as they are types but not part of an enumeration. Thus, we followed BARR’s annotations for other examples as Proteins C and S, that were identified as abbreviations of ‘peak C’ and ‘Seattle”.

Short forms containing letters and numbers are classified as acronyms when their long form contains more than one word. Short forms containing letters and numbers are classified as ABBR when the letter (or letters) is itself an abbreviation and as symbols when the letters are neither abbreviations nor acronyms (i.e., the elements do not have a long form), but the term as a unit has a long form as in the case *M1 Segmento esfenoidal* (M1 Sphenoidal segment).

4.5.3 Parts of phrases

For terms formed by an abbreviation and a full word, e.g., “E. Coli”, “S. Neumoniae”, “S. Aureus”, “E. Faecium”, “E. Faecalis”, “E. Epidermidis”, only the abbreviation was annotated with its corresponding long form.

Hyphenated words were annotated as a unit, as both parts compose the term. However, if those words lack the hyphen, they are annotated separately.

4.5.4 Misspellings

As already mentioned by (Benavent and Iscla, 2001) incorrect variations of known abbreviations are quite frequent in clinical records. Incorrect forms together with other misspellings such as missing letters and wrong letter order were classified as unknown. However, the correct abbreviation together with the long form, taking the context into account, were suggested in the notes section of annotation. For instance, in Figure 3 we see the proposal for the case of *mgr* instead of *mg* (‘miligram’). Only in 20 cases, they were absolute unknown terms, for instance

"Orientado en espacio y persona, PINR, no refiere diplopía" (Space and person oriented, PINR, does not refer diplopia).

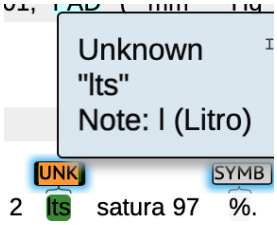


Figure 3. Detail of Unknown annotation: the note might contain the correct short form and corresponding longform.

We have handled the use of commas instead of periods as misspellings. For instance, "E. Coli" was also found in the texts as: "E.coli", "E coli", "E COLI", and "e, coli". They were annotated as abbreviations, whether it is a capital or lower-case character, and not including the comma.The incorrect use of upper and lower-case letters, such as "mmii" instead of "MMII" for *miembros inferiors* (lower limbs) or "eeii" instead of "EEII", for *extremidades inferiores* (lower extremities) is too frequent to be considered an occasional misspelling. We decided to annotate it as a correct form.

Another rather frequent misspelling was when the last part of a word wrongfully joins the next word. In our corpus, very often, the contraction "del" or "al" ('of the' and 'to the', literally) got separated and the letter "l" joins the next word, which can be very confusing and may hinder the identification of the term. In this case, we annotated the abbreviation and ignored the "l".

4.5.5 Mathematical symbols

Mathematical symbols are annotated for consistency. Some mathematical symbols are very ambiguous, but the context helps in deciding about the corresponding long form. Thus, we decided annotating the symbols and coding the appropriate Long Form according to the context. Some symbols deserve special clarification. We classified "+" as a symbol, and the actual long form for each case depends on the context. It can be:

- Positive
- Addition (mathematical symbol). Also, + is commonly used instead of "and".
- Intensity (as in edemas and swellings): + (mild, "leve"), ++ (moderate, "moderado"), +++ (severe, "intenso"), ++++ (serious, "muy severo")
- Levels in mg/dl: it is used in tests to determine the presence of proteins in urine.

We classified the "-" symbol as *negative* or *subtraction*, depending on the context, although, note it could be a hyphen too, but in this case there is no annotation. Other symbols were quotes (' and "), that were classified as symbols of minute and second respectively in the appropriate cases. Finally, we also had to make a distinction between roman numbers (for us, symbols like *IV ventrículo*, fourth ventricle) and acronyms (like *VI*, *ventrículo izquierdo*, left ventricle). We decided whether the term was an acronym or a roman number symbol according to context.

5 Conclusions

In this article, we have introduced the IULA-SCRC–ABBR corpus. It is a dataset of 3,194 sentences extracted from anonymized clinical records and annotated for abbreviation identification, including shortened forms, acronyms and symbols. The corpus was revised and validated by two human annotators. We have also described the annotation guidelines for the annotators, and the underlying criteria that motivated the choice of the three classes: ABBR, ACRO and SYMB. These underlying criteria were based on the characteristics of Spanish abbreviation and in relation with other abbreviation annotated corpora of clinical records already available, although for other languages, that is English, German and Swedish. To our knowledge, the IULA-SCRC-ABBR corpus is the first corpus of Spanish authentic clinical records annotated for abbreviations that is freely accessible under a Creative Commons BY-SA 3.0 license as this resource has been created for supporting the development of natural language processing systems for Spanish and their evaluation.

Acknowledgements

We would like to thank Dr. Pilar Bel, Laura Bernard, Miquel Cornudella, Jorge Vivaldi, and Montserrat Marimon for their collaboration in the annotation task and for their valuable comments and assessment. Research reported in this publication was partially supported by the Project PID2019-104512GB-I00 funded by Ministerio de Ciencia e Innovación (Spain).

References

Benavent, R.A., and A.A. Iscla, A.A. 2001. Problemas del lenguaje médico actual. (ii) abreviaciones y epónimos. *Papeles Med* 10(4), 170–6 (2001).

Castaño, J., P. Ávila, D. Pérez, H. Berinsky, Park, L. Gambarte, and D. Luna. 2018. A Simple Approach to Abbreviation Resolution at BARR2, IberEval 2018. In Rosso et al. (eds) *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*. http://ceur-ws.org/Vol-2150/ Accessed 6-12-2019.

Cuadros, M., N. Pérez, I. Montoya, and A. García Pablo. 2018. Vicomtech at BARR2: Detecting Biomedical Abbreviations with ML Methods and Dictionary-based Heuristics. In Rosso et al. (eds) *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*. http://ceur-ws.org/Vol-2150/ Accessed 6-12-2019.

Dallianis, H. 2018. Clinical Text Mining. Secondary Use of Electronic Patient Records. Springer. doi: 10.1007/978-3-319-78503-5.

Gorinski, Ph., H. Wu, C. Grover, R. Tobin, C. Talbot, H. Whalley, C. Sudlow, W. Whiteley, and B. Alex. 2019. Named Entity Recognition for Electronic Health Records: A Comparison of Rule-based and Machine Learning Approaches. *arXiv preprint arXiv:1903.03985.*

Hua X., P. Stetson, and C. Friedman C. 2007. A Study of Abbreviations in Clinical Notes. *AMIA 2007 Symposium Proceedings*: 821-825.

IATE, Interactive Terminology for Europe. http://iate.europa.eu, accessed 10-07-2021.

Isenius, N. 2012. Abbreviation Detection in Swedish Medical Records. The Development of SCAN, A Swedish Clinical Abbreviation Normalizer. Master's thesis, Department of Computer and Systems Sciences, Stockholm University.

Islamaj Doğan, R., D. C Comeau, L. Yeganova, and W.J. Wilbur. 2014. Finding abbreviations in biomedical literature: three BioC-compatible modules and four BioC-formatted corpora. *Database: The Journal of Biological Databases and Curation*, bau044. doi:10.1093/database/bau044

Intxaurrondo, A, M. Marimon, A. Gonzalez-Agirre, J.A. Lopez-Martin, H.M. Rodriguez, J. Santamaria, M. Villegas, and M. Krallinger, M. 2018. Finding mentions of abbreviations and their definitions in Spanish Clinical Cases: the BARR2 shared task evaluation results. In Rosso et al. (eds) *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages* (IberEval 2018). http://ceur-ws.org/Vol-2150/ Accessed 6-12-2019.

Intxaurrondo, A, J. C. de la Torre, H. Rodriguez, M. Marimon, J.A. Lopez-Martin, A. Gonzalez-Agirre, J. Santamaria, M. Villegas, and M. Krallinger. 2018. Resources, guidelines and annotations for the recognition, definition resolution and concept normalization of Spanish clinical abbreviations: the BARR2 corpus. Available at http://temu.bsc.es/BARR2/. Accessed 3-12-2021.

Kreuzthaler M., M. Oleynik, A. Avian, S. Schulz. 2016. Unsupervised Abbreviation Detection in Clinical Narratives. In *Proceedings of the Clinical Natural Language Processing Workshop*, 91-98.

Kvist M., and S. Velupillai (2014) SCAN: A Swedish Clinical Abbreviation Normalizer. In Kanoulas et al. (eds.): *CLEF 2014*, LNCS 8685: 62-73.

Lai, K.H., M. Topaz, F.R. Goss, and L. Zhou. 2015. Automated misspelling detection and correction in clinical free-text records, *Journal of Biomedical Informatics*, 55.

Marimon, M., J. Vivaldi, and N. Bel. 2017. Annotation of negation in the IULA Spanish Clinical Record Corpus. In *Proceedings of the Workshop Computational Semantics Beyond Events and Roles, SemBEaR*.

MESH. Spanish translation of Medical Subject Headings. https://www.nlm.nih.gov/mesh/. Accessed 10-07-2021.

Névéol, A., H. Dallianis, S. Velupillai, G. Savova, and P. Zweigenbaum. 2018. Clinical Natural Language Processing in languages other than English: opportunities

and challenges. *Journal of Biomedical Semantics* 9.

Pathak, J., K.R. Bailey, C.E. Beebe, S. Bethard, D.C. Carrell, P.J. Chen, D. Dligach, C. M. Endle, L.A. Hart, P.J. Haug, S.M. Huff, V.C. Kaggal, D. Li, H. Liu, K. Marchant, J. Masanz, T. Miller, T. Oniki, M. Palmer, K.J. Peterson, and C.G. Chute. 2013. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. *Journal of the American Medical Informatics Association: JAMIA*, *20*(e2), e341–e348.

Padró, Ll. and E. Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)* ELRA.

Rubio-López, I., R. Costumero, H. Ambit, C. Gonzalo-Martín, E. Menasalvas, and G.A. Rodríguez. 2017. Acronym disambiguation in Spanish electronic health narratives using machine learning techniques. *Studies in health technology and informatics:* 235-251.

Sánchez, Ch., and P. Martínez. 2018. A Simple Method to Extract Abbreviations Within a Document Using Regular Expressions. In Rosso et al. (eds) *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018).* http://ceur-ws.org/Vol-2150/ Accessed 6-12-2019.

Sánchez León, F. 2018. ARBOREx: Abbreviation Resolution Based on Regular Expressions for BARR2. In Rosso et al. (eds.) *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018).* http://ceur-ws.org/Vol-2150/ Accessed 6-12-2019.

SNOMED CT Spanish browser, http://browser.ihtsdotools.org/. Accessed 10-07-2021.

Soto Montalvo, S., R. Martínez, M. Almagro, S. Lorenzo. 2018. MAMTRA-MED at Biomedical Abbreviation Recognition and Resolution - IberEval 2018. In Rosso et al. (eds) *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018).* http://ceur-ws.org/Vol-2150/ Accessed 6-12-2019.

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 Session at EACL 2012*. (The tool is available at http://brat.nlplab.org/)

Wang, Y., L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, and H. Liu, 2018. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, *77*: 34-49

Wu, Y., S. T. Rosenbloom, J.C. Denny, R.A. Miller, S. Mani, D.A. Giuse, and H. Xu. 2011. Detecting Abbreviations in Discharge Summaries using Machine Learning Methods. *AMIA Annual Symposium Proceedings,* 1541–1549.

Yetano Laguna J. and V. Alberola Cuñat. 2002. Diccionario de siglas médicas y otras abreviaturas, epónimos y términos médicos relacionados con la codificación de las altas hospitalarias. Madrid: Ministerio de Sanidad y Consumo.

Procesamiento del Lenguaje Natural, Revista nº 68, marzo de 2022, pp. 111-121 recibido 19-12-2021 revisado 20-01-2022 aceptado 25-01-2022

A Methodology for the Automatic Annotation of Factuality in Spanish

Una metodología para la anotación automática de la factualidad en español

Irene Castellón Masalles,[1] **Ana Fernández Montraveta**[2] **Laura Alonso Alemany**[2]
[1]Universitat de Barcelona
icastellon@ub.edu
[2]Universitat Autònoma de Barcelona
Ana.Fernandez@uab.cat
[3]Universidad Nacional de Córdoba, Argentina
lauraalonsoalemany@unc.edu.ar

Abstract: In the last decade, factuality has undeniably been an area of growing interest in Natural Language Processing. This paper describes a rule-based tool to automatically identify the factual status of events in Spanish text, understood with respect to the degree of commitment with which a narrator presents situations. Factuality is represented compositionally, considering the following semantic categories: commitment, polarity, event structure, and time. In contrast with neural machine learning approaches, this tool is entirely based on manually created lexico-syntactic rules that systematize semantic and syntactic patterns of factuality. Thus, it is able to provide explanations for automatic decisions, which are very valuable to guarantee accountability of the system. We evaluate the performance of the system by comparison with a manually annotated Gold Standard, obtaining results that are comparable, if not better, to machine learning approaches for a related task, the FACT 2019 challenge at the IBERLEF evaluation forum.
Keywords: Factuality, event annotation, lexico-syntactic patterns, rule-based systems.

Resumen: La información factual es un área de investigación de creciente interés en el Procesamiento del Lenguaje Natural. Este artículo describe una herramienta basada en reglas para la identificación automática en español de la clase factual de los eventos en un texto, entendida con respecto al grado de compromiso con el que un narrador presenta las situaciones. En esta aproximación la información factual se representa compositivamente, considerando las siguientes categorías semánticas: compromiso, polaridad, estructura del evento y tiempo. A diferencia de los enfoques de Machine Learning, esta herramienta se basa por completo en reglas léxico-sintácticas y semánticas creadas manualmente que sistematizan los patrones semánticos y sintácticos de la información factual. Así, este sistema es capaz de proporcionar explicaciones para las decisiones automáticas, que son muy valiosas para garantía de la responsabilidad del sistema. Evaluamos el rendimiento del sistema mediante la comparación con un Gold Standard anotado manualmente, obteniendo resultados que son comparables, si no mejores, a los enfoques de aprendizaje automático para una tarea relacionada: el reto FACT 2019 del foro de evaluación IBERLEF.
Palabras clave: Factualidad, anotación de eventos, patrones lexico-sintácticos, sistemas basados en reglas.

1 Introduction and Motivation

The identification of factuality in corpora, i.e., recognizing the factual status of propositions, has been a research area of growing interest in Natural Language Processing (NLP) (Saurí, 2008; Saurí and Pustejovsky, 2009; Diab et al., 2009; Narita, Mizuno, and Inui, 2013; Soni et al., 2014), among others). In our project, following Saurí (2008)'s proposal, events' factual status is understood as the degree of commitment with which situations are presented by the narrator of the text.

The detection of this type of semantic information is extremely relevant for the semantic interpretation of texts and constitutes the base of several more complex processes and applications, such as fact-checking, fake news detection or information retrieval, among others, that need to be able to differentiate situations described as real from utterances of opinion or belief.

The work we present in this paper aims to build an automated annotator of factuality for Spanish texts exclusively based on linguistic knowledge. Unlike other annotators (Wonsever, Rosá, and Malcuori, 2016; Diab et al., 2009; Huang et al., 2019), our methodology only uses contextual linguistic knowledge, our algorithmic solution is solely based on linguistic cues. Currently, most automatic analyses of language are approached with purely statistical methods, machine learning using word embeddings, large neural language models and classifiers to perform the task (Wonsever, Rosá, and Malcuori, 2016; Huang et al., 2019; Qian et al., 2019; Rosá et al., 2020).

However, systems based on neural networks are obscure artifacts, which do not allow practitioners to understand how a given annotation has been made. For applications involving critical decision making, like fact checking, explainability is a must (Minh et al., 2021). It is important to be able to assess why a given text might be expressing a fact or a speculation, in order for people to ground their decisions with all relevant information. Indeed, regulations around the world are beginning to require that automated decision making systems can account for how decisions were reached, as in Spanish so-called rider law, which requires that companies disclose algorithms that they use to make decisions concerning labour rights [1]. In contrast with neural-based approaches, rule-based systems built upon relevant linguistic concepts provide adequate explanations, understandable by users. Some machine learning approaches, like decision trees or logistic regressions, can also provide some interpretability with respect to their decisions, however, they are dependant on big amounts of annotated text. Such big amounts of annotated text are usually not available, and, moreover, they may contain stereotypes and biases that are subsequently reproduced and amplified by the technologies based upon them, and are very difficult to detect and mitigate. In contrast, rule-based systems allow for explicit policing of biases, which makes it easier to implement positive policies and existing regulations.

Since the method this paper presents is solely based on linguistic knowledge, a prior thorough analysis of texts has been necessary to be able to identify the relevant knowledge, formalize it and systematize it as a system of rules. These rules basically exploit lexico-syntactic and morphological information that is able to capture the relevant semantic and syntactic phenomena that are related to factuality.

As will be developed below, this approach reaches good performance as evaluated on a gold standard test dataset. The domain chosen for this project, the written press, presents wide lexical diversity but is less complex in terms of the syntactic structures used, which we believe facilitates the approach of this task by means of conditions-actions.

Besides providing a tool for automated factuality analysis for Spanish, the good performance obtained by this approach allows us to automatically generate annotated corpora which will help compensate for the lack of annotated corpora at the factuality level and, in general, at the whole semantic level, for languages other than English.

The rest of the paper is organized as follows. In the next section, we describe the categories and tagset to annotate the different aspects of factuality. Then, Section 3 presents the methodology. In Section 4, we discuss some experiments to assess the performance of the system and analyze the results obtained, both quantitatively and qual-

[1] https://www.boe.es/diario_boe/txt.php?id=BOE-A-2021-7840

itatively. We then conclude with some future directions for this work.

2 Aspects of factuality to be annotated

To begin with, only declarative sentences are considered for annotation, given that interrogative or exclamatory sentences never assert facts. Only propositional content is annotated, not implications or implicatures. However, not all declarative sentences are annotated with respect to factuality, since statements describing desires or some conditional situations, for example, are not. The speaker of these types of sentences is not committing themselves to the truth of the proposition by asserting them because they do not describe 'real' situations (*realis*) and cannot be, therefore, said to be true or false (*irrealis*).

In this project, facts are understood as those present or past situations, presented by the author with commitment (that is, they are depicted as true by the speaker), and marked with positive polarity. Those situations that share these characteristics but are expressed with negative polarity are understood to depict counterfacts. All the rest of the situations belong to future or uncertain worlds and, therefore, are not considered facts. Event structure helps us to determine the eventive or stative nature of the situation described.

We use the tagset proposed within the TagFact project (Alonso Alemany et al., 2018) to describe these different aspects of facts or counterfacts, detailed in Table 1.

The combination of these four levels of linguistic description contribute to the factual interpretation of a proposition:

(1) *La presidenta madrileña que ha hecho de su hiperactividad mediática su capital político...*[2]
'The Madrid president who has made her political capital out of her media hyperactivity...'
Commitment, past, positive, event

(2) *Durante toda la jornada en la sede del Gobierno en la Puerta del Sol se aguardó por una rueda de prensa que nunca se produjo.*[3]
'Throughout the day at the Government headquarters in Puerta del Sol, they waited for a press conference that never took place.'
Commitment,past, negative, event

(3) *La Guardia Civil citará, además, como investigado para este jueves al exvicepresidente autonómico y exdirector general de la policía...*
'The Civil Guard will cite, in addition, as investigated for this Thursday the former regional vice president and former general director of the police...'
Commitment, future, positive, event

-2***Commitment**	Commitment
	Non-commitment
-2***Polarity**	Positive
	Negative
-3***Reference time**	Past
	Present
	Future
-5***Event Structure**	Event
	Mental
	Property-event
	Property-non-event
	Absolute truth

Table 1: Labels used in TagFact.

This tagset can be mapped easily to the ones used in other projects (Saurí, 2008) or (Wonsever, Rosá, and Malcuori, 2016). Below, Table 2 presents the mapping between our outcome and three other standard tagsets, Factbank, Fact Task (Rosá et al., 2019; Rosá et al., 2020) and (Qian et al., 2019), which is also based on FactBank.

This section has briefly described the tagset we used, proposed by the TagFact project, and how it relates to other similar projects. For a more detailed characterization of the tags and contexts see (Vázquez García and Montraveta, 2020).

[2] https://www.eldiario.es/politica/horas-Cifuentes-apago-focos_0_753474740.html

[3] https://www.elperiodico.com/es/politica/20180522/zaplana-detenido-por-la-guardia-civil-6832200

TagFact	**FACT task (Iberlef)**	**FactBank**	**Qian et al.**
Commitment Positive Present & Past Event - State	FACT	CT+ (certain) (incl. future situations)	CT+
Commitment Negative Present & Past Event - State	COUNTERFACT	CT-	CT-
Future Non-Commitment bv Event - State	UNDEFINED[4]	PR (probability) PS (possibility) U (undefined) PSu	PS[5]
Not applicable	UNDEFINED	U (undefined)	U

Table 2: Mapping of tagsets between TagFact and Fact Task, FactBank and Quian et al. (2019).

3 Methodology

The process leading to the creation of the automatic analyzer has been developed in the following three phases:

1. **Linguistic analysis and formalization**: The first stage consisted in the classification and characterization of the linguistic phenomena in factuality found in the TagFact corpus, a set of articles collected from several Spanish newspapers (Alonso Alemany et al., 2018; Fernández-Montraveta et al., 2020). This analysis has provided the necessary input for the implementation of the rules.

2. **Automatic processing**: This stage consisted in two steps. First, finding and evaluating the tools required for the linguistic preprocessing of the text and, second, the implementation and prioritisation of the rules. In this second step, we followed an incremental methodology: rules were developed and tested continuously, with a benchmark of representative cases for immediate assessment of the impact of each new rule, reordering or refactoring of the modules. Members of the implementation team regularly met with members of the linguistic analysis team to include new characterizations of the targeted phenomena, assess unclear cases and add further cases to the assessment benchmark.

3. **Evaluation**: Three different evaluations have been carried out, two during the development phase and the last one once the implementation was completed. The first kind of evaluation was qualitative, aimed to assess the impact of changes in the implementation: while rules were being developed, they were continuously tested against a benchmark obtained from the development corpus and some cases included ad hoc to monitor the behavior of the tool with respect to some phenomena of particular interest. The other two evaluations were quantitative and performed automatically by comparison with a manually tagged corpus. For this purpose the TagFact corpus, totalling 59.514 words, was divided into two parts, a development corpus with 82,6% of the total corpus, with 49.202 words, without manual annotations, and a test corpus, the Gold Standard, a 17,3% of the total corpus, with manual annotations. The test corpus was divided in two parts, the first consisting of three articles (1,9% of the total corpus, 1.141 words) used for the evaluation of the first implementation of the annotator, and the second (15,4% of the total corpus, 9.171 words) was reserved for the final evaluation. The evaluation was carried out quantitatively and qualitatively, the latter performed by members of the linguistic analysis team.

4 *Architecture of the automatic annotator for factuality*

The automatic annotation system detects candidate facts (situations) in text as any string tagged as a verb by the Freeling morphosyntactic annotator, and, for each situation, it assigns a value for each of the aspects of factuality described in Section 2. As a result of the combination of these aspects, the factual value of the situation is determined. The system is rule-based and works with the linguistic information available in the scope of the sentences. The rules have been implemented in Python3.6.

Within the first step of the second phase described in Section 3, the starting step of the automatic process consists in a morphosyntactic analysis of the text is carried out with Freeling (Padró et al. 2012). The output format of the analysis chosen is ConLL.As a result of this pre-process, all the lexical items that are annotated as verbs are identified as candidate facts.

The annotator of factuality consists of a sequence of sub-processes that incrementally characterise the different aspects of factuality for each of the identified situations. In the final stage, each different combination of values for the different aspects provides a standard factual value (Fact, Counterfact, Undefined).

The modules that make up the process are the following:

Module 1 selects the situations to be further annotated. In this module hypotheses, conditions and unreal worlds (irrealis) are discarded.

Module 2 assigns a polarity label to those situations selected in module 1.

Module 3 assigns a degree of commitment with which the situation is presented.

Module 4 is in charge of the analysis of referential time.

Module 5 assigns to each situation the label corresponding to the type of event denoted.

4.1 Applies

The first task is to decide whether factual analysis is applicable or not to each predicate. This is one of the most complex analyses since sentences may be describing situations in irreal worlds. In conditional constructions for example, not all hypotheses express unreal situations. A sentence, such as (8), is expressing two counterfacts, which are consistent with the real world, not situations in irreal worlds.

(4) *Si hubiera estudiado, habría aprobado.*
'Had I studied, I would have passed.'

The values assigned by this module are: Applies, Does_not_Apply and Non_pred, for those lexical items wrongly annotated as predicates in the pre-process. This assignment requires a complex analysis, which has been developed in three parts:

Local annotation: annotation of predicates in simple sentences. It takes into account the verb tense of the predicate and its context, except in the case of complex sentences with subordinate clauses, where there may be interference between the different predicates.

Conditional annotation between events: annotation inferred from the interaction of predicates between main sentences and subordinate sentences. For certain cases of non-personal forms, an inheritance mechanism has been implemented between the non-personal form and the verb that governs it.

Univocal predicate annotation: finally, some lexical items marked as predicates by the morphosyntactic pre-process are fixed forms that should not be annotated. This is the case of mira (look), used as an exclamation that formally corresponds to the imperative form of mirar (to look).

The linguistic information this module requires are: verbal tense, adverbs, conjunctions, prepositions, constructions, syntactic dependencies and syntactic functions. A total of 274 rules were developed, covering more phenomena than those appearing in the development corpus. This is because in the phase devoted to the linguistic analysis, other sources were consulted and an attempt was made to generalise the rules.

The predicates annotated with the category Applies continue the annotation process in the following modules. Events anno-

tated with the category Does_not-Apply and Non_pred are not further analyzed.

4.2 Polarity

The module dealing with the annotation of Polarity is composed of 38 rules. The label Positive is applied by default unless some triggers for Negative polarity are found in the co-text. Some examples of these kinds of triggers are adverbs of negation (5) such as no, nunca (never) or jamás (never, at no time), dependencies and syntactic functions to identify subjects or determinants (3 and 4) -nadie, ninguno, ningún + Noun (nobody, none, no + Noun) and some verb tenses (past perfect subjunctive -see (1)).

(5) *...creen que el dinero realmente nunca regresó a las arcas públicas.*[6]
'...they believe that the money never returned to the public coffers'

(6) *Nadie pone en duda la gran capacidad de trabajo que siempre ha demostrado Calvo,...*[7]
'Nobody doubts the great capacity for work that Calvo has always shown,...'

(7) *Aunque esta posibilidad en ningún momento ha sido confirmada.*[8]
'Although this possibility has never been confirmed.'

4.3 Commitment

The module in charge of assigning a Commitment value has 89 rules. It focuses on detecting expressions of doubt or uncertainty from triggers such as: creo que (I believe that), quizás (maybe), seguramente (surely), parece que (it seems that), etc. Some lexico-syntactic patterns restricted to some items have also been considered. Patterns such as the following:

(8) Existe (there exist) + det + Noun[trigger] + de (of) + que (that)
No + V0 + duda de que (there is no doubt that)
Verbs of opinion + que (that) +Verb

help us detect expressions such as sentences in (6-7):

(9) *Existe la certeza de que acudieron de noche.*
'There is a certainty that they came at night'

(10) *No le cabe la menor duda de que los empleados robaron en la sede.*
'He has no doubt that the employees robbed the headquarters.'

(11) *Considera que la solución no fue buena.*
'He considers that it was not a good solution.'

4.4 Time

In order to annotate referential time a total of 40 rules have been created. These rules deal with the recognition of referential time of simple and compound verb tenses, verb periphrases and non-personal verb forms. Besides, some rules have been developed to account for syntactic dependencies, as for example, a verb of communication in the present, if it has an animated subject refers to a past event (12):

(12) *Esa es su intención, afirma decidido, "cuando todo pase".*[9]
'That is his intention, he affirms decisively, "when everything passes".'

4.5 Event

Last, the module Event allows us to distinguish, basically, between states and events (50 rules). This module works with lists of event types. We distinguish between events, such as aprobar (pass), mental events, such as considerar (consider) and states such as tener (have). Starting from this category, the rules apply from triggers such as frequency adverbs (cada día, -every day) or specific verb forms (suele -used to or hay -there is). Some of the rules have to consult the analysis of syntactic dependencies (10) so that the category takes into consideration the co-text:

[6] https://www.elperiodico.com/es/politica/20181008/guardia-civil-desvela-gasto-32000-euros.prostibulo-fundacion-empleo.andalucia-7077756

[7] https://www.publico.es/politica/carmen-calvo-sera-vicepresidenta-del-gobierno.ministra-igualdad.html

[8] https://www.larazon.es/internacional/la-union-europea-y-reino-unido-podrian-haber.alcanzado-un-acuerdo-sobre-el-brexit.JE20174409/

[9] https://www.eldiario.es/desalambre/despues-juicio-queremos-mediterraneo-central_1_2138530.html

If a communication verbs has an inanimate subject the predicate is annotated as a state (property_non_event)

(13) *El artículo explica muy detalladamente el proceso de unificación.*
'The article explains in great detail the unification process.'

One problem that still remains to be addressed is differentiating between states (property_non_event) (14) and absolute truths or beliefs (15). Up to this moment we have not been able to formally differentiate them since, generally speaking, they share the same structure.

(14) *El presupuesto es alto.*
'The budget is high.'

(15) *La tierra es redonda.*
'The Earth is round.'

5 Gold Standard

The performance of the automatic annotator was evaluated by comparing automatic predictions against a manually annotated Gold Standard corpus.

We use a part of the Gold Standard corpus created within the TagFact project (Curell et al., 2020). It is composed of 22 press articles from Spanish generalist newspapers[10]. It contains a total of 10.272 words collected between June and September of 2020 (a mean of 553.7 words per article). The articles were mostly extracted from the Politics Section (70%) with the remaining 30% from other sections such as Economy, Sports or Technology, among others.

The corpus was first morpho-syntactically parsed and predicates were automatically identified by Freeling. Of a total of 1.696 words automatically marked as predicates only 1.319 remained after the manual phase. Then they were manually labelled (Section 2.1) by six senior linguists.

The interrater reliability was measured using Cohen's Kappa coefficient (Fernández-Montraveta Castellón in press) scoring 0,61 for the category Applies, 0,64 for Time, 0,55 for Polarity, and 0,35 for Event. The kappa value of the Commitment module could not be calculated because of lack of examples of one of the categories (Non-commitment).

[10]The articles were extracted from the following Spanish newspapers: ABC, El Diario, El Periódico and La Vanguardia.

	accuracy	**support**
Applies	81,1%	127
Time	81,35%	59
Commitment	100%	59
Polarity	98,30%	59
Event	66,1%	59

Table 3: Performance of automatic annotation in comparison with the Gold Standard.

6 Evaluation

We present here the mid-term evaluation of the project, aimed to detect how to improve the automatic annotator. This evaluation was carried out with a corpus manually annotated to calculate the inter-annotator agreement, with 127 predicates (Fernández-Montraveta et al., 2020). In what follows, we present a quantitative (6.1) and qualitative analysis (6.2) of results comparing the automatic and manual annotation.

6.1 Quantitative Analysis

Table 3 shows the general results of the comparison between the Gold Standard and the outcome of the automated process:

As can be observed, Commitment (100%) and Polarity (98,3%) are the categories showing the best behavior in terms of agreement. Second, the annotation of Applies and Time could be improved since both show around 81% accuracy, which is not a bad result but leaves room for improvement. Finally, Event is the category that shows the worst agreement rate. This fact could be explained because, first, it is the module that has more categories, some of them holding a type-subtype relation and, second, as mentioned above, the formal marks between some of them are blurred.

In a more detailed analysis, we can see the performance across the different classes for each level of analysis, as displayed in Table 3. We can appreciate that some of the proposed categories have not been evaluated because they were not found in the Gold Standard. This is the case of: non-commitment, future, property-event, mental and absolute truth.

Concerning the distinction between Applies / Does_not_Apply, we can see that the class Applies presents an F1 of 0.76, with high (0.92) recall but somehow lower 0.75 precision. Conversely, the category

'Does_not_Apply' shows a good precision (0.85) but recall drops (0.69).

Something similar happens for the Predicate / Non-Predicate distinction , with 0.90 precision but recall below 0.70. In this sense, the detection of predicates that require a factuality annotation needs to be improved.

Regarding verb tenses, present (F1 0.88) and past (F1 0.82) show good performance, with complementary distributions of precision and recall that suggest that errors in one category are confusions with the other, that is, predicates that should have been labelled as present are labelled as past and vice versa. Improvement is needed again in recall for the past tense and precision for the present. Future is underrepresented in the corpus so F1 cannot be calculated.

The annotation of Polarity reaches a very good performance, with 0,99 positive and 0,90 negative F1), as is the case with the Commitment tags (that reach 100%). Examples of non-commitment are not represented in the corpus.

Lastly, the category that shows poorer results is Event. Events have an acceptable 0.76 F1 but States perform much worse, with an F1 of 0.58, and the rest of the stative categories not even represented.

For the sake of comparison with related tasks, we have translated the annotations in the Gold Standard Corpus to the Iberlef FACT task, following the correspondence shown in Table 2. Results with this tagset can be seen in Table 5. The obtained F1 macro average is 75.6, which is better than the results obtained by machine learning approaches within the Iberlef FACT Task 1, shown in Table 6, albeit with a different corpus. We will apply the final version of this annotator to the Iberlef FACT corpus to have a more comparable assessment of performance.

Task 2 of the FACT 2019 challenge, Event Identification, is comparable to the Predicate aspect identified by our analyzer. Again, our results are comparable to those obtained by machine learning systems. We obtain 77% F1, while the only participating system for this task at FACT 2019 obtains 86.5% F1 and the baseline obtains 60%.

Therefore, our rule-based approach is competitive with machine learning approaches for a similar task, if not performing better. Nonetheless, the quantitative analysis shows that there is ample room for improvement. The categories requiring most effort to improve are Applies, Predicate and Event, and Time to a lesser extent. In what follows we carry out an analysis of errors on those categories to determine how to improve the performance of the analyzer.

6.2 Qualitative Analysis

We have carried out a systematic analysis of the cases where the automatic annotation fails, which has allowed us to create an inventory of system errors and elaborate a classification of the cases in which the analyzer fails. In order to present this classification, we describe the errors for each category.

6.2.1 Applies

The greatest number of errors in this category are predicates that should be tagged as Does_not_apply but are instead tagged as Applies. This is the case of some verb periphrases, infinitive clauses and conditional structures. Some of these problems, like the right interpretation of conditional sentences in example (17) are not an easy task to formalize in a systematic rule:

(16) *Si no se produce un acuerdo para devolver de oficio los intereses demás cobrados, el cliente bancario que esperaba la sentencia europea tiene la oportunidad de reclamar. "Primero tiene que hacerlo por vía extrajudicial, acudiendo al defensor del cliente,...".*
'If no agreement is reached to return ex officio the interest charged, the bank client who was waiting for the European judgment has the opportunity to claim. "First you have to do it extrajudicially, going to the client's ombudsman, ..

Other cases difficult to treat are infinitive clauses that do not inherit the category of the main verb because of errors in the pre-process of automatic parsing or the lack of a rule that runs through the syntactic structure.

(17) *Las entitades financieras han aprovechado la indefinición jurídica en torno a la retroactividad de las cláusulas para plantear a sus clientes cambios....*
'Financial entities have taken advantage of the legal uncertainty around the retroactivity of the

	Precision	Recall	F1
Applies			
Applies	0,753	0,920	0,828
Does_not_Apply	0,853	0,686	0,760
Predicate			
Non_pred	0,909	0,667	0,769
Time			
Present	0,795	1	0,886
Past	1	0,708	0,829
Future	na	na	na
Polarity			
Positive	1	0,981	0,990
Negative	0,833	1	0,909
Commitment			
commitment	1	1	1
non-commitment	na	na	na
Event			
event	0,851	0,696	0,766
property-non-event	0,555	0,625	0,588
property-event	na	na	na
mental event	na	na	na
absolute truth	na	na	na

Table 4: Precision, recall and F1 of the different classes for each category. When no cases were found in the corpus, "na" is reported.

	Prec	Rec	Acc	F1
Counterfact	0,833	0,555	0,954	0,667
Fact	0,924	0,710	0,812	0,803
Undefined	0,694	0,943	0,806	0,800

Table 5: Precision, recall and F1 of the automatic annotator in the Gold Standard corpus, where categories have been translated to the Iberlef Fact Task Category.

Participant	Macro-F1
t.romani	60.7
guster	59.3
accg14	55.0
trinidadg	53.6
premjithb	39.3
garain	36.6
FACT_baseline	24.6

Table 6: Results obtained in FACT 2020 Task 1, Factuality Determination.

clauses to propose changes to their customers.'

Thus, it seems difficult to address these errors in the next version of the annotator. Other errors, however, can be addressed, by incorporating additional rules to the annotator. For example, past participles premodified by a determinant ("lo cobrado" – *what is charged*) are currently tagged as Applies but a rule will be added so that they are tagged as Does_not_Apply. Additional rules will be incorporated to treat some modal periphrases that have been incorrectly labelled as Does_not_Apply.

6.2.2 Time

Most of the errors in the detection of the referential time are produced by the rule that

asserts that, in the press domain, a diction-communication verb with an animated subject [human], although a present indicative morphologically, is assigned a past value in the category referential time (18).

(18) *"Hay mucha información útil en YouTube, pero también mucha información errónea", afirma en declaraciones a The Guardian la profesora y autora del estudio Asheley LLandrum, ...* [11]
'"There is a lot of useful information on YouTube, but also a lot of misinformation," professor and study author Asheley LLandrum affirms (told) The Guardian, ...'

This temporal change does not happen when the entity is inanimate (19).

(19) *Las entrevistas realizadas a estas personas demuestran, según el estudio de la Texas_ Tech_ University, que la mayoría basan sus creencias en los vídeos que han visto en YouTube.*[12]
'The interviews carried out with these people show, according to the Texas_ Tech_ University study, that most base their beliefs on the videos they have seen on YouTube'.

In order to apply this rule, a list of animated entities was created, but still the rule fell short to account for the following cases:

- Some entities denoting collective entities were not in the list of animated entities, although they behave as such with respect to time: associations or offices, among others. These will be included in the updated list of animated entities.
- When ellided subjects were not retrievable, the rule could not apply properly.
- Errors in the syntactic pre-processing to detect the subject.
- Verbs that were not in the list of diction-communication, which will be included in the updated list for the improved version of the annotator.

[11] https://www.lavanguardia.com/tecnologia/20190219/46572983466/asi-alimenta-youtube-teorias-afirman-tierra-plana.html

[12] id. supra

6.2.3 Event

It is the category where the most errors have been detected. The annotation of this category has required creating lists of verbs lexically classified as states or events as the basis of the rules. That notwithstanding, contextual information might change the lexical event structure. In general, errors in this module come from the following factors:

- The verb of the sentence is not in the corresponding list. These have been included in the improved version.
- Lack of specific rules: for example, an inanimate object plus a communication verb produces a stative interpretation. This rule has been included in the improved version of the annotator.
- Some words were not included in the list of animated entities, and thus the relevant contextual rules could not be applied. They have now been included.

Another source of error with respect to events is that the detection of a special subkind of states, namely absolute truths, is beyond the scope of the automatic analyzer. This, however, cannot be properly addressed in the updated version of the analyzer either.

7 Conclusions and Future Work

In this paper we have presented a symbolic, rule-based system to automatically annotate factuality in Spanish text. Factuality is annotated compositionally, distinguishing different aspects of its semantics: commitment, time, eventuality and polarity. The total number of rules developed is 491, where 274 deal with searching for annotable candidates and 217 rules annotate values for the four categories.

We have shown that this approach performs comparably, if not better to machine learning approaches for the same task, but it still has room for improvement. An extensive error analysis shows where to direct efforts for future improvements, by including further rules or enhancing lists of words. Limitations of the approach have also been clearly depicted, for example, lack of accuracy due to errors in the morphosyntactic pre-processing. A future version of the analyzer will include these improvements, and will be evaluated in a holdout annotated dataset, as well as in the standard Iberlef FACT dataset.

References

Alonso Alemany, L., I. Castellón Masalles, H. Curell, A. Fernández Montraveta, S. Oliver, and G. Vázquez García. 2018. Proyecto tagfact: Del texto al conocimiento. factualidad y grados de certeza en español. *Procesamiento del Lenguaje Natural*, 61:151–154.

Curell, H., G. Vázquez, I. Castellón, A. Fernández-Montraveta, and L. Barrios. 2020. Un gold standard sobre factualidad para el español. In *III Congreso Internacional de Lingüística Computacional y de Corpus*, Colombia. Universidad de Antioquía.

Diab, M., L. Levin, T. Mitamura, O. Rambow, V. Prabhakaran, and W. Guo. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 68–73, Suntec, Singapore, August. Association for Computational Linguistics.

Fernández-Montraveta, A., H. Curell, G. Vázquez, and I. Castellón. 2020. The tagfact annotator and editor: A versatile tool. *Research in Corpus Linguistics*, 8(1):131–146, May.

Huang, R., B. Zou, H. Wang, P. Li, and G. Zhou, 2019. *Event Factuality Detection in Discourse*, pages 404–414. 09.

Minh, D., H. Wang, Y. Li, and T. Nguyen. 2021. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, 11.

Narita, K., J. Mizuno, and K. Inui. 2013. A lexicon-based investigation of research issues in japanese factuality analysis. In *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan*, pages 587–595.

Qian, Z., P. Li, Q. Zhu, and G. Zhou. 2019. Document-level event factuality identification via adversarial neural network. 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 2799–2809, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Rosá, A., L. Alonso, I. Castellón, L. Chiruzzo, H. Curell, A. Montraveta, S. Góngora, M. Malcuori, G. Vàzquez, and D. Wonsever. 2020. Overview of fact at iberlef 2020: Events detection and classification. *CEUR Workshop Proceedings*, 2664:197–205.

Rosá, A., I. Castellón, L. Chiruzzo, H. Curell, M. Etcheverry, A. Fernández, G. Vázquez, and D. Wonsever. 2019. Overview of fact at iberlef 2019 factuality analysis and classification task. *CEUR Workshop Proceedings*, 2421:105–110.

Saurí, R. and J. Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268.

Saurí, R. 2008. *A Factuality Profiler for Eventualities in Text*. Ph.D. thesis, 01.

Soni, S., T. Mitra, E. Gilbert, and J. Eisenstein. 2014. Modeling factuality judgments in social media text. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 415–420, Baltimore, Maryland, June. Association for Computational Linguistics.

Vázquez García, G. F. and A. M. Montraveta. 2020. Annotating factuality in the tagfact corpus. In M. Fuster-Márquez, C. Gregori-Signes, and J. S. Ruiz, editors, *Multiperspectives in Analysis and corpus design*. Comares, Granada. ISBN 9788413690094.

Wonsever, D., A. Rosá, and M. Malcuori. 2016. Factuality annotation and learning in Spanish texts. In *LREC'16*, pages 2076–2080, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Procesamiento del Lenguaje Natural, Revista nº 68, marzo de 2022, pp. 123-132 recibido 19-12-2021 revisado 04-02-2022 aceptado 09-02-2022

A Discourse Marker Tagger for Spanish using Transformers

Etiquetador automático de Marcadores Discursivos mediante Transformers

Ana García Toro, Jordi Porta Zamorano, Antonio Moreno-Sandoval
Universidad Autónoma de Madrid
{ana.garciatoro, jordi.porta, antonio.msandoval}@uam.es

Abstract: We present an automatic discourse particle (DM) tagger developed using manual annotation and machine learning. The tagger has been developed on a dataset of financial letters, where human annotators have reached an 0.897 agreement rate (IAA) on the indications of a specific annotation guide. With the annotated dataset, a prototype has been developed using the pre-trained Transformers, adapting it to the task (fine-tunning), reaching an F1-score of 0.933. An evaluation of the results obtained by the tagger is included.
Keywords: Discourse Markers, Spanish, fine-tuning Transformers.

Resumen: Presentamos un etiquetador automático de partículas discursivas (DM) desarrollado mediante etiquetado manual y aprendizaje automático. El etiquetador se ha desarrollado en un *dataset* de cartas financieras. Las anotadoras humanas han alcanzado un 0,897 de tasa de acuerdo (IAA) sobre las indicaciones de una guía de anotación específica. Con el *dataset* anotado se ha desarrollado un prototipo usando modelos de Transformers pre-entrenados adaptándolos a la tarea (*fine-tuning*) con un F1 de 0,933. Al final se da una evaluación de los resultados obtenidos por el *tagger*.
Palabras clave: Discourse Markers, Spanish, fine-tuning Transformers.

1 Why a Discourse Marker Tagger?

1.1 What is a DM?

Discourse Markers (DMs) are a large and heterogeneous group of invariable linguistic units that constitute intra- and supra-speech links for textual cohesion and coherence. Their primary function is to mark and define the relationship between the parts of the speech and to guide the inferences of the discourse from a procedural approach (Zorraquino and Portolés, 1999; Pons, 2000; Montolío, 2001; Briz et al., 2008; Fuentes 2009; Landone, 2012). Among the inferences that DMs guide are structuring information (1), counter-arguing opposite ideas (2), adding information (3), focusing on relevant nuances (4), introducing new arguments or statements (5), and reinforcing elements of the speech. In any case, they cohere and structure the discourse to be satisfactorily understood.

Here are some examples from our FinT-esp financial corpus (Moreno-Sandoval et al., 2020):

(1) *En 2015 nuestros dos objetivos fundamentales son:* ***por un lado****, seguir mejorando la franquicia comercial para estar en disposición de ganar cuota en una economía en crecimiento (...)*
(2) ***Por el contrario****, los resultados por operaciones financieras caen un 16% afectados por la volatilidad del mercado*
(3) ***También*** *es destacable el aumento del crédito, por primera vez desde 2008, por el impulso de empresas y pymes,* ***así como*** *el fuerte crecimiento en la producción de nuevas hipotecas*
(4) *Se trata de un resultado impulsado* ***principalmente*** *por el impacto de los 605 millones de euros de plusvalía obtenidos con la venta del 34% de Cellnex Telecom*
(5) ***Con respecto a*** *la tecnología, estamos invirtiendo fuertemente para ser más eficientes y abaratar los procesos*

Following previous definitions of DMs for written texts, we tried to find the best description

for our financial corpus. In financial texts, the writer's aim of using them is that the reader arrives at a particular interpretation of the utterances through certain inferences (persuasion). DMs are, therefore, essential keys to financial discourse.

It is important to consider that most of the definitions given by the literature have been applied to oral and written texts in general. For this reason, they are broader and less accurate definitions to include all DMs, despite their very different characteristics. They all have a semantic feature in common (with a few exceptions): DMs are characterised by a lack of referential or propositional content. Some authors (Llamas et al., 2010) have focused their taxonomy on one discourse's type or genre. In the case of Llamas et al. (2010), they have classified DMs in academic texts, while others have done it for oral discourse (Briz et al., 2008). In any case, their definition and classification have been a controversial field for scholars (Loureda and Acín, 2010) because each author considers different elements, concepts, and properties to categorise the DMs. Following previous definitions of DMs for written texts, we tried to find the best definition according to our financial, written, and formal discourse. Discourse Markers are a large and heterogeneous group of linguistic units that constitute intra- and supra-speech links for textual cohesion and coherence. The aim of using these elements by the sender of the text is that the interlocutor arrives at a particular interpretation of the utterances through certain inferences. Given the lack of studies on discourse markers in financial narratives, we provide this definition.

It should be noted that DMs are not grammatical elements, nor do they have a conceptual meaning. In other words, they do not have a defined place in the syntax; they act at various levels of discourse (depending on their function) and do not provide lexical information. DMs give information on how ideas in discourse are related.

Because of the diversity of the original categories (adverb, preposition, conjunction) and their behaviour in discourse, it isn't easy to establish a boundary between what is and what is not a DM. This makes the functional category of DMs a semi-open category. Let us say that not everything can function as a marker, but that, as the name suggests, they are words that mark discourse, and guide certain inferences. Even so, the list of DMs is neither closed nor defined according to certain established features; in fact, in this work we have been able to verify that the influence between languages can generate new incorporations of these elements into a language, as has happened in Spanish with *adicionalmente*, which we think comes from the English *additionally*. So, which words can mark the discourse and perform the functions of a DM? The complexity of this DM-tagger is, precisely, that we are dealing with a task that is difficult for linguists to define.

1.2 Why a DM-Tagger?

To understand the role of these particles, which may or may not be integrated into the sentence, we must know they form an essential part of it. The function of a DM-tagger goes beyond distinguishing them in context. It allows the reader to understand how discourse structures work, their distribution, and their purpose. Specifically, regarding the issue at hand, they can also give clues about the company's financial results. DMs seem to show imperceptible inferences in the text, guiding our thoughts and beliefs about the company. In short, DMs guide discourse and are tools of persuasion and manipulation, which are of great interest to speakers in business discourse.

A DM-tagger and its automatic annotation may involve the introduction of an objective measuring instrument that can resolve theoretical discussions. This tagger aims to reduce the inherent subjectivity that results from studies carried out with introspective methods and with which they end up doing manual and controversial classifications.

This tagger, applied to financial discourse, can be a tool applicable to any other type of discourse for identifying DMs. Knowing its distribution, functions, or behaviour are the first steps to a better understanding of the structure and construction of the discourse and, above all, to see how the discourse can be transmitted, manipulated, or lied through. DMs help study all discourse structures since they are part of them, whether in the sentence or outside it.

Of course, as we said, they are essential for discourse comprehension, but not indispensable; that is to say, they are a great help to the reader, reducing errors of interpretation and textual ambiguity.

Besides theoretical consequences, this work has several practical applications such as

discourse segmentation, information extraction, automatic summaries, or machine translation.

1.3 Previous Work on DMs Tagging

The main problem for the classification of DMs, and even more so for automatic classification, is the lack of consensus among scholars as to what is a DM and in which contexts an item is considered a DM, and in which are not. In our case, we also must consider the annotator bias in shaping the annotation guide. Proposals have been made since the 1990s to detect and systematise DMs. But neither the first investigations nor those carried out at the beginning of the century achieved results with a high percentage of precision and accuracy, due, once again, to the lack of consensus that exists when it comes to defining the units that do or do not fall into this group. Alonso et al. (2002), and, subsequently, Muller et al. (2016) undertook the construction of a computational lexicon of previously hand-coded DM, using the clustering technique to group markers (mostly connectors) that shared syntactic contexts or, in other words, that had similar behaviour. In both cases, the categories collected were hardly verifiable in the corpus because they had not been compiled based on an actual text but on a predefined lexicon.

Hernán et al. (2017) present a proposal for automatic induction of classifications of DMs that behave parenthetically (at the margin of the sentence separated by punctuation marks). They used a parallel corpus to automatically induce DMs categories according to the similarity between Spanish and English elements, without any prior annotation, which has not been done to date. In their update, Hernán and Nazar (2018) achieved high DM/non-DM decision accuracy.

Lastly, Rogelio Nazar, following his previous work in DMs identification and classification in Hernán et al. (2017) and Hernán and Nazar (2018), presents in Nazar (2021) a methodological proposal for the automatic induction of a multilingual taxonomy of DMs through parallel corpora. Using statistical calculations, he separates DMs from the rest of the units because, due to their low amount of referential information, they act "randomly" when grouped with other units in the text (as opposed to lexical units, which syntactically behave in a regular way). Then, once the DM candidates are selected automatically, they are aligned in pairs with their equivalents in other languages without any human intervention. In this way, the DMs in one language and similar DMs in others belong to the same category because they behave similarly. At first, these categories into which the DMs are grouped are not labelled with any name, but once they already contain a considerable number of DMs, the terminology followed is the one contributed a couple of decades ago by Zorraquino and Portolés (1999). Finally, this clustering technique is used to obtain and classify new units from these categories.

They collected 2636 items divided into 70 categories, which human annotators then reviewed. The review revealed that the model had 95% accuracy in the languages chosen for the experiment: English, Spanish, Catalan, French and German, except the latter, with 84%, probably due to the morphological characteristics of this language.

One of the disadvantages of this type of methodology is the 100% automatic selection and classification of DMs, in which the context is not considered. The information around a functional element such as these is important because they function as DMs in some contexts, while in others do not. Discerning contexts in which a DM functions seems to be a task that requires previous human annotation.

As for us, we provide an approach to the study of DMs in financial narrative from an actual perspective of their behaviour and distribution. Our analysis is based on a corpus annotated and contrasted by two annotators. However, it is still subject to certain underlying theoretical conjectures of linguistic introspection and the foundations laid by experts.

2 Dataset

The documents used in this research belong to the financial domain, characterised by a specialised language and a particular communicative exchange. The interlocutors are usually specialists in the field of finance and business. We will focus on letters written by managers to their investors (see 2.1).

The financial narrative in Spanish, in contrast to English, presents an excessively technical discourse with a significant contribution of English terminology (Mateo 2007, Vargas and Carbajo 2021). Mateo (2007) goes so far as to state that financial texts in Spanish are obscure and complex and that the reading of the financial

press is so dense that its content is not within reach of the non-specialist reader.

On the other hand, the exhortative communicative function predominates in the particular documents considered here (see 2.1) The sender intends to convince the receiver of their company's benefits so that he/she invests in it.

For this reason, we have found it to be a good testing ground for the use of DMs in argumentation.

2.1 Letters to Shareholders Corpus

Letters to Shareholders (LTS) is a sub-type of the financial narrative genre. They are the summaries that appear in companies' annual reports. It has recently attracted some interest in the NLP field: El-Haj et al. (2019), Moreno et al. (2019), and Bel et al. (2021).

Gisbert (2021) describes the two argumentative strategies used by managers:

a. Emphasising the company's good results, thanks to good management.

b. Hiding negative information that affects the expectations and reputation of the company and its managers.

For the work presented in this paper, we have chosen a subset of 397 letters in Spanish, with a length of 462,189 words and 16,800 sentences.

2.2 Annotated Dataset

Linguists have manually annotated the LTS corpus with DM tags in different stages, explained in section 3.2. In the complete annotation process (see section 3.2), 3170 DMs have been annotated, which appear in a total of 6432 sentences, containing a total of 154219 tokens. The distribution in each phase is shown in Table 2 (see section 4).

3 Annotation Process

3.1 Guidelines

The task of our annotation guide was to collect only the Discourse Marker (DM) category; this is to say, annotating only terms that were discourse markers.

Multi-label annotation to account for sub-categories has not been handled in this phase and is left for further work since we only wanted to approach DMs annotation. Revising previous work on DMs classification, we noticed the classification criteria' issues. Defining a Discourse Marker and its limits as a functional element is complex enough, considering that the classification used for a type of text is useless for others (notice the differences between all the DMs used in oral discourse that are never used in written texts). In addition, researchers disagree with groups of DMs and their sub-groups. These were the reasons why, for this study, as a first step, it was decided to annotate the binary task of DM/non-DM.

We train the model with a more significant number of DMs, although some of them may be under-represented. We prioritised coverage over accuracy.

The criteria followed in this annotation guide aim to reduce the complexity for the machine of learning contextual nuances. It should be noted that no grammatical rules are involved in this functional class, so it is more evident that True Positives (TP) must be deduced from a broader context.

Besides, in some cases, we had many difficulties in agreeing to consider items as DM o non-DM because they did not appear in any DMs classification for Spanish nor English. Adverbs ending in *-mente* are one of those cases, as they should not be systematically considered DMs.

For instance, *especialmente* works as an adverb in some contexts (6):

(6) *En España destaca* ***especialmente*** *el negocio de Automóviles* (= ‘de manera especial’)

In other contexts, it has the function as a DM of highlighting (7), and it can be rephrased as a quantity adverb, standing out a member of discourse:

(7) *En nuestro caso la ejecución ha sido* ***especialmente*** *difícil* (= ‘muy’)

Another example of ambiguous DM is *con respecto al*, that functions as DM introducing a topic when it is at the beginning of a sentence (8), but not when it has a comparison function (9) or when it is in the middle of the sentence (10). We decided to annotate *con respect al* only when it appears at the beginning of a phrase or paragraph introducing a new idea:

(8) ***Con respecto a*** *la tecnología, estamos invirtiendo fuertemente para ser más eficientes y abaratar los procesos (DM)*

(9) *Ha mantenido el volumen de actividad* ***con respecto al*** *año anterior* (no DM)

(10*) En atención al compromiso adquirido hace un año* ***con respecto al*** *cumplimiento de todas las recomendaciones* (no DM).

The guidelines (available for consultation here[1]) are organised following three types of criteria: General criteria (general rules), inclusion criteria (positive rules) and exclusion criteria (negative rules). Furthermore, there is a section where all DMs annotated in the corpus are collected (288 in total).

The most relevant criteria are provided in the following paragraphs.

General criteria: As a general rule, we annotate discourse markers included in general classifications and others that do not appear in taxonomies. Still, we have added some DMs typical of the financial language (*adicionalmente*). We annotate all the words that are part of a discourse marker: those which include prepositions and articles, as *además*/*además* ***de/*** *además* ***del***; or those followed by a nexus: *de tal forma* ***que***. No punctuation marks are incorporated in the annotations. Discourse markers are collected without commas or dots. As an exception, a comma should be included following the marker in enumeration with ordinals to avoid the ambiguities caused by these elements functioning as determinative adjectives (*primero, segundo*).

Inclusion criteria: DMs included in the annotation guide follow the three classification criteria established by researchers: a) semantic criteria, since their inferences help us to group them into homogeneous types; b) syntactic criteria, because these characteristics let us limit the DMs when they are part of a larger constituent; and c) morphological criteria, related to their nature, form and grammaticalisation process. The inclusion criteria of a DM have also been decided according to the given contexts.

Exclusion criteria: The main principle in the negative rules is that we do not include DMs with a low degree of grammaticalisation[2]. Besides, we have not annotated those items whose form is identical to DMs, but which function as modifiers in other parts of the speech. As regards to specific negative rules, we are not including metatextual or anaphoric markers (*en este contexto, a partir de ahí, sobre esta base, hasta el punto de, dicho lo cual, centrándonos en*, etc.); only *todo ello,* considering its degree of gramaticalization. Due to their variability, some DMs, particularly those which are addressed to the audience, have multiple combinations: *como ven, como bien conoce, no cabe ninguna duda de que*, etc., or others alluding to personal opinions: *en nuestro caso, a mi juicio, a nuestro juicio, en mi opinión*, etc. In these cases, we do not include them either.

Another negative rule is not annotating discontinuous discourse markers: *no solo… sino también* or comparative structures: *tan… como, más…que*, etc. We also do not annotate markers that incorporate an element that modifies only part of the marker, not the whole marker. This means that the DM has a small degree of grammaticalisation or is not grammaticalised in that example, so such cases are not included: ***gracias****, en cierta medida,* ***a; con el objetivo*** *claro* ***de***. However, they are not exceptions to those particles that could have two parts (discontinuous DMs): *por un lado… por otro; por una parte… por otra*, since they can work independently from the other part (we can only have *por un lado*, or *por otra parte*, and they are doing a function by themselves).

3.2 Manual Annotation

This process was divided into three phases:

a. Training the two annotators with the guide and the tool (Doccano[3]). In this phase, both annotators could consult each other's annotations to reach a consensus and prove they had acquired the required skills. This process helped to modify some definitions in the annotation guide. In total, 100 LTS were annotated, one-quarter of the dataset.
b. Creation of the Gold Standard (GS). Each linguist annotated 40 LTS in an utterly blind way (i.e. without knowing the annotation of the other linguist and

[1]http://www.lllf.uam.es/ESP/Publicaciones/guia_anotacion.html

[2] The grammatizalisation process consists in the acquisition of a new grammatical value for these lexical units, which implies a shift from a more referential meaning to a less referential one. For instance, *es más* does not mean the beginning of a comparison structure, if not it appears alone in the speech guiding an inference reinforcing the following information. Compare: ***Es más****, en el siglo XXI en el que ya nos adentramos, el avance cada vez más rápido de la tecnología en combinación con la gestión más profesionalizada de la economía // nuestro mínimo regulatorio* ***es más*** *bajo porque nuestro modelo está menos interconectado y es más fácil de resolver*.

[3] https://doccano.herokuapp.com/

consulting only with the guide). This part is the one that has been used to calculate IAA (see 3.3.). The GS has been generated by joint approval of the two annotators after knowing the IAA results. It was not necessary for a judge to decide discrepancies. A first DM tagger has been created with the 100 + 40 LTS.

c. Manual revision of the automatic tagging generated by the initial DM tagger model. Each annotator has post-edited 130 LTS and corrected the assigned tags per tagger. Each annotator has acted as an expert judge in deciding the final version. There is no cross-checking between annotators. The result is a Silver Standard (SS) of 260 LTS.

The DM tagger has been trained on the first and third datasets, leaving the GS for evaluation (see 5).

3.3 Interannotator Agreement (IAA)

The inter-annotator agreement (IAA) measures how well different annotators can make the same annotation decision for a specific category. IAA also reveals how clear the annotation guidelines are and how reproducible the annotation task is. Cohen's kappa coefficient (κ) is a statistic to measure the reliability between annotators. It is more robust than the simple per cent of agreement (or accuracy) since κ considers the possibility of agreement by chance: $\kappa = (P_o - P_e) / (1 - P_e)$ where P_o is the relative observer agreement among annotators and P_e is the probability of agreement by chance.

Two annotators, we will refer to as A and B, worked with 40 documents, accounting for 52,890 tokens (words and punctuations) in 1,759 sentences. Annotators A and B recognised, respectively, 850 and 756 discourse markers agreeing in 732 cases (annotator A identified 118 cases not recognized by B, and B 33 cases not recognized by A). The IAA computed with κ was 0.897, which can be interpreted as a remarkably high degree of agreement.

Annotators went back to agree on their disagreements to build a reliable set to measure the classifier's performance. The number of discourse markers finally agreed was 856. The comparison between the original annotations and the new agreed set, calculated with κ, were 0.957 for annotator A and 0.916 for B. As human classifiers, the performance of the annotators is shown in Table 1, and it will be used as a reference when evaluating the performance of an automatic classifier. The formulas used to calculate this performance of a classifier are precision = TP / (TP + FP), recall = TP / (TP + FN), and F1-score = 2 * precision * recall / (precision + recall), where TP are the number of true positives, FP the number false positives, and FN the number of false negatives. We used seqeval (Nakayama, 2018) to calculate them.

Annotator	κ	Precision	Recall	F1
A	0.957	0.955	0.949	0.952
B	0.916	0.970	0.867	0.915

Table 1: Annotator performance on the test set after agreement.

4 Model Training and Selection

We used pre-trained transformer-based language models to approach the problem of discourse marker detection as a token classification task with a IOB (Input-Outside-Beginning) annotation scheme and one category (DM).

The annotated data was split into the training and validation sets. This data was annotated before the IAA experiment. The data finally used in the experiments are shown in Table 2.

Set	Sentences	Tokens	DMs
Training	3,735	118,406	1,880
Validation	938	30,524	440
Test	1,759	52,890	856

Table 2: Annotated sets.

We experimented with BSC-BNE[4], a Spanish Roberta model (Gutiérrez-Fandiño et al., 2021), mBERT[5] (Devlin et al., 2019) and BETO[6] (Cañete et al., 2020), and XLM-Roberta[7] (Conneau et al., 2020).

[4]https://huggingface.co/BSC-TeMU/roberta-base-bne

[5]https://huggingface.co/bert-base-multilingual-cased

[6]https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased

[7] https://huggingface.co/xlm-roberta-base

Model	STL	LR	Epochs	BS	WU	Avg. F1
BSC-BNE	all	5e-5	3	8	0.0	0.928
BETO	all	5e-5	4	8	0.0	0.927
mBERT	first	7e-5	4	16	0.1	0.927
BETO	first	6e-5	4	8	0.1	0.926
BETO	all	6e-5	3	8	0.1	0.926

Table 3: Models performance on the validation set, where STL is the sub-token labelling strategy (first or all), LR is the learning rate, BS is the batch size, and WU is the warmup ratio.

Following the recommendations in the Appendix A.3 for fine-tunning mBERT models (Devlin et al., 2019), we performed a grid search of hyperparameters for each language model with learning rates: 2e-5, 3e-5, and 5e-5; epochs: 2, 3, and 4; batch sizes: 8, 16, 32; warmup ratios: 0 and 0.1; and three different seeds (0, 3, and 5). We used the AdamW optimiser with no weight decay and a warmup of 0 and 10% steps. For each model, we have two versions: one that only labels the first sub-token delivered by the model's internal tokeniser and a second version where all the sub-tokens of a token are labelled. We averaged the F1-score of the three runs with different seeds to assign performance to a classifier. F1-scores were calculated with the SeqEval package[8] on a DM basis, i.e., a DM is correct if all the tokens in a DM have received the correct IOB tag.

The five best-performing systems, shown in Table 3, had a remarkably similar average F1-score, and the worst-performing system of the 2160 systems tested had a 0.882 F1-score[9].

5 Evaluation and error analysis

Finally, we trained a system with the model and the hyperparameters of the best performing system in Table 3 (with the seed set to 0), and it was evaluated with the test set. Results for this system were a precision of 0.941, a recall of 0.925, and an F1-score of 0.933, which is right in the middle of the range defined by the two human annotators (0.915–0.952).

Regarding false positives (FP) on the test set (38 cases in total), 40.67% of the cases were considered true false positives by the human annotators:

- 24.01% of the FP cases corresponded to ill-formed IOB sequences, mostly tokens labelled with inside tags (I-DM) without a beginning tag (B-DM) on the preceding token.
- 16.66% of the FP cases, despite being well-formed according to the OIB scheme, were considered valid false FP.

The rest of the FP cases (59.33%) were regarded as actual discourse markers, and they were distributed as follows:

- 42.59% of the cases were actual discourse markers that went unnoticed by the two annotators in the test dataset but were present in the training set and the annotation guideline.
- 3.7% were valid right-side extensions of other known discourse markers also present in the training set: *además de*, *en consecuencia de* or *de tal forma que*.
- 12.96% were accurate discourse markers, overlooked by the annotators without any occurrence in the training set. The tagger has been able to generalise that they can be DMs. These are the most interesting results, as they show that the model has been able to resolve doubts that arise for human annotators.

Some of these "new" discourse markers can be considered generalizations done by the model: *a continuación*, *al margen del*, *con ello*,*del mismo modo que*, *en total*, and *posiblemente*. Some of these particles were identified during the design of the guidelines. However, we did not annotate them in the GS as proper DMs because we considered them fuzzy. But the ML model has been able to learn in fuzziness.

On the other hand, False Negatives (FN) are DMs that were annotated by the linguists in the GS but were not detected by the ML model. In total, there were 52 FNs, of which:

- 75% (39 cases) were actual DMs. Hence model errors.

[8] https://github.com/chakki-works/seqeval

[9] Experiments with Bi-LSTM models hardly reach 70% F1-score on test set.

- 25% (13 cases) were genuine DMs, which the human annotators did not detect, and the model did.

In summary, the model has improved the performance of humans proportionally more on the FP side than on the FN side. Out of 90 cases, the ML model hits 22 (24.44%) versus human annotators.

6 Conclusions and future work

The proposed model shows an F1-score (0.933) in the range defined by two annotators (0.915–0.952), and error analysis of the false positives cases in the test set reveals that the model was able to recognise a significant number of discourse markers that went unnoticed by the annotators, some of them seen in the training set and a few discovered by the model.

In conclusion, we can say that DMs are units that mark the discourse and give it cohesion and coherence to facilitate the reader the comprehension and interpretation of the text. We also have concluded that they are an open or semi-open functional category. That is, they are not a grammatical category, although most of the DMs are in a grammaticalisation process. We know, for sure, that they are units of the speech that mark the discourse. So, which words can mark the discourse and perform the functions of a DM? The complexity of this DM-tagger and everything related to DMs is, precisely, that we are dealing with a difficult task for linguists to define.

We assume this work is challenging, and it wasn't easy to define the criteria for considering an element DM or no-DM. There were and still are some doubts about the definition and the limits of these discourse units. Overall, we had difficulties with DMs coming from adverbial phrases because their context tends to be ambiguous. The guideline of this study, especially its negative criteria, must be revised. Nevertheless, our model can discover or annotate new DMs that were not initially annotated by humans, which means that NLP can somehow develop the capacity of detection DMs functionality beyond their form.

Further work will be, indeed, a Discourse Marker Tagger that classifies DMs into their types and subtypes (following the work begun by Hernán and Nazar (2018) and Nazar (2021), section 1.3) because this would provide us more information about the financial text and its factual inferences. We will look to study derived from usage data, with less reliance on language knowledge, using the methodology proposed by these authors. These steps may make us closer to defining Discourse Markers better than we used to do through human introspection.

The DMs tagger (under development) will be used in the annotation of argumentative structures. In particular, we are mainly interested in CAUSE-EFFECT (This has happened. *Consequently*, this other thing has happened) and counter-argumentative structures (This has happened. *However*, this other thing has also happened).

A DM-tagger and its automatic annotation may be an objective measuring instrument to help resolve theoretical discussions.

Acknowledgements

The research has been carried out within the CLARA-FINT project (PID2020-116001RB-C31), funded by the Spanish Ministry of Science and Innovation.

References

Alonso, L., Castellón, I., Gibert, K. and Padró, L. 2002. Lexicón computacional de marcadores del discurso. Procesamiento del lenguaje natural, 29:239–246.

Bel, N., Bracons, G., and Anderberg, S. (2021). Finding Evidence of Fraudster Companies in the CEO's Letter to Shareholders with Sentiment Analysis. *Information*, *12*(8), 307. http://dx.doi.org/10.3390/info12080307.

Briz, A., Pons, S. and Portolés, J. (coords.) 2008. *Diccionario de partículas discursivas del español*. Retrieved from www.dpde.es.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In PML4DC at ICRL-2020.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, Veselin. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the ACL*.

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

understanding. In *Proceedings of the NAACL*.

El-Haj, M., Rayson. P., Walker, M., Young, S., and Simaki, V. (2019). In search of meaning: lessons, resources, and next steps for computational analysis of financial discourse. *Journal of Business, Finance and Accounting*, *46*(3-4), 265-306. https://doi.org/10.1111/jbfa.12378.

Fuentes Rodríguez, C. 2009. *Diccionario de conectores y operadores del español*. Madrid: Arco Libros.

Gisbert, A. 2021. Financial narratives. In Moreno-Sandoval, (coord.), *Financial narrative processing in Spanish*. Valencia: Tirant lo Blanc, 15-50.

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Pio-Carrino, C., Gonzalez-Aguirre, A., Armentano-Oller, C., Rodriguez-Penagos, C. and Villegas, M. 2021. Spanish Language Models. https://arxiv.org/abs/2107.07253.

Landone, E. 2012. La clasificación de los marcadores del discurso y su valor operativo. In *Cassol, A., XXIV Congresso AISPI*, 431-444. Roma: AISPI Edizioni.

Llamas Saíz, C., Martínez Pasamar, C., and Tabernero, Sala, C. 2012. La comunicación académica y profesional. Usos, técnicas y estilo. Pamplona: Thomson Reuters / Aranzadi (pp. 140-141).

Loureda, Ó. and Acín, E. (coords.) 2010. *Los estudios sobre marcadores del discurso en español hoy*. Madrid: Arco/Libros.

Martín Zorraquino, M.A. and Portolés, J. 1999. Los marcadores del discurso. In *Gramática descriptiva de la lengua española*. Madrid: Espasa Calpe, 4051-4214.

Mateo Martínez, J. 2007. El lenguaje de las ciencias económicas. In E. Alcaraz, J. Mateo and F. Yus (Eds.), *Las lenguas profesionales y académicas* (pp. 191-203). Barcelona: Ariel.

Montolío, E. 2001. *Conectores de la lengua escrita. Contraargumentativos, consecutivos, aditivos y organizadores de la información*. Barcelona: Ariel.

Moreno-Sandoval, A., Gisbert, A., Haya, P.A., Guerrero, M. and Montoro, H. 2019. Tone analysis in Spanish financial reporting narratives. In M. El-Haj, P. Rayson, S. Young, H. Bouamor and S. Ferradans (Eds.), *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019*) (pp. 42-50). Turku: Linköping University Electronic Press.

Moreno-Sandoval, A. Gisbert, A. and Montoro, H. 2020. FinT-esp: a corpus of financial reports in Spanish. In Fuster-Márquez, Gregori-Signes and Santaemilia-Ruiz (eds.) *Muliperspectives in analysis and corpus design*. Granada: Comares, 89-102.

Muller, P., Conrath, J., Afantenos, S. and Asher, N. 2016. Data-driven discourse markers representation and classification. In *TextLink– Structuring Discourse in Multilingual Europe Károli Gáspár University of the Reformed Church*. Hungary, Budapest.

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

Nazar, R. (2021). Inducción automática de una taxonomía multilingüe de marcadores discursivos: primeros resultados en castellano, inglés, francés, alemán y catalán. In *Procesamiento del Lenguaje Natural*, (67), 127-138.

Pons, S. 2000. Los conectores. En A. Briz and Val.Es.Co, (eds.), *¿Cómo se comenta un texto coloquial?* Barcelona: Ariel, 193–220

Robledo, H., Nazar, R and Renau, I. 2017. Un enfoque inductivo y de corpus para la categorización de los marcadores del discurso en español. In *Proceedings of the 5th International Conference "Discourse Markers in Romance Languages: Boundaries and Interfaces"*, 91–93. Université Catholique de Louvain, Belgium.

Robledo, H. and Nazar, R. 2018. Clasificación automatizada de marcadores discursivos. In *Procesamiento del Lenguaje Natural*, (61), 109–116.

Vargas-Sierra, C. and Carbajo-Coronado, B. 2021. Anglicisms in Financial Narrative: The Case of the Annual Reports and Letters to Shareholders. In Moreno-Sandoval, (coord.) *Financial narrative processing in Spanish*. Valencia: Tirant lo Blanc, 99-134.

Procesamiento del Lenguaje Natural, Revista nº 68, marzo de 2022, pp. 133-142 recibido 29-11-2021 revisado 08-02-2022 aceptado 10-02-2022

Readers versus Re-rankers in Question Answering over COVID-19 scientific literature

Readers versus Re-rankers para la Búsqueda de Respuestas sobre COVID-19 en literatura científica

Borja Lozano, Javier Berná, Anselmo Peñas
UNED NLP and IR Group
Universidad Nacional de Educación a Distancia (UNED)
{blozano, jberna, anselmo}@lsi.uned.es

Abstract: In this work we present a comparison between the two most used neural Question Answering (QA) architectures to solve the problem of information overload on COVID-19 related articles. The span extraction (reader) and the re-ranker. We have found that there are no studies that compare these two methods even though they are so widely used. We also performed a search of the best hyperparameters for this task, and tried to conclude whether a model pre-trained with biomedical documents such as bioBERT outperforms a general domain model such as BERT. We found that the domain model is not clearly superior to the generalist one. We have studied also the number of answers to be extracted per context to obtain consistently good results. Finally, we conclude that although both approaches (readers and re-rankers) are very competitive, readers obtain systematically better results.
Keywords: Question Answering, Information Retrieval, Transformers based pre-trained models, BERT, COVID-19.

Resumen: En este trabajo presentamos una comparación entre las dos arquitecturas neuronales de Respuesta a Preguntas (QA) más utilizadas para resolver el problema de la sobrecarga de información en los artículos relacionados con COVID-19: extracción de respuestas (reader) y el reordenamiento (re-ranker). Hemos encontrado que no hay estudios que comparen estos dos métodos a pesar de que son tan ampliamente utilizados. También realizamos una búsqueda de los mejores hiperparámetros para esta tarea y tratamos de concluir si un modelo pre-entrenado con documentos del dominio biomédico como bioBERT supera a un modelo de dominio general como BERT. Encontramos que el modelo de dominio biomédico no es claramente superior al generalista. También hemos estudiado el número de respuestas a extraer por contexto para obtener resultados consistentemente buenos. Finalmente, concluimos que aunque ambos enfoques (readers y re-rankers) son muy competitivos, los readers obtienen sistemáticamente mejores resultados.
Palabras clave: Búsqueda de Respuestas, Recuperación de Información, Modelos pre-entrenados basados en transformers, BERT, COVID-19.

1 Introduction

Since the COVID-19 outbreak, a huge number of scientific articles have been published making the effective acquisition of new knowledge difficult. There are emerging requests from the medical research community for efficient management of the information about COVID-19 from this huge number of research articles[1]. Therefore, Information Systems are needed to assist biosanitary experts in analyzing these publications.

In this work, we explore full Question Answering (QA) systems, systems that given a question and a document collection, rank all the relevant answers that come from different sources. The collections used in the COVID-19 domain are large enough to require a two-stage pipeline (Chen et al., 2017) that combines an Information Retrieval (IR) step with a neural QA module.

There are two main neural strategies for combining both IR and QA in the state-of-

[1] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

the-art: readers and re-rankers. Both receive a preliminary ranking of contexts given by the IR module. Readers approach scan these contexts looking for the text spans that answer the question. Readers assign a score to each answer, so the final ranking of answers (across different sources) comes from this score or its combination with the IR scores. In the re-rankers approach, the neural model is used to directly re-rank the initial list of paragraphs or sentences given by the initial retrieval.

The advantage of re-rankers is that they provide a final ranking in a sound way. In the other side, they can't go beyond the information retrieved by the IR module. However, readers can scan larger contexts looking for answers with bigger lexical gaps to the question, missed by the classical IR engines.

We have found that, although neuronal reranking is a common method to apply after an ad-hoc information retrieval and prior to a reader, there are no studies that compare these two methods independently and fairly.

Therefore, our goal in this work is two fold:

1. Compare readers and re-rankers to determine their differences in performance, and
2. Determine which is the best configuration for answering questions about COVID-19.

The coupling of IR and QA modules hinders the independent evaluation of the QA approaches (readers vs. re-rankers). To isolate their performance and be able to compare the most popular QA architectures we will use the relevance judgements (qrels) to fix the IR variable to the subset of documents, paragraphs and sentences that contain the actual answers to the test questions.

2 Previous work

Open-domain Question Answering (QA) aims to answer questions by finding answers in a large collection of documents (Voorhees and others, 1999). Early approaches to solve this problem consisted in elaborated systems with pipelined components dealing with question analysis, document retrieval and answer extraction (Brill, Dumais, and Banko, 2002; Ferrucci, 2012). Recent advances of Machine Reading Comprehension (MRC) leaded a two-step pipeline, the *retriever-reader* (Chen et al., 2017).

State-of-the-art models for both architectures (readers vs. re-rankers) are based on pretrained models like BERT (Devlin et al., 2018) which are then finetuned for a specific task.

2.1 Readers

The two-stage pipeline for open-QA was first proposed by (Chen et al., 2017). In this architecture the retriever first extracts a small subset of contexts from a large collection. Then the second component of the pipeline, the reader, scans each context thoughtfully in search for an the answer to the question. (Chen et al., 2017) encode the retrieved contexts and the questions using different Recurrent Neural Networks (RNN). For each question-context pair, two distributions over the contexts tokens are computed using bilinear terms, one for the start of the span and the other for the end. The final answer maximizes the probability of the start and end tokens. With the advent of transformers and pre-trained language models (Devlin et al., 2019) many systems adapted them as their reader (Hao et al., 2022). These systems, although effective at extracting correct answers from a context, process each question-context pair as independent of each other. To improve on this issue (Wang et al., 2019) normalizes the probabilities of the span start and end for all tokens in all contexts whereas (Karpukhin et al., 2020) adds another distribution over the [CLS] token representation of all contexts. Recently some authors proposed generative models with enough parameters to create the answer instead of extracting it (Roberts, Raffel, and Shazeer, 2020). Although competitive in some benchmarks large generative models are expensive to train and make inferences on. To tackle this problem (Izacard and Grave, 2021) combines evidence from the retrieved passages to generate the answer.

2.2 Re-rankers

Other approaches substitutes the reader by a answer re-ranking module where the retrieved passages are divided into plausible sentences and re-ranked by a BERT based cross-encoder (Nogueira and Cho, 2019; Yang, Zhang, and Lin, 2019). In those approaches the neural model is used to rerank

an initial ranking generated by a classical information retrieval model based on term-matching techniques. Specifically, they fine-tune the BERT Large model for the task of binary classification, adding a single layer neural network fed by the [CLS] vector in order to obtain a relevance probability. It has been demonstrated that fine-tuning BERT and treating ranking as a classification problem outperforms existing neural information retrieval models by large margins (Pradeep, Nogueira, and Lin, 2021). A known issue of such neural architectures is that require a large number of query relevances (qrels) for training, but their manual generation is very expensive. Some authors (Nogueira and Cho, 2019; Yang, Zhang, and Lin, 2019) use qrel data oriented to passage retrieval such as MS-Marco (Nguyen et al., 2016) and TREC-CAR (Dietz et al., 2017). Another alternative is to generate relevance judgements automatically. (Dehghani et al., 2017), for example, propose to train neural models for ranking using pseudo-qrels generated by unsupervised models like BM25. The TREC-CAR dataset (Dietz et al., 2017) itself is automatically generated from the structure (article, section and paragraph) of the Wikipedia articles. (MacAvaney, Hui, and Yates, 2017) generate pseudo-qrels from a news collection, using the titles as pseudo-queries and their content as relevant text.

2.3 QA on COVID-19

The model vocabulary and its transfer knowledge capabilities depend on the corpus where it has been pretrained. In the same way general domain models are pretrained using general domain corpus like Wikipedia, we hypothesize that models pretrained with in-domain knowledge such as bioBERT (Lee et al., 2019) should improve the performance of downstream tasks related to biomedical information such as the COVID-19 domain is.

With the rise of the COVID-19 Pandemic the value of open-domain QA systems increased as the academic literature about the virus became unmanageable. Many systems, like Vespa[2], AWS search[3] (Bhatia et al., 2020), Google[4] (Bendersky et al., 2020) or Waterloo[5] (Zhang et al., 2020) arose during the first months of the pandemic. Albeit useful in aiding scientific search of COVID-19 literature they all lacked proper domain evaluation, which is usually perform by comparing the correct span of text with the predicted one using a set metric like *F1* or an *Exact Match* (Rajpurkar, Jia, and Liang, 2018). This evaluation is well suited for short and factoid answers but fails to capture complex responses to diverse information needs within the same question.

The Epidemic Question Answering (EPIC-QA) (Goodwin et al., 2020) was organized to aid in the creation of COVID-19 QA systems. The track evaluates capable of automatically answering ad-hoc questions about the disease COVID-19 by extracting answers from the CORD-19 dataset (Wang et al., 2020), a resource of over 400,000 scholarly articles, including over 150,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. The CORD-19 dataset is regularly updated and represents the most extensive machine-readable coronavirus literature collection.

One way complex question answering scenarios have been evaluated has been through the use of *nuggets*, a set of atomic "facts" that answer the question. Old evaluation scenarios differentiated between "vital" nuggets and "non-vital" nuggets (Dang, Lin, and Kelly, 2008) whereas new evaluation methods consider all nuggets equally relevant and score answers based on how diverse (in terms of number of *nuggets*) their answers are (Goodwin et al., 2020).

3 Models under evaluation

Since we want to compare the reader approach versus the re-ranker approach, we will fix the retrieval variable by using directly the relevant documents with the different correct answers per question that provide the EPIC-QA dataset.

The models we will compare for the reader and the re-ranker are the following ones in the state-of-the-art:

3.1 Reader

The span extraction module is based on pretrained BERT models (Devlin et al., 2018) with two additional parameters vectors for the span start (S) and span end (E), both $S, E \in R^h$ with h being the hidden size of the last layer. The probability for a span to

[2]https://cord19.vespa.ai/
[3]https://cord19.aws/
[4]https://covid19-research-explorer.appspot.com/
[5]https://covidex.ai/

be an answer is computed in two steps:

1. The $soft$ highest logits in the start and end logits vectors are combined to form $soft^2$ plausible answers, scored by the sum of the start token and end token logits.
2. Iterate each of the answers ranked by score to discard non-valid answers (e.g. end token before start token) until the $soft$ answers are valid. Compute the probability of each answer by a softmax over their scores.

After scoring the best $soft$ answers for each document only the qa_cut best are ranked in the final ranking, up to 1000 answers per question.

The number $soft$ of scored answers for each document and the number qa_cut of selected answers for each document are hyperparameters. We experimented with $soft \in \{10, 20, 50, 100, 200, 500\}$ and $qa\_cut \in \{1 : 20\}$.

In our experimentation we consider 2 pretrained BERT models: The original BERT (Devlin et al., 2018) trained with Wikipedia and Book Corpus, a dataset containing +10,000 books of different genres and BioBERT (Lee et al., 2019) trained on large-scale biomedical corpora.

Four different datasets where consider to finetune the models for span extraction:

1. SQuAD2.0 (Rajpurkar, Jia, and Liang, 2018), which is a reading comprehension dataset widely used in the QA research community.
2. QuAC (Choi et al., 2018) a conversational QA dataset containing a higher rate of non-factoid questions than SQuAD.
3. Merge, a combination of SQuAD2.0 and QuAC with the examples shuffled.
4. Seq, a combination of SQuAD2.0 and QuAC where the model is first finetuned with SQuAD2.0 and then with QuAC.

3.2 Re-ranker

The re-ranking module is based on finetuned BERT models on the MSMARCO dataset (Nguyen et al., 2016), a passage ranking dataset which contains one million queries from real users and their respective relevant passages annotated by humans.

The documents are divided into small sentences that are re-ranked using this BERT-based relevance classifier, following a strategy similar to the one proposed by (Nogueira and Cho, 2019).

Then, as with the reader, only the qa_cut best are ranked in the final ranking, up to 1000 answers per question. The number qa_cut of selected answers for each document is $qa\_cut \in \{1 : 20\}$.

We consider two pretrained BERT models: The original BERT (Devlin et al., 2018), and BioBERT (Lee et al., 2019) trained on large-scale biomedical corpora. Both finetuned with MSMARCO for the re-ranking task. The input to the cross-encoder is formed by concatenating the question and sentence into a sequence separated by the [SEP] token. BERT then computed the probability of the sentence being relevant to the query.

4 Evaluation setting

4.1 Dataset

CORD-19 (Wang et al., 2020) is a resource of over 400,000 scholarly articles, including over 150,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. The CORD-19 dataset represents the most extensive machine-readable coronavirus literature collection. It is used extensively for research, including international shared tasks in the IR and QA fields, such as the CORD-19 Challenge at Kaggle[6], TREC-COVID(Roberts et al., 2020) or EPIC-QA[7]

Epidemic Question Answering (EPIC-QA) track aims to develop systems capable of automatically answering ad-hoc questions in English about COVID-19. EPIC-QA involves two tasks, Expert QA and Consumer QA. Experiment in this work are conducted with the data related to the Expert QA task, aimed to answer questions posed by experts.

The questions have three fields: a keyword-based query, a natural language question, and narrative or background. They are evaluated through the use of *nuggets*, a set of atomic "facts" that answer the question. Two datasets were compiled for the task:

[6]https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

[7]https://bionlp.nlm.nih.gov/epic_qa/

The Preliminary Round dataset uses a snapshot of CORD-19 from June 19, 2020, and it includes 45 expert level questions used in the 4th round of the TREC-COVID IR shared task. EPIC-QA Organizers annotated human-generated answers and sentence-level answer annotations (judgements for short) for 21 of those questions as evaluation set in the preliminary round.

The Primary Round dataset is compiled using a snapshot of CORD-19 from October 22, 2020, and it includes 30 expert level questions and their respective relevance judgements.

In this work we have merged the two annotated sets (21 from preliminary round plus 30 from primary round) into one single dataset (epicQA) to gather more evidence in the evaluation results (Table 1).

In order to evaluate the *reader* and *re-ranker* modules in isolation we constructed an *ideal IR*. The organization of the event released some judgements with the correct nuggets for each question and the sentence where they were. By this it is possible to generate an ideal documents level *ideal IR*. We use this *ideal IR* to evaluate our final systems in 5.

Questions	51
Docs	1446
Tokens/Doc	3351
Sentences/Doc	124
Tokens/Sentence	27
Relevant Sentences/Docs	4

Table 1: EPIC-QA dataset statistics.

4.2 Metrics

The evaluation metric, *Normalized Discount Novelty Score* (NDNS), was provided in the EPIC-QA track as a modified version of Normalized Discounted Cumulative Gain. For each answer in a ranking for a question the *Novelty Score* measures the relevant information not yet seen in previous answers of the ranked list.

$$NS(a) = \frac{n_a * (n_a + 1)}{n_a + f_a} \tag{1}$$

where n_a is the number of novel nuggets of answer a and f_a is the *sentence factor*. Three different variants of NDNS are consider based on how this factor is computed:

- **Exact**: Answers should express novel nuggets in as few sentences as possible. This scenario is more suited to evaluate system where brevity is a priority, like a chat bot which can only give one answer.

$$f_a = n_{sentences} \tag{2}$$

- **Relaxed**: Length doesn't penalise answers as long as every sentence contains novel nuggets. This variant of the NDNS metric rewards systems where brevity is not a requirement but non-redundancy is.

$$f_a = n_{non-relevant} + n_{redundant} + 1 \tag{3}$$

- **Partial**: Redundant information is not penalized which makes this metric well suited for systems solving tasks like a state-of-the-art research about a topic where some overlap in the relevant answers is expected.

$$f_a = n_{non-relevant} + 1 \tag{4}$$

The final metric is computed as the cumulative NS of answers up to rank $k = 1000$

$$NDNS(\mathbf{a}) = \frac{1}{NDNS_{ideal}} * \sum_{r=1}^{k} \frac{NS(a_r)}{\log_2(r+1)} \tag{5}$$

where $NDNS_{ideal}$ is the optimal ranking of answers that could have been found in the document collection for the given question, computed using a beam-search with a width of 10 over the annotated sentences.

4.3 Random baseline

For the creation of the baseline we randomly sorted all sentences in the ideal IR documents into groups of 1000 and evaluated them until a convergence score was reached (Table 2).

epicQA	**Baseline**
NDNS-Partial	0.1726
NDNS-Relaxed	0.1736
NDNS-Exact	0.1948

Table 2: Baseline for the three metrics.

5 Experimentation

In order to make a fair comparison between architectures, we first explore the best set

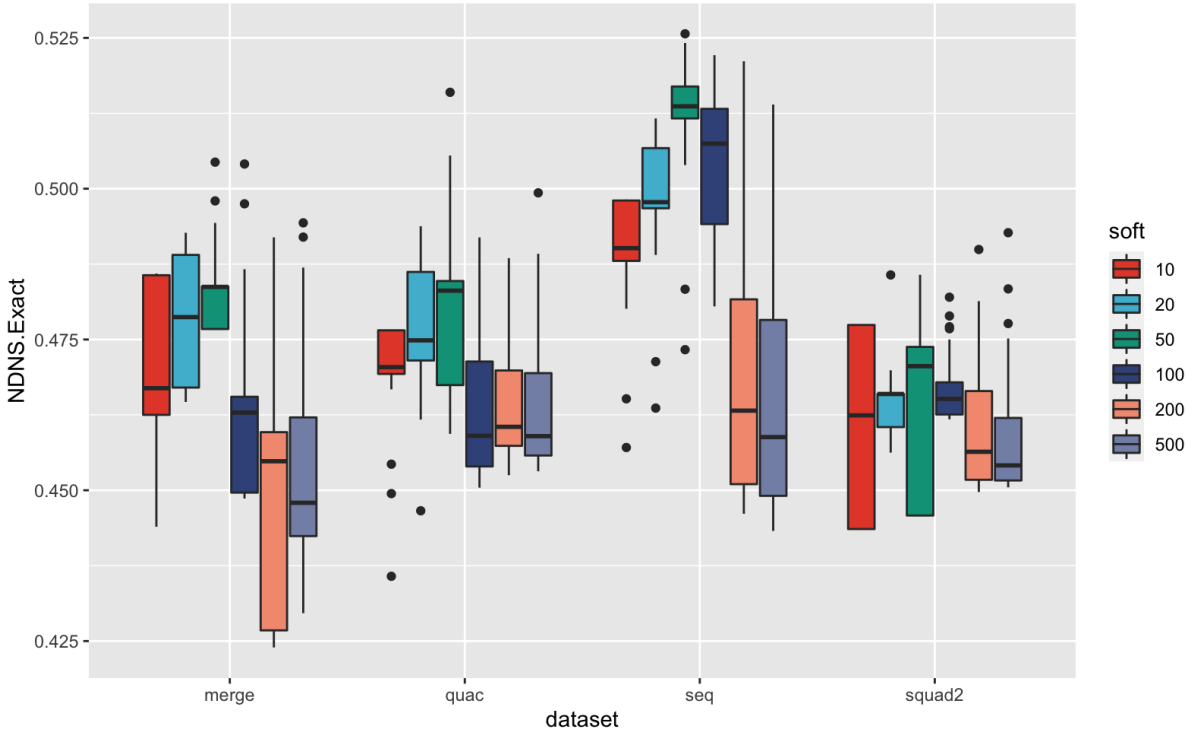


Figure 1: Average NDNS-Exact box-plot across models (BERT and BioBERT) for the reader architecture. Number of considered answers per document (qa_cut 1..20) for each fine-tuning model, breakdown by softmax size. See 3.1 for an explanation of the datasets.

of hyperparameters for each system independently. Those hyperparameters shared by both architectures (the base model and the number of answers consider per document) will be explored jointly. The three different metric scenarios (NDNS-Exact, NDNS-Relaxed and NDNS-artial) have a very strong correlation. We will use mainly NDNS-Exact for comparison since it is the most restrictive and similar to the common metrics used in MRC evaluation.

5.1 Reader

The configuration of the reader depends on two elements. First, the dataset used for tuning the pre-trained model. Second, how to calculate the score of each answer that will determine the final ranking of answers. This score depends on the size of the softmax over the scores of pre-candidate answers for each context.

5.1.1 Dataset used for tuning

The first parameter decision in the reader pipeline is choosing which dataset will be used to finetune the model. We consider four different datasets, detailed in 3.1: SQuAD 2.0, QuAC, random merge of both, and training in sequence (first SQuAD and then QuAC). Figure 1 presents the results breakdown by softmax. Training BERT models with two consecutive datasets is shown to yield the overall best results, even better than just randomly merge both datasets. This result was somehow expected: First, using both dataset gives us more training data. Second, QuAC answers are longer than the kind of factoid-like answers of SQuAD. In this sense, they are closer to the kind of complex answer we need in the COVID-19 domain. So, ending the training with QuAC benefits the model we need.

5.1.2 Softmax

The second step in the reader is to compute the probability of the answers in a document and obtain the scores we need for the final ranking. Given a context and the question, each candidate answer span is first scored by the sum of its start and end token logits. Once all the possible answers in the context are scored, then the probability is computed by a softmax. The size of this softmax, i.e. number of answers per document over which probability is distributed determine the later rank of all answers for a question. The bigger

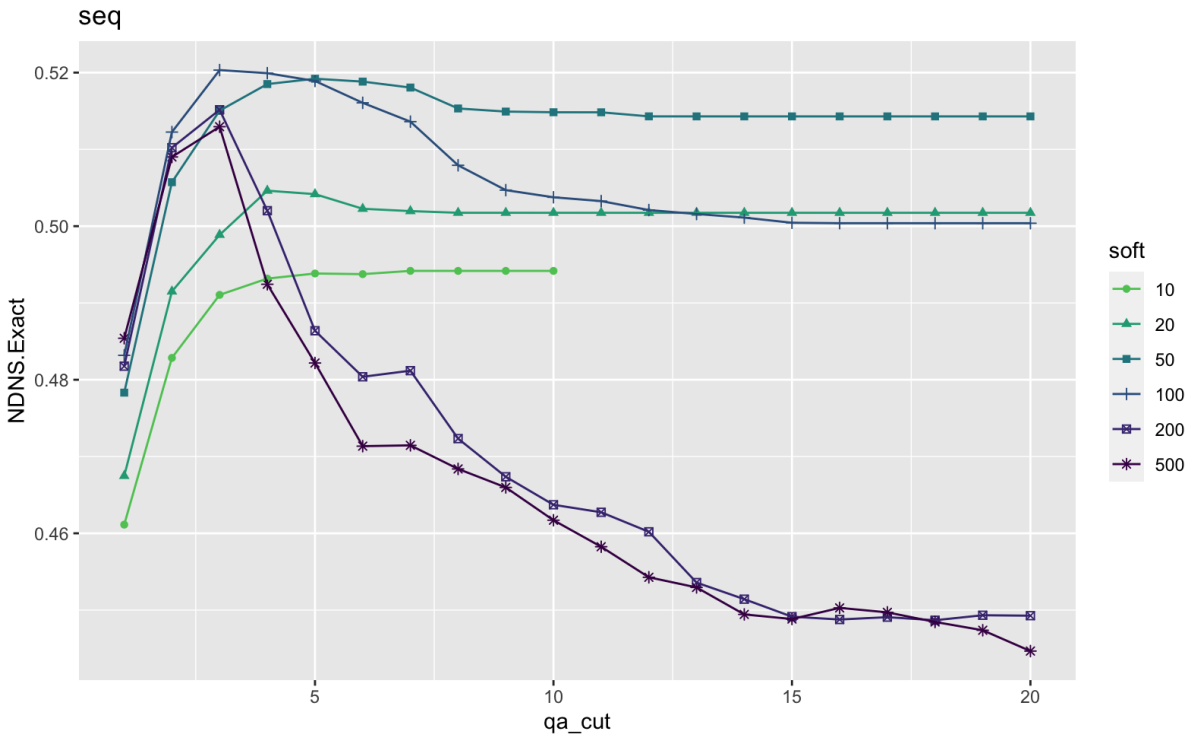


Figure 2: Average NDNS-Exact across models (BERT and BioBERT) for the reader, depending on the number of answers per document (qa_cut), breakdown by softmax size.

the size of the softmax the smaller the probability of each individual answer in absolute terms. However answers in which the model has a high confidence may stand out more from the rest. Results are ploted in Figure 2 with each softmax having a separate curve for all levels of qa_cut, the number of answers selected from each document to be ranked. All softmax sizes follow a similar trend at the beginning of the curve with bigger sizes (> 20) peaking all above 0.51 NDNS-Exact. Interestingly the highest sizes experience a step drop down in their scores. The rest of the sizes experience a small decrease followed by a flat convergence. Overall most softmax sizes perform better at qa_cuts smaller than 10. In this range both softmax sizes of 50 and 100 had the highest scores among all sizes. For this reason we have used them in the rest of the experimentation.

5.2 Re-ranker

The only hyperparameter to explore in the re-ranking is the number of responses per document (qa_cut). In Figure 3 we can see as scores improve drastically when taking between 2 and 10 responses per document, reaching the top in 8, and from 10 responses per document the quality drop decreases in a linear way. Between a qa_cut of 2 and 10, scores above 0.44 are consistently obtained.

5.3 Reader vs Ranker

Finally we compare both architectures filtering qa_values above 10 as both methods results worsen after. Results are plotted in 4. Both methods beat the random baseline 2 by a large margin of more than 25 points, proving its effectiveness.

Results show that the reader approach constantly outperforms the re-ranker one, even for its lowest score with one answer for document. We observe also that reader scores have a smaller variance over the range of qa_cut whereas the re-ranker is surprisingly bad with only one response per document. Another interesting observation is that these results are robust to the use of different pretrained models (BERT and BioBERT), and to the softmax size of the reader.

Contrary to what it might be expected, the domain model bioBERT does not outperform the generalist model BERT, specially in the case of the reader approach. This result rises questions on whether the QA task on COVID-19 benefits from domain-trained networks or if generalists are sufficient.

In the case of the re-ranking method the

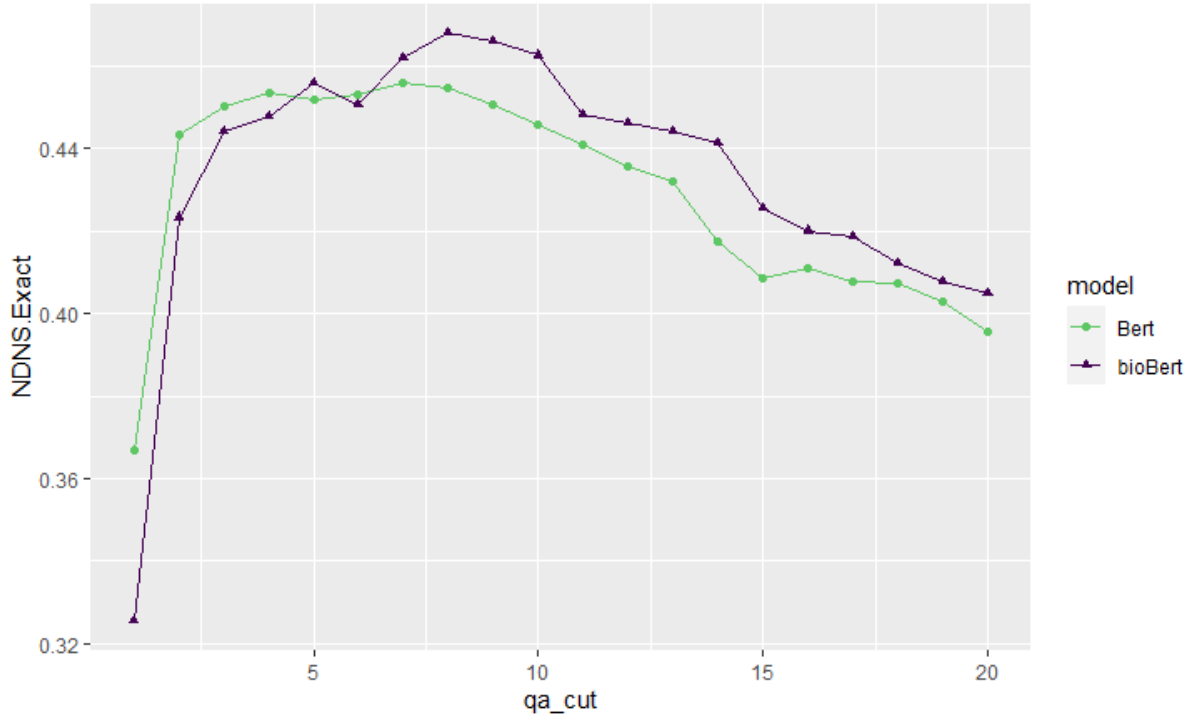


Figure 3: NDNS-Exact results for the re-ranker by number of answers per document (qa_cut) up to 20.

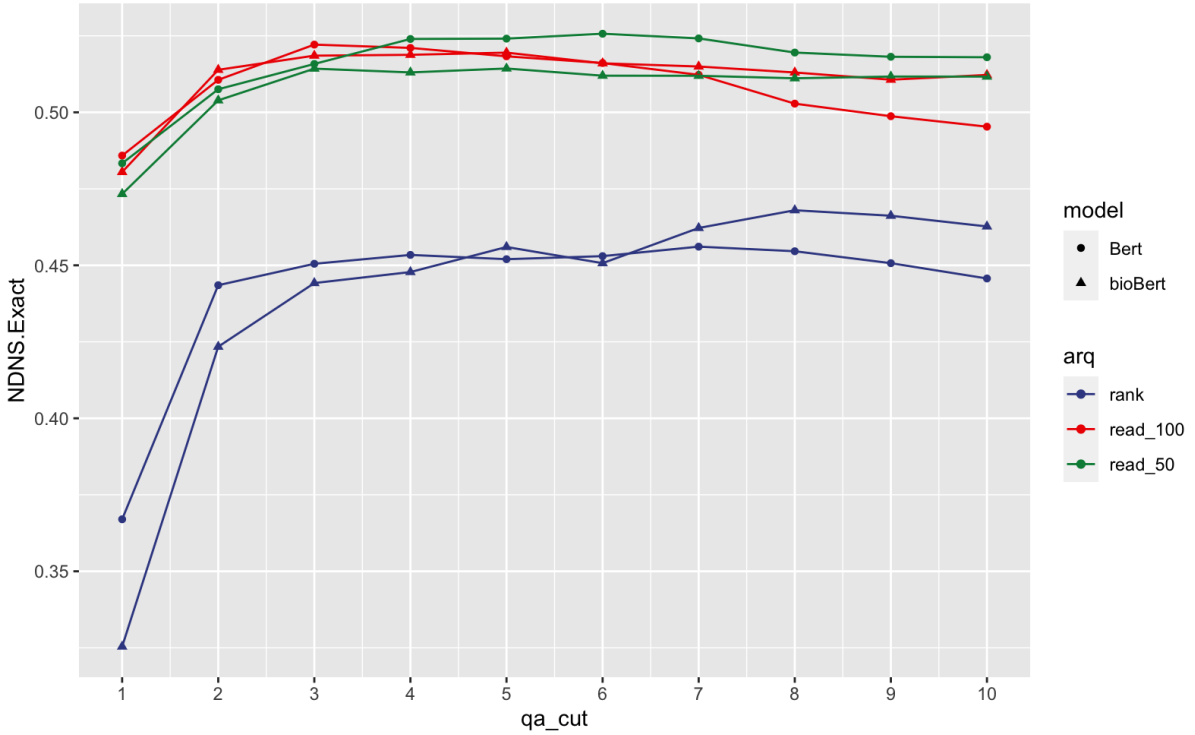


Figure 4: NDNS-Exact results for the Reader vs the Re-ranker by number of answers per document (qa_cut) up to 10.

domain model does outperform the generalist. So we can not come to a global conclusion, but it would be worthwhile to investigate further in this direction.

6 Conclusions and future work

In this work we compare two of the most popular neural QA architectures, retriever-reader and retriever-reranker. We have tested them in the domain-specific scenario of COVID-19.

Although both approaches have shown competitive results, the reader has proven to yield better results than the re-ranker.

Both architectures rely on a previous retrieval step and both return a ranking of sentences answering a given question. However, the retriever-reader approach allows the retrieval of broader contexts than a single sentence and then scan that context looking for the best match. In this case, the final selected sentence may not contain all the exact terms used for the retrieval step, but other related terms according to the language models behind. Therefore, it seems more robust to the initial keyword base retrieval step.

We also concluded that regardless of the method to be used it is always better to take several responses per document, specially in open-domain QA. We conclude that a good range is between 3 and 10 responses per document.

Both domain and generalist models have obtained similar results. We believe that there is an overestimation of the capabilities of domain models and it would be interesting to continue the research in this direction.

As future work, we plan to extend this work to other BERT models and new datasets.

Acknowledgments

This work has been partially funded by VIGICOVID project[8] FSuperaCovid-5 (Fondo Supera COVID-19/CRUE-CSIC-Santander) and by the Spanish Ministry of Science, Innovation and Universities (DeepReading RTI2018-096846-B-C21, MCIU/ AEI/FEDER, UE).

References

Bendersky, M., H. Zhuang, J. Ma, S. Han, K. Hall, and R. McDonald. 2020. RRF102: Meeting the TREC-COVID challenge with a 100+ runs ensemble. *arXiv preprint arXiv:2010.00200*.

Bhatia, P., L. Liu, K. Arumae, N. Pourdamghani, S. Deshpande, B. Snively, M. Mona, C. Wise, G. Price, S. Ramaswamy, X. Ma, R. Nallapati, Z. Huang, B. Xiang, and T. Kass-Hout. 2020. AWS CORD-19 Search: A Neural Search Engine for COVID-19 Literature.

Brill, E., S. Dumais, and M. Banko. 2002. An analysis of the AskMSR question-answering system. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 257–264.

Chen, D., A. Fisch, J. Weston, and A. Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Choi, E., H. He, M. Iyyer, M. Yatskar, W.-t. Yih, Y. Choi, P. Liang, and L. Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184.

Dang, H., J. Lin, and D. Kelly. 2008. Overview of the TREC 2006 Question Answering Track, 2008-11-05.

Dehghani, M., H. Zamani, A. Severyn, J. Kamps, and W. B. Croft. 2017. Neural ranking models with weak supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 65–74.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

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.

Dietz, L., M. Verma, F. Radlinski, and N. Craswell. 2017. TREC Complex Answer Retrieval Overview. In *TREC*.

[8] http://nlp.uned.es/vigicovid-project

Ferrucci, D. A. 2012. Introduction to "This is Watson". *IBM Journal of Research and Development*, 56(3.4):1–1.

Goodwin, T. R., D. Demner-Fushman, K. Lo, L. L. Wang, W. R. Hersh, H. T. Dang, and I. M. Soboroff. 2020. Overview of the 2020 Epidemic Question Answering Track. Technical report, Text Analysis Conference (TAC) 2020.

Hao, T., X. Li, Y. He, F. L. Wang, and Y. Qu. 2022. Recent progress in leveraging deep learning methods for question answering. *Neural Computing and Applications*, 34(4):2765–2783.

Izacard, G. and E. Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online, April. Association for Computational Linguistics.

Karpukhin, V., B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Lee, J., W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 09.

MacAvaney, S., K. Hui, and A. Yates. 2017. An approach for weakly-supervised deep information retrieval. *arXiv preprint arXiv:1707.00189*.

Nguyen, T., M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng. 2016. MS MARCO: A human-generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Nogueira, R. and K. Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

Pradeep, R., R. Nogueira, and J. Lin. 2021. The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models. *arXiv preprint arXiv:2101.05667*.

Rajpurkar, P., R. Jia, and P. Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Roberts, A., C. Raffel, and N. Shazeer. 2020. How Much Knowledge Can You Pack into the Parameters of a Language Model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426.

Roberts, K., T. Alam, S. Bedrick, D. Demner-Fushman, K. Lo, I. Soboroff, E. Voorhees, L. L. Wang, and W. R. Hersh. 2020. TREC-COVID: rationale and structure of an information retrieval shared task for COVID-19. *Journal of the American Medical Informatics Association*, 27(9):1431–1436, 07.

Voorhees, E. M. et al. 1999. The TREC-8 question answering track report. In *Trec*, volume 99, pages 77–82. Citeseer.

Wang, L. L., K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. M. Kinney, et al. 2020. CORD-19: The COVID-19 Open Research Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Wang, Z., P. Ng, X. Ma, R. Nallapati, and B. Xiang. 2019. Multi-passage BERT: A Globally Normalized BERT Model for Open-domain Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5881–5885.

Yang, W., H. Zhang, and J. Lin. 2019. Simple applications of BERT for ad hoc document retrieval. *arXiv preprint arXiv:1903.10972*.

Zhang, E., N. Gupta, R. Tang, X. Han, R. Pradeep, K. Lu, Y. Zhang, R. Nogueira, K. Cho, H. Fang, et al. 2020. Covidex: Neural ranking models and keyword search infrastructure for the covid-19 open research dataset. *arXiv preprint arXiv:2007.07846*.

Tesis

Procesamiento del Lenguaje Natural, Revista nº 68, marzo de 2022, pp. 145-148 recibido 25-11-2021 revisado 25-01-2022 aceptado 30-01-2022

Dependency Syntax in the Automatic Detection of Irony and Stance

Sintaxis de dependencias en la detección automática de ironía y posicionamiento

Alessandra Teresa Cignarella[1,2]
[1]PRHLT Research Center, Universitat Politècnica de València, Spain
[2]Dipartimento di Informatica, Università degli Studi di Torino, Italy
alessandrateresa.cignarella@unito.it

Abstract: PhD thesis in Computer Science written by Alessandra Teresa Cignarella under the supervision of Dr. Cristina Bosco (University of Turin) and Prof. Dr. Paolo Rosso (Universitat Politècnica de València). This thesis was developed under a cotutelle between the PRHLT Research Center of the Universitat Politècnica de València, Spain and the Computer Science Department of the University of Turin, Italy. The thesis defense was held in Torino (*online*), Italy on October 26th, 2021. The doctoral committee was composed by: Prof. Dr. Joakim Nivre (Department of Linguistics and Philology, Uppsala University, Sweden), Dr. Saif M. Mohammad (National Research Council Canada, Ottawa, Canada), and Prof. Dr. Veronique Hoste (Department of Translation, Interpreting and Communication, Ghent, Belgium). An international mention was achieved after having spent 18 months at the Università degli Studi di Torino and 18 months at the Universitat Politècnica de València.
Keywords: Universal Dependencies, Irony Detection, Stance Detection, NLP, Multilingual, User-Generated Content, Twitter, BERT.

Resumen: Tesis doctoral en Informática realizada por Alessandra Teresa Cignarella y dirigida por la Dra. Cristina Bosco (University of Turin) y el Prof. Dr. Paolo Rosso (Universitat Politècnica de València) en el marco de un convenio de cotutela entre el PRHLT Research Center de la Universitat Politècnica de València, España y el Departamento de Informatica de la Universidad de Turin, Italia. La defensa de la tesis fue en Turin (*en linea*), Italia el 26 de octubre de 2021 ante un tribunal compuesto por: Prof. Dr. Joakim Nivre (Department of Linguistics and Philology, Uppsala University, Sweden), Dr. Saif M. Mohammad (National Research Council Canada, Ottawa, Canada), and Profa. Dra. Veronique Hoste (Department of Translation, Interpreting and Communication, Ghent, Belgium). Se obtuvo la mención internacional tras una estancia de 18 meses en la Università degli Studi di Torino y 18 meses en la Universitat Politècnica de València.
Palabras clave: Dependencias universales, detección de ironía, detección de posicionamiento, PNL, multilingüe, contenido generado por el usuario, Twitter, BERT.

1 Introduction

The present dissertation is part of the broad panorama of studies of Natural Language Processing (NLP). In particular, it is a work of Computational Linguistics (CL) designed to study in depth the contribution of syntax in the field of sentiment analysis and, therefore, to study texts extracted from social media or, more generally, online content.

Furthermore, given the recent interest of the scientific community in the Universal Dependencies[1] (UD) project (De Marneffe et al., 2021), which proposes an annotation format aimed at creating a "universal" representation of the phenomena of morphology and syntax in a manifold of languages, in this work we made use of this format, thinking of a study in a multilingual perspective (Ita-

[1]https://universaldependencies.org/.

lian, English, French and Spanish). Although the UD format was originally conceived to be applied to texts that are more "standard" from the point of view of morphosyntactic norms and punctuation, in more recent years the same scheme has begun to be applied also to user-generated content (UGC), i.e. texts extracted from social media, blogs, forums and microblogging platforms, such as Reddit, Twitter or Wikipedia pages. Inevitably, the application of this annotation framework to such a peculiar textual genre, in which the texts are accompanied by multimedia elements such as links, photos and videos, emojis and non-standardized punctuation, has opened up several problems in the Universal Dependencies community, many of which are still the subject of open and heated debate today.

In this work we provide an exhaustive presentation of the morphosyntactic annotation format of UD, in particular underlining the most relevant issues regarding their application to UGC. Two sub-areas of NLP will be presented, and used as case studies, in order to test the research hypotheses: the first case study will be in the field of *Irony Detection* (Van Hee, Lefever, and Hoste, 2018) and the second in the area of *Stance Detection* (Mohammad et al., 2016). In both cases, historical notes are provided that can serve as a context for the reader, the problems faced are introduced and the activities proposed in the computational linguistics community are described. Furthermore, particular attention is paid to the resources currently available as well as to those developed specifically for the study of the aforementioned phenomena. Finally,through the discussion of a set of experiments performed within or outside evaluation campaigns, we describe how syntax can contribute to the resolution of such tasks.

1.1 Motivation and Objectives

My main purpose is to explore the impact of morpho-syntactic information in sentiment analysis related tasks.

Firstly, for both irony and stance, I focused on the importance of the formulation of a clear problem statement, and the subsequent computational modeling of it. Secondly, I highlighted my experience in the creation of annotated corpora for those problems, my contribution to the organization of shared tasks and the important lessons learned, in terms of research understanding. Later on, I proposed my first approaches to solve both tasks from a shallow perspective, starting to explore the most feasible way to represent morphosyntactic information, to extract it, and exploit it for classification purposes. I ended this process by relying on the UD Dependencies annotation format. Finally, after having encountered a satisfactory combination of features, I exploited the best sets of them – some of which are encoded in UD format – and I performed a handful of experiments in a variety of settings.

In the whole thesis a multilingual scenario is kept in mind, exploring four different language settings: English, Spanish, French, and Italian, for both irony detection and stance. Furthermore, due to the availability of benchmark datasets, regarding stance detection, I also experiment on a fifth language, i.e., Catalan.

The research questions that I aimed answering to are as follows:

RQ-1: *Could features derived from morphology and syntax help to address the task of irony detection?*

RQ-2: *To what extent does using resources such as treebanks for training NLP models improve the performance in irony detection?*

RQ-3: *Could features derived from morphology and syntax help to address the task of stance detection?*

RQ-4: *To what extent does using resources such as treebanks for training NLP models improve the performance in stance detection?*

2 Thesis Overview

This thesis consists in a reorganized collection of the most relevant investigations extracted from some research projects in which I was involved during my Ph.D. studies.

A brief overview of the contents of the thesis is presented below, summarizing all the work done and resuming the results obtained in the framework of this three-year-long research path. In Chapter 3, I also show some unpublished results regarding stance detection with dependency syntax and neural networks. Lastly, I draw some conclusions and discuss future work in the final chapter.

Chapter 1 – Introduction: In this chapter I introduced the reader to the main topics that will be discussed in the thesis, starting with a broad description of Natural Language Processing and automatic text classification, followed by an introduction on Universal Dependencies, morphology and syntax. I also proposed a brief discussion on the issues that can arise while applying the UD format to social media data, mainly referring to the following work: Sanguinetti et al. (2022).

Chapter 2 – Irony Detection: In the second chapter, which deals with the topic of irony detection, I described several works regarding such topic. In particular, in Section 2.1.1, I mainly referred to Cignarella et al. (2018) in order to describe described the organization of the *IronITA 2018* shared task. In Section 2.1.2, I described the creation of the multilayered corpus TWITTIRÒ-UD, annotated both of irony and morphology and dependency syntax.

In Section 2.3, I finally described some experiments performed in which I leveraged morpho-syntactic information for irony detection, mainly referring to what was done in the participation of the IroSvA 2019 shared task (Cignarella and Bosco, 2019; Cignarella et al., 2020).

Chapter 3 – Stance Detection: In the third chapter I dealt with the task of stance detection. In Section 3.1.1, I described the organization of the *SardiStance* shared task at EVALITA 2020. In Section 3.2.1, I presented the work done in Lai, Cignarella, and Hernandez Fariás (2017), for describing the participation in the *StanceCat 2017* shared task at IberEval 2017. In Section 3.3, I described my participation in *RumorEval 2019* where I first applied a syntax-based approach to the task of stance detection.

In Section 3.3.2, I presented a completely new research, specifically done for the PhD thesis, where I introduced a BERT-based approach leveraging morphosyntactic information for the automatic detection of stance in different languages.

Chapter 4 – The Interaction of Irony and Stance: In the fourth chapter I proposed a new part of my research, in which I explored the interaction between irony and stance, through the analysis of the *SardiStance* dataset, which has been annotated accordingly to both phenomena.

Chapter 5 – Conclusions and Future Work – In the last chapter, I finally summarized all the important lessons learned and I proposed new research directions for future work.

3 Conclusions

This thesis collocates within the growing trend of studies devoted to make Artificial Intelligence results more explainable, going beyond the achievement of highest scores in performing tasks, but rather making their motivations understandable and comprehensible for experts in the domain.

The novel contribution of this work mainly consists in the exploitation of features that are based on *morphology* and *dependency syntax*, which were used in order to create vectorial representations of social media texts in various languages and for two different tasks. Such features have then been paired with a manifold of machine learning classifiers, with some neural networks and also with the language model BERT.

Results suggest that fine-grained dependency-based syntactic information is more informative for the detection of irony, and less informative for what concerns stance detection. Nonetheless, dependency syntax might still prove useful in the task of stance detection if, firstly, irony detection is considered as a pre-processing step. I also believe that the approach based on dependency syntax that I proposed could help in understanding and explaining a such a complex phenomenon like irony.

In fact, the several studies presented here allowed to investigate whether syntactic structures, independently from the target language, may provide information useful to understand whether a message is ironic or not.

Although it has been duly noted that syntax does not seem to be particularly informative regarding directly the task of stance detection (the second case study, presented in Chapter 3). On the other hand, also supported by some previous linguistic studies, syntax seems to play an important role in the detection of irony. Therefore, a new speculation that comes to mind, is that it could be more useful to perform a *"cascade task"*. Meaning that, firstly it might be useful to predict irony, with the help of morphosyntactic cues (step 1), and only then (as step 2),

proceeding in the detection of stance. In general, my assumption, is that predicting irony could be the first step in numerous other tasks, even shallow sentiment analysis, or the identification of fake news.

This outcome is something that should not be ignored, but obviously carrying out supervised studies in this sense would also mean dedicating a great effort and consuming much time in the creation of annotated datasets (that ought to be annotated on various layers, for different dimensions and phenomena). In fact, to further study this line of investigation, in Chapter 4, I proposed a shallow analysis of the Italian dataset regarding the Sardines Movement, which is only a small and limited beginning, but it is also certainly opening a new research perspective.

My work has certainly many limitations. Firstly, I needed to deal with the scarcity of data annotated in some adequate way and with the reduced size of the few datasets that are indeed available or those I helped to develop. Furthermore, this kind of investigation, mostly based on morphosyntactic cues that are applied to NLP tasks, is a rather new one. In fact, there are very few studies going towards this direction up to these days.

By having looked at some of the results obtained in the wide variety of experiments performed in this thesis, it is fundamental to stress that we did not solely want to appreciate the outcomes in terms of numerical performances, but rather being more focused in the more profound linguistic reasons behind them. And the same is valid also for why sometimes results are poorer and why features do not make improvements on a certain task.

I am positive that if we manage to understand what is the linguistic knowledge a certain approach, or a group of features, leverages when it produces good (or poor) results, among many possible approaches, it could allow us to make more mature choices for following work. Indeed, the future of NLP research needs to go towards approaches that better integrate different types of knowledge (such as syntactic knowledge, for once) and that manage to be more versatile for certain types of data and in different application contexts.

Acknowledgments

This work has been funded by the scholarship "Be Positive!" (under the 2019 "Google.org Impact Challenge on Safety" call) and also supported by the European project "STEREOTYPES - STudying European Racial Hoaxes and sterEOTYPES" funded by the Compagnia di San Paolo and VolksWagen Stiftung under the "Challenges for Europe" call for Project (CUP: B99C20000640007).

References

Cignarella, A. T., V. Basile, M. Sanguinetti, C. Bosco, F. Benamara, and P. Rosso. 2020. Multilingual Irony Detection with Dependency Syntax and Neural Models. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*. ACL.

Cignarella, A. T. and C. Bosco. 2019. ATC at IroSvA 2019: Shallow Syntactic Dependency-based Features for Irony Detection in Spanish Variants. In *Proceedings of IberLEF 2019*. CEUR-WS.org.

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 *Proceedings of EVALITA 2018*. CEUR-WS.org.

De Marneffe, M.-C., C. D. Manning, J. Nivre, and D. Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Lai, M., A. T. Cignarella, and D. I. Hernandez Fariás. 2017. iTACOS at IberEval2017: Detecting Stance in Catalan and Spanish Tweets. In *Proceedings of IberEval 2017*. CEUR-WS.org.

Mohammad, S., S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of SemEval 2016*. ACL.

Sanguinetti, M., C. Bosco, L. Cassidy, Ö. Çetinoğlu, A. T. Cignarella, T. Lynn, I. Rehbein, J. Ruppenhofer, D. Seddah, and A. Zeldes. 2022. Treebanking User-generated Content: A UD based Overview of Guidelines, Corpora and Unified Recommendations. *Language Resources and Evaluation*, pages 1–52.

Van Hee, C., E. Lefever, and V. Hoste. 2018. SemEval-2018 Task 3: Irony Detection in English Tweets. In *In Proceedings of SemEval 2018*. ACL.

Procesamiento del Lenguaje Natural, Revista nº 68, marzo de 2022, pp. 149-153 recibido 18-12-2021 revisado 22-01-2022 aceptado 29-01-2022

Biomedical entities recognition in Spanish combining word embeddings

Reconocimiento de entidades biomédicas en español combinando word embedding

Pilar López-Úbeda
SINAI, Department of Computer Science, CEATIC, Universidad de Jaén,
Campus Las Lagunillas s/n, 23071, Jaén (Spain)
plubeda@ujaen.es

Abstract: This is a summary of the Ph.D. thesis written by Pilar López Úbeda at Universidad de Jaén under the supervision of PhD. M. Teresa Martín Valdivia, Ph.D. L. Alfonso Ureña López and PhD. Manuel Carlos Díaz Galiano. The defense was held in Jaén on April 22, 2021. The doctoral committee was integrated by PhD. Rafael Muñoz Guillena from Universidad de Alicante, PhD. Paloma Martínez Fernández from Universidad Carlos III de Madrid, and Manuel Montes y Gómez from National Institute of Astrophysics, Optics and Electronics (Mexico). The thesis obtained the grade of *Summa Cum Laude* and the international mention.
Keywords: Natural language processing, Spanish corpora, biomedical entity recognition, word embeddings, deep learning.

Resumen: Este es un resumen de la tesis doctoral realizada por Pilar López Úbeda en la Universidad de Jaén bajo la dirección de los doctores Dña. M. Teresa Martín Valdivia, D. L. Alfonso Ureña López y D. Manuel Carlos Díaz Galiano. La defensa se realizó en Jaén el 22 de abril de 2021. La comisión de doctorado estuvo integrada por el PhD. Rafael Muñoz Guillena de la Universidad de Alicante, la PhD. Paloma Martínez Fernández de la Universidad Carlos III de Madrid, y Manuel Montes y Gómez del Instituto Nacional de Astrofísica, Óptica y Electrónica (México). La tesis obtuvo la calificación de Sobresaliente Cum Laude y mención de doctorado internacional.
Palabras clave: Procesamiento del lenguaje natural, corpus en español, reconocimiento de entidades biomédicas, representación de palabras, aprendizaje profundo.

1 Introduction

One of the main purposes of clinical text mining is the possibility to process and analyze the large volumes of textual information contained in medical records. Through this treatment of the information, we attempt to answer questions such as, which patients presented a certain condition? What kind of conditions were used to detect the disease? What were the results of the tests performed? What was the treatment given? These questions could seem quite simple for some medical professionals, but they become extremely complex when managed automatically by computational systems.

In the biomedical domain, we can find large collections of free textual information (medical reports, Electronic Health Records - EHR, scientific papers, among others) that contain very relevant data that need to be studied in depth. However, current health information systems are not prepared to analyze and extract this knowledge due to the time and cost involved in processing it manually. The field of artificial intelligence known as Natural Language Processing (NLP) is being applied to medical documents to build applications that can understand and analyze this huge amount of textual information automatically (Friedman y Johnson, 2006)

Many researchers in the NLP field focus on the area of Information Extraction (IE) in the biomedical domain to address these challenges. IE systems take natural language text as input and produce structured infor-

mation specified by certain criteria and that is relevant to a particular application. Depending on the different inputs of IE systems and expected outputs, many sub-tasks can be defined such as Named Entity Recognition (NER).

In this thesis, we focus on information extraction from Spanish biomedical texts, more specifically, on the NER task. Spanish has more than 480 million native speakers and nowadays there is a worldwide interest in processing medical texts in this language. With this study, we aim to advance the task of biomedical NER in this relevant language and thus answer the above-mentioned questions (López Úbeda, 2021).

To accomplish this study, we propose a methodology based on deep learning. Furthermore, different word embeddings are used in combination to obtain a better representation of each word. With this approach, we aim to achieve the desired final goal: to recognize biomedical entities accurately in different scenarios.

1.1 Motivation

Over the years, the recognition of biomedical entities has motivated the scientific community to continue developing automatic systems to facilitate the extraction of medical knowledge. NER is a difficult task to solve that can help in many other medical-related systems such as those presented below:

- **Clinical decision support**. Automated NER systems can provide real-time results, which means that entities such as diseases can be detected immediately. This evidence can be used to help professionals identify emerging health problems, for instance, to alert them to the presence of certain unexpected findings (López-Úbeda et al., 2020b).
- **Entity representation**. In the NER task, different words can have similar meanings. This problem is caused by the multiple ways in which a particular entity can be represented and written. For instance, "*adriamicina*"(adriamycin) and "*doxorrubicina*"(doxorubicin) refer to the same drug widely used in cancer chemotherapy.

  On the other hand, an acronym does not always have a unique description, it can be interpreted as two different entities depending on the context. For instance, in Spanish, PCR can be referred to "*parada cardiorrespiratoria*"(cardiorespiratory arrest) or "*Reacción en Cadena de la Polimerasa*"(Polymerase Chain Reaction). Finally, as we see in the examples, biological entities may also have multi-word names, so the problem is additionally complicated by the need to determine name boundaries and resolve overlap of candidate names.
- **Basis for other NLP tasks**. Biomedical entity recognition serves as the basis for many other crucial areas of information management, such as classification tasks, question answering, information retrieval, and text summarization (López-Úbeda et al., 2020a). For instance, the use of NER becomes important for analyzing the clinical text and obtaining the most relevant tags in each report, allowing the classification of documents.
- **Extracting structured information**. Biomedical NER is a task that facilitates medical professionals in structuring reports contributing to solutions such as providing a summary of patient conditions or serving as a tool to organize the documentation of the physician's decision-making process, plan development, and patient outcomes.

1.2 Objectives

The main objective of this thesis focuses on the study, analysis, and development of NLP techniques and tools for the NER task in the biomedical domain in Spanish. Specifically, it focuses on the study and applicability of different combinations of word embeddings as word representations.

This general objective has been defined through the following specific objectives:

- Collect resources available in Spanish annotated with biomedical entities used in different challenges.
- Study and select the existing word embeddings in Spanish serving as input to the network.
- Propose a deep learning-based method for NER in the biomedical domain that

can take a combination of different word embeddings as input.

- Generate a new word embedding for Spanish focused on the biomedical domain to see how effective it is in comparison to existing ones.
- Evaluate the performance of the proposed method on the NER problem using three application scenarios: pharmacological domain, oncological domain, and knowledge discovery in biomedical texts.
- Conduct a results analysis comparing our system with the state-of-the-art.
- Perform an error analysis to understand the capabilities and drawbacks of our system.
- Identify open issues from the conclusions in order to propose future research.

1.3 Hypotheses

In this thesis, we address the problem of biomedical entity extraction in Spanish through deep learning and combinations of word embeddings using NLP methods. Based on the objectives set out above, our general hypothesis can be summarized as follows:

NLP techniques applied to the NER task can improve biomedical systems.

However, since this hypothesis is very ambitious, we have decided to subdivide it into three specific hypotheses:

Hypothesis 1 *(H1). Deep neural networks in NLP leverage the advantage of existing relevant information from the Spanish biomedical textual data and the NER task, outperforming models that do not integrate this information properly.*

Hypothesis 2 *(H2). Combining different types of word embeddings by concatenating each embedding vector to form the final word vectors is an important part of the biomedical entity recognition task. The probability of recognizing a specific entity in a text should increase as optimal representations of that word are combined because they are more comprehensively represented and integrate relevant knowledge.*

Hypothesis 3 *(H3). Integrating domain-specific knowledge into the training corpus can be beneficial for improving the quality of word embeddings. Thus, this resource provides a more accurate representation of words in a particular context and domain.*

2 Thesis outline

This thesis is organized into six chapters and an appendix as described below:

Chapter 1 contains an introduction explaining the motivation and objectives that led us to carry out the study. Furthermore, we have presented the hypotheses with the research questions we intend to solve and the methodology we will carry out.

Chapter 2 presents an overview of the methodologies based on ML commonly used in the NER task and which are necessary to understand the later parts of this thesis.

Chapter 3 summarizes previous work on NLP tasks based on ML in the biomedical domain and shows an extensive literature review of the NER task with regard to present state-of-the-art studies. Since the interest of this thesis lies in word representation, this chapter details the review of existing methods for word representations up to the moment.

Chapter 4 describes the proposed model to solve the biomedical entity extraction problem. After an extensive review of previously applied methodologies, we propose an approach based on a Bidirectional Long Short-Term Memory (BiLSTM) neural network with a final CRF layer.

Chapter 5 presents the experimentation carried out using the approach proposed. The experimental framework was developed in three scenarios belonging to different biomedical sub-domains including pharmacology, oncology, and knowledge discovery. For each scenario, this chapter contains a description of the problem, the dataset, the results obtained, error analysis, and a discussion.

Chapter 6 contains our conclusion where we summarize our findings and main contributions. Moreover, this chapter provides an outlook into the future, the publications derived from the study, and the research results transferred.

Finally, **Appendix A** contains additional results of the NER task performance in the different scenarios proposed.

3 Main contributions

This research has carried out a series of studies, analyses, and development of NLP tech-

niques designed to address the task of NER in Spanish biomedical texts. This has resulted in several contributions to the research that we have considered on the basis of the hypotheses.

To support hypothesis H1, we can summarize the following contributions:

Contribution 1 We have investigated and implemented different machine learning approaches. First, we have reviewed unsupervised models and then advanced to supervised models using traditional models such as CRF and deep neural networks.

Contribution 2 In our review of the state-of-the-art in deep learning, we have exposed what kind of architectures are used by the scientific community interested in NER.

Contribution 3 We have proposed a model based on neural networks. Specifically, the architecture is composed of a BiLSTM network and a CRF layer (López-Úbedaa et al., 2020).

To support hypothesis H2, we provide the following contributions:

Contribution 4 In our review of related literature, we have found that word representations and, more specifically, word embeddings are the most commonly used methods.

Contribution 5 We have selected different word embeddings to include in the neural network to address the NER problem in biomedicine.

Contribution 6 We have presented a model based on a combination of word embeddings for a more exhaustive representation of the words, thus improving entity identification systems.

The contributions that support hypothesis H3 can be summarized as follows:

Contribution 7 We have collected an unannotated corpus by extracting documents from different corpora and websites related to the biomedical domain, obtaining a vocabulary of 1,704,151 words.

Contribution 8 We have generated new word embeddings specifically for the biomedical domain in Spanish (López-Úbeda et al., 2020c).

Acknowledgements

This work has been partially supported by a grant from Fondo Europeo de Desarrollo Regional (FEDER), LIVING-LANG project [RTI2018-094653-B-C21], and the Government of Andalusia [PY20_00956].

Bibliografía

Friedman, C. y S. B. Johnson. 2006. Natural language and text processing in biomedicine. En *Biomedical Informatics*. Springer, páginas 312–343.

López Úbeda, P. 2021. Biomedical entities recognition in spanish combining word embeddings.

López-Úbeda, P., M. C. Díaz-Galiano, T. Martín-Noguerol, A. Luna, L. A. Ureña-López, y M. T. Martín-Valdivia. 2020a. Covid-19 detection in radiological text reports integrating entity recognition. *Computers in Biology and Medicine*, 127:104066.

López-Úbeda, P., M. C. Díaz-Galiano, T. Martín-Noguerol, A. Ureña-López, M.-T. Martín-Valdivia, y A. Luna. 2020b. Detection of unexpected findings in radiology reports: A comparative study of machine learning approaches. *Expert Systems with Applications*, 160:113647.

López-Úbeda, P., M. Dıaz-Galiano, M. T. Martın-Valdivia, y L. A. Ureña-López. 2020c. Extracting neoplasms morphology mentions in spanish clinical cases through word embeddings. *Proceedings of IberLEF*.

López-Úbedaa, P., J. M. Perea-Ortegab, M. C. Díaz-Galianoa, M. T. Martín-Valdiviaa, y L. A. Ureña-Lópeza. 2020. Sinai at ehealth-kd challenge 2020: Combining word embeddings for named entity recognition in spanish medical records.

Procesamiento del Lenguaje Natural, Revista nº 68, marzo de 2022, pp. 153-156 recibido 18-12-2021 revisado 01-02-2022 aceptado 09-02-2022

The Observational Representation Framework and its Implications in Document Similarity, Feature Aggregation and Ranking Fusion

El Marco de Representación Observacional y su Implicación en Similaridad de Documentos, Agregación de Características y Fusión de Rankings

Fernando Giner Martínez
Research Group in Natural Language Processing and Information Retrieval
Universidad Nacional de Educación a Distancia (UNED)
C/ Juan del Rosal, 16, 28040 - Madrid, Spain
fginer3@gmail.com

Abstract: This is a summary of the Ph.D. thesis written by Fernando Giner Martínez at National Distance University ETSI - UNED, under the supervision of Ph.D. Enrique Amigó Cabrera. The author was examined on Thursday, September 23th, 2021 by a committee formed by Ph.D. Fermín Moscoso del Prado Martín from Lingvist and Radbound University Nijmegen, Ph.D. Julián Urbano from University of Delf and Ph.D. Victor Fresno Fernández from UNED. The Ph.D. thesis obtained Summa Cum Laude.
Keywords: Document representation, information theory, similarity, ranking fusion.

Resumen: Este es un resumen de la tesis doctoral realizada por Fernando Giner Martínez en la Universidad Nacional de Educación a Distancia ETSI - UNED bajo la dirección del doctor D. Enrique Amigó Cabrera. El acto de defensa tuvo lugar el jueves 23 de septiembre de 2021 ante el tribunal formado por los doctores D. Fermín Moscoso del Prado Martín de Lingvist y de Radbound University Nijmegen, D. Julián Urbano de la Universidad de Delf y D. Victor Fresno Fernández de la UNED. La tesis obtuvo la calificacion de Sobresaliente Cum Laude por unanimidad.
Palabras clave: Representación de documentos, teoria de la información, similaridad, fusión de ránkings.

1 Introduction

Information Access is a research area which involves many tasks such as Text Mining, Information Retrieval or Text Categorization. In all these tasks, document representation is a key step. Document features can be binary, such as word occurrence, named entities, links, or any kind of linguistic structure. Other features are defined in a continuous range, such as time stamp, topicality, sentiment polarity, etc.

We can highlight three main issues in document representation. First, features have a certain importance in the information access process. For instance, the word "Obama" has more weight than "said" when managing news. The second issue is the analysis of feature dependencies. For instance, expected words do not provide new information. In the news domain, "Obama" does not contribute substantially to the information provided by "Barak Obama", given that "Obama" is informative enough in this context. The third issue is feature scaling. For instance, time stamps and word occurrences are completely different scales.

These three issues are tackled in a different way depending on whether we are in a supervised or unsupervised scenario. In the first case, manually annotated output samples are available and features can be weighted, reduced or projected on the basis of their predictive power. In other words, the

training process adapts the learned model to the statistical dependencies and scale properties of features. For instance, a supervised classifier learns that "Obama" is more relevant than "said" when classifying news by topic. It can also infer that "Barak" does not provide additional evidence regarding "Obama", and also that news published less than 72 hours ago are more relevant for readers. Although in some contexts supervised approaches have shown to be highly effective, their drawbacks have been widely discussed in the literature, such as overfitting, domain dependency, data bias, annotation cost, etc. Another important drawback is that supervised learning does not provide mechanisms to manage information pieces, e.g., aggregation or comparison operators.

On the other hand, in the absence of human annotated data, the weight, dependence or scale of features is determined according to their distribution in a document collection. Typically, unexpected features have more presence in the representation than expected feature values. For instance, the word-feature "Obama" has more weight in the representation than frequent common words. The feature dependency can be also inferred from coocurrence. For instance, "Barak" and "Obama" are two word features which tends to appear together. As discussed in the first chapters of the thesis, the unexpectedness and coocurrence of features is the basis of the popular tf.idf feature weighting, stopwords removal or word sequence perplexity in language models. However, this paradigm is not compatible with the management of continuous feature values. The reason is that estimating expectedness in terms of occurrence requires some kind of value discretization. That is, we can estimate the probability of a word, n-gram, tag, etc. However, the likelihood of values in continuous features, for instance time stamp, depends on the granularity in which time is discretized (i.e. days, minutes, etc.). As far as we know, there are not standard criteria to quantify the likelihood of continuous feature values in the context of document representation for Information Access.

In order to overcome this challenge, this thesis presents the Observational Representation Framework (ORF). This approach integrates properties from representation frameworks based on feature set, vector spaces and information theory. Just like vector spaces representations, it captures continuous values. Just like feature set based representations, it allows apply operators such as inclusion or union, and just like information theory based representations and weighting functions, ORF weights features in terms of their likelihood.

ORF has relevant implications in different lines. In this thesis we delved into three of them. First, it provides a common theoretical framework to analyse, compare and generalise document similarity functions which are based on different representation schemes. Second, it allows to integrate intrinsic and extrinsic document features in the same representation. Intrinsic features includes words, n-grams, etc. Extrinsic features can be the output of a clustering process or category membership values generated by classification systems. And third, it provides us a theoretical foundation and mechanism for ranking fusion.

2 General outline of the dissertation

This thesis is organized in eight chapters. A brief summary of the content of each chapter is provided below.

Chapter 1 It provides a motivation for the formalization of document representation as a task of information access and establishes the contributions of this thesis.

Chapter 2 We review the main representation approaches in unsupervised tasks. We highlight their strengths ans weaknesses, analysing their ability to capture: (i) specificity, which establishes that the less common aspects of the information pieces should have greater relevance, since they are the features that distinguish them from the rest of the information pieces, (ii) diversity, which establishes the existence of relationships between the different features of the information pieces; the elimination of redundancies facilitates the study of these relationships and (iii) quantitativity, which establishes the need to capture binary and quantitative characteristics.

Chapter 3 Our representation framework ORF is presented (Giner, Amigó, and Verdejo, 2020; Giner and Amigó, 2016). It deals to an extension of the traditional Shannon's notion of information content, the one we have called *Observational Information*

Quantity (OIQ). This extension is able to manage continuous feature values. ORF not only fulfils the three properties highlighted in the previous chapter (specificity, dependence and quantitivity), but also verifies others, such as monotonicity with respect to values and features, as well as monotonicity with respect to union and the combination of inverse features. It is also able to generalize the most used representation models.

Chapter 4 We present a revision of the similarity axiomatic between pieces of information, such as distances in a metric space, Tversky's feature-based similarity, etc. Based on the hypothesis that there is a universal set of similarity principles that must be observed with respect to the space of features and the representations of pieces of information, we define a set of restrictions: *identity*, *identity specificity*, *unexpectedness* and *dependency*. These restrictions can be summarized in a single axiom: *similarity information monotonocity* (SIM), which considers *pointwise mutual information* (PMI) and conditional probability as two complementary aspects (Amigó et al., 2020; Amigó et al., 2017a).

Chapter 5 In this chapter, similarity functions are classified according to their representation paradigm. Based on ORF, we propose a similarity measure called information contrast model (ICM) (Amigó et al., 2011). ICM generalizes both the Pointwise Mutual information and the set-based models considering additions and joints of information quantities. We also present a study case on sentence similarities based on statistics in a popular image description corpus.

Chapter 6 We focus on the reputation-monitoring scenario, in which social media messages are analysed to identify conversations or events that can affect the reputation of a company or brand. The proposed ORF model is compared with different representation frameworks, using as baseline common schemes, such as bag of words and *tf.idf*. In order to measure the proximity between information pieces, similarity measures common in the literature are used (pointwise mutual information, Jaccard and Lin's distances), in addition to the similarity measure proposed in this work: ICM. Our experiments confirm the hypothesis that adding heterogeneous features under the same ORF-based weighting criterion increases progressively the similarity estimation performance, even when features include both discrete and continuous values and have different scale properties. Finally, a small study is carried out to improve the performance of the approaches through the parameterization of the proposed model (Giner, Amigó, and Verdejo, 2020).

Chapter 7 Based on experimental results, we highlight a set of desirable properties that any ranking fusion procedure should satisfy. We then analyse whether the main ranking fusion methods, such as averaging, Borda's rule, the family of Condorcet's methods, etc, satisfy them. Then, we observe that the ORF model presented in this work can be adapted as a ranking fusion method (assuming item scores as features). In addition, ORF satisfies all the desired properties, and moreover, we see under which conditions the ranking fusion algorithms approximate OIQ. Finally, we also present the performance of the ranking fusion methods in the experimental part (Amigó et al., 2017b; Amigó et al., 2018).

Chapter 8 A summary of each chapter can be seen, and some conclusions are drawn.

In addition, the thesis contains an appendix with the formal demonstrations of the statements established in previous chapters.

3 Contributions

The first contribution in this thesis is an in-depth study of the benefits and limitations of existing representation models. In particular, we analyse their ability to capture feature specificity, diversity and quantitativity (discrete vs. continuous feature values). After formalising a number of desirable properties, we observe that none of the families of document representation frameworks (e.g. set-based, metric spaces, language models, etc.) complies with all constraints.

On the basis of this analysis, the second and main contribution in this thesis is the definition of the Observational Representation Framework (ORF), which extends the traditional Shannon's notion of Information Content ($-log(P(x))$) to the management of continuous feature values. This is called the Observational Information Quantity (OIQ) and is grounded on feature fuzzy sets and inclusion relationships between document observation outcomes in a document collection. We study in a comprehensive way the for-

mal properties of ORF and OIQ as well as their generalization power regarding traditional representation approaches.

The third contribution is the analysis of similarity functions and their foundations (i.e. cosine, euclidean, feature overlap, etc.). We will see, through the study of counterexamples and evidence provided in the literature, that euclidean axioms, as well as set-based axioms (Tversky's model) do not capture similarity properly in the context of information access systems. On the basis of ORF, we review the axiomatic in which traditional similarity functions are based. Again, our analysis shows that different families of similarity functions comply with different constraints. Based on this analysis, we present a general and parametrisable similarity function called Information Contrast Model (ICM). ICM, besides satisfying desirable formal constraints, it generalises traditional functions such as PMI, conditional probability, euclidean distance or Tversky's Linear Contrast Model.

The fourth contribution is related with the capability of ORF to aggregate heterogenous features in a document representation. For this, we develop a study case: clustering of tweets in the context of on-line reputation management. We prove empirically that the model integrates effectively discrete features (words) with continuous feature values. In our study case, continuous values are the proximity to pre-annotated categories of tweets and previously generated clusters. The results show that adding heterogeneous features increases the similarity predictive power between tweet representations. In this sense, ORF allows us to integrate explicit features (i.e. words) with features extracted from supervised processes (class membership).

Finally, the fifth contribution is a study of the foundations of unsupervised fusion ranking fusion on the basis of OIQ and ORF. The application of our framework in ranking fusion is developed on the basis that rank scores can be interpreted as quantitative document features. We verify that the Observational Information Quantity (OIQ) generalises traditional ranking fusion algorithms and explains the effectiveness of existing approaches under different situations. We study empirically these phenomena on six different ranking fusion scenarios.

Acknowledgements

The work was supported by the Ministerio de Economía y Competitividad, TIN Program (Vemodalen), under Grant Number: TIN2015−71785−R.

References

Amigó, E., F. Giner, J. Gonzalo, and F. Verdejo. 2017a. An axiomatic account of similarity. In *Proceedings of the SIGIR'17 Workshop on Axiomatic Thinking for Information Retrieval and Related Tasks (ATIR)*, SIGIR '20, New York, NY, USA. ACM.

Amigó, E., F. Giner, J. Gonzalo, and F. Verdejo, 2017b. *A Formal and Empirical Study of Unsupervised Signal Combination for Textual Similarity Tasks*, pages 369–382. Springer International Publishing, Cham.

Amigó, E., F. Giner, J. Gonzalo, and F. Verdejo. 2020. On the foundations of similarity in information access. *Inf. Retr. J.*, 23(3):216–254.

Amigó, E., F. Giner, S. Mizzaro, and D. Spina. 2018. A formal account of effectiveness evaluation and ranking fusion. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 123–130.

Amigó, E., J. Gonzalo, J. Artiles, and F. Verdejo. 2011. Combining evaluation metrics via the unanimous improvement ratio and its application to clustering tasks. *J. Artif. Intell. Res. (JAIR)*, 42:689–718.

Giner, F. and E. Amigó. 2016. General representation model for text similarity. In *Proceedings of FETLT'16*.

Giner, F., E. Amigó, and F. Verdejo. 2020. Integrating learned and explicit document features for reputation monitoring in social media. *Knowledge and Information Systems*, 62(3):951–985.

Información General

SEPLN 2022

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

20-23 de septiembre 2022
https://sepln2022.grupolys.org/

1 Presentación

La XXXVIII edición del Congreso Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) se celebrará los días 20, 21, 22 y 23 de septiembre de 2022.

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

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

2 Objetivos

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

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

3 Áreas Temáticas

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

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

ISSN: 1135-5948

- Generación de Resúmenes.
- PLN para la generación de recursos educativos.
- PLN para lenguas con recursos limitados.
- Aplicaciones industriales del PLN.

4 Formato del Congreso

La duración prevista del congreso será de tres días, con sesiones dedicadas a la presentación de artículos, proyectos de investigación en marcha y demostraciones de aplicaciones. Además, tendrá lugar la cuarta edición de IberLEF el día 20 de septiembre.

5 Comité ejecutivo SEPLN 2022

Presidenta del Comité Organizador

- Miguel A. Alonso Pardo (Universidad de La Coruña).

Colaboradores

- Margarita Alonso Ramos (Universidad de La Coruña).
- Carlos Gómez Rodríguez (Universidad de La Coruña).
- Jorge Graña Gil (Universidad de La Coruña).
- Nancy Vázquez Veiga (Universidad de La Coruña).
- David Vilares Calvo (Universidad de La Coruña).
- Jesús Vilares Ferro (Universidad de La Coruña).

6 Consejo Asesor

Miembros:

- Xabier Arregi (Universidad del País Vasco, España).
- Manuel de Buenaga Rodríguez (Universidad de Alcalá, España).
- José Camacho Collados (Cardiff University, Reino Unido).
- Sylviane Cardey-Greenfield (Centre de recherche en linguistique et traitement automatique des langues, Lucien Tesnière. Besançon, Francia).
- Irene Castellón Masalles (Universidad de Barcelona, España).
- Arantza Díaz de Ilarraza (Universidad del País Vasco, España).
- Antonio Ferrández Rodríguez (Universidad de Alicante, España).
- Koldo Gojenola Galletebeitia (Universidad del País Vasco, España).
- Xavier Gómez Guinovart (Universidad de Vigo, España).
- José Miguel Goñi Menoyo (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).
- Bernardo Magnini (Fondazione Bruno Kessler, Italia).
- Nuno J. Mamede (Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa, Portugal).
- M. Teresa Martín Valdivia (Universidad de Jaén, España).
- Patricio Martínez Barco (Universidad de Alicante, España).
- Eugenio Martínez Cámara (Universidad de Granada, España).
- Paloma Martínez Fernández (Universidad Carlos III, España).
- Raquel Martínez Unanue (Universidad Nacional de Educación a Distancia, España).
- Ruslan Mitkov (University of Wolverhampton, Reino Unido).
- Manuel Montes y Gómez (Instituto Nacional de Astrofísica, Óptica y Electrónica, México).
- Mariana Neves (German Federal Institute for Risk Assessment, Alemania).
- Lluis Padró Cirera (Universidad Politécnica de Cataluña, España).
- Manuel Palomar Sanz (Universidad de Alicante, España).
- Ferrán Pla (Universidad Politécnica de Valencia, España).
- Germán Rigau Claramunt (Universidad del País Vasco, España).
- 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).
- Encarga 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).
- M. Felisa Verdejo Maíllo (Universidad Nacional de Educación a Distancia, España).
- Manuel Vilares Ferro (Universidad de la Coruña, España).
- Luis Villaseñor-Pineda (Instituto Nacional de Astrofísica, Óptica y Electrónica, México).

7 Fechas importantes

Fechas para la presentación y aceptación de comunicaciones:

- Fecha límite para la entrega de comunicaciones: 31 de marzo de 2022.
- Notificación de aceptación: 6 de mayo de 2022.
- Fecha límite para entrega de la versión definitiva: 20 de mayo de 2022.

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 página web de la Sociedad Española para el Procesamiento del Lenguaje Natural (http://www.sepln.org)
- Para los trabajos con formato LaTeX se mandará el archivo PDF junto a todos los fuentes necesarios para compilación LaTex
- Para los trabajos con formato Word se mandará el archivo PDF junto al DOC o RTF
- Para más información http://www.sepln.org/index.php/la-revista/informacion-para-autores

Información Adicional

Funciones del Consejo de Redacción

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

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

El consejo de redacción está formado por los siguientes miembros

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

Funciones del Consejo Asesor

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

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

El Consejo Asesor está formado por los siguientes miembros:

Xabier Arregi	Universidad del País Vasco (España)
Manuel de Buenaga	Universidad de Alcalá (España)
Sylviane Cardey-Greenfield	Centre de recherche en linguistique et traitement automatique des langues (Francia)
Irene Castellón	Universidad de Barcelona (España)
José Camacho Collados	Cardiff University (Reino Unido)
Arantza Díaz de Ilarraza	Universidad del País Vasco (España)
Antonio Ferrández	Universidad de Alicante (España)
Koldo Gojenola	Universidad del País Vasco (España)
Xavier Gómez Guinovart	Universidad de Vigo (España)
José Miguel Goñi	Universidad Politécnica de Madrid (España)
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)

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