

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 Rodríguez-Penagos,¹
 Aitor Gonzalez-Agirre,¹ Marta Villegas¹
¹Barcelona 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 extraídas 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-TL¹ 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

¹<https://plantl.mineco.gob.es/>

²<https://github.com/PlanTL-GOB-ES/1m-spanish>

* Equal contribution.

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 BERTIN⁴, 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.

³The multilingual 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 BNE⁵) 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 library⁷ 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 char-

acters 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 identifier¹⁰ 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 corpus¹² (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 corpus¹³ (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/1m-spanish>

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

¹³https://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

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.

having material originally written in 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

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), BERTIN¹⁵ 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,

¹⁵<https://huggingface.co/bertin-project/bertin-roberta-base-spanish/tree/v1-512>

¹⁶<https://huggingface.co/mrm8488/electricidad-base-generator>

and a total of 110M parameters.¹⁷ However, the actual version uploaded to HuggingFace¹⁸ 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 OPUS¹⁹ 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 ELECTRICIDAD 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.

¹⁷Note 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/>

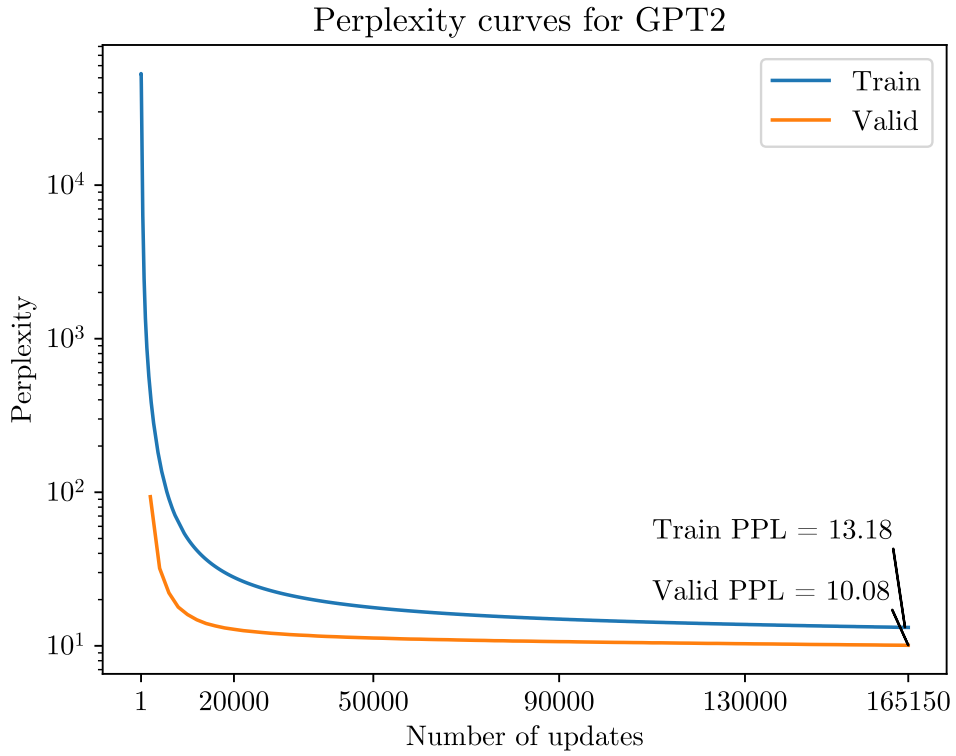


Figure 1: Perplexity curves for GPT2 model.

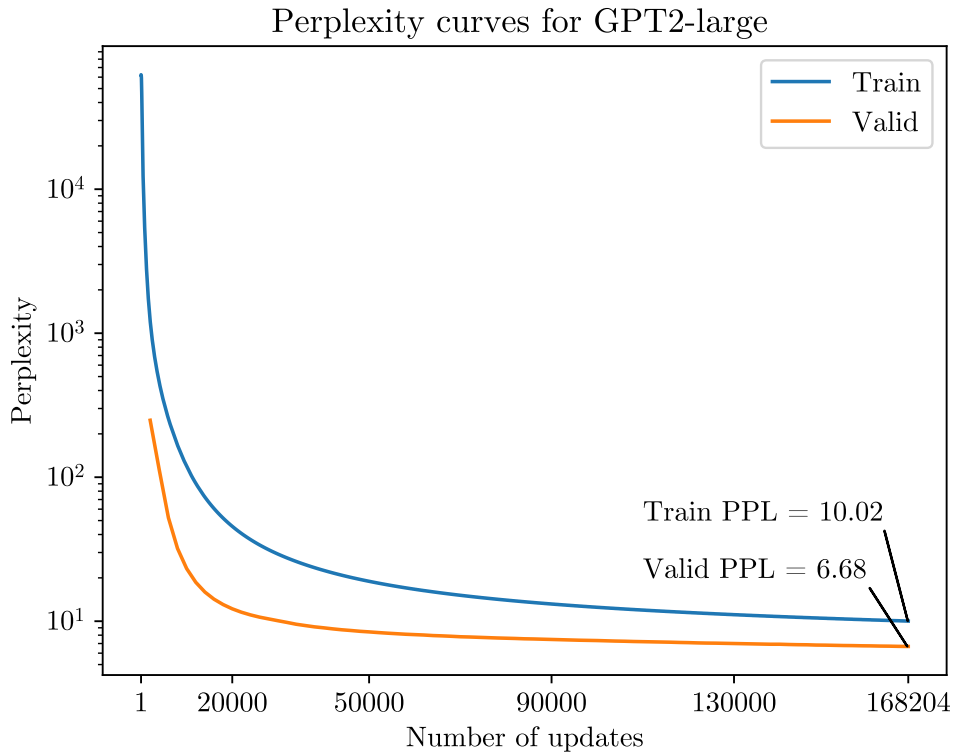


Figure 2: Perplexity curves for GPT2-large model.

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

²¹<https://github.com/PlanTL-GOB-ES/lm-spanish>

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.

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	barríos	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 CBOW²² and Skip-gram²³ 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 node²⁴ 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 tienes 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 tienes 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.