A Language Model Trained on Uruguayan Spanish News Text

**Juan Pablo Filevich[1], Gonzalo Marco[1],
Santiago Castro[2], Luis Chiruzzo[1], Aiala Rosá[1]**
[1]Universidad de la República – Uruguay [2]University of Michigan – Ann Arbor, USA
{juan.filevich,gonzalo.marco.mohotse}@fing.edu.uy

Abstract

This paper presents a language model trained from scratch exclusively on a brand new corpus consisting of about 6 GiB of Uruguayan newspaper text. We trained the model for 30 days on a single Nvidia P100 using the RoBERTa-base architecture but with considerably fewer parameters than other standard RoBERTa models. We evaluated the model on two NLP tasks and found that it outperforms BETO, the widely used Spanish BERT pre-trained model. We also compared our model on the masked-word prediction task with two popular multilingual BERT-based models, Multilingual BERT and XLM-RoBERTa, obtaining outstanding results on sentences from the Uruguayan press domain. Our experiments show that training a language model on a domain-specific corpus can significantly improve performance even when the model is smaller and was trained with significantly less data than more standard pre-trained models.

Keywords: Uruguay, News Corpus, Pre-trained Language Model

1. Introduction

In recent years, the Natural Language Processing community has witnessed considerable improvements in several areas – including Question Answering (Izacard et al., 2022; Zhang et al., 2021), Machine Translation (Takase and Kiyono, 2021; Liu et al., 2020a), and Sentiment Analysis (Raffel et al., 2020; Yang et al., 2019) – largely due to the advances in the pre-training methodology and the availability of data and pre-trained models to build upon (Jia et al., 2022; Liu et al., 2020b; Tian et al., 2020). Even though most of these advances have focused on English (Brown et al., 2020; Devlin et al., 2019; Liu et al., 2019), several efforts have considered multiple languages, including Spanish (Cañete et al., 2020; Pérez et al., 2022; De la Rosa et al., 2022; Xue et al., 2021; Conneau et al., 2020).

Current approaches that employ the Spanish language focus on pre-training on data dominated by Spanish varieties from countries with the most speakers or the most resources (i.e., Mexico, USA, and Spain). For example, the corpus used for BETO (Cañete et al., 2020; Cañete, 2019) employs many European source texts, hinting at a strong presence of Peninsular Spanish. Low-resource Spanish varieties have been broadly left behind, even when the Spanish language, like other languages, varies significantly from country to country (and even by region) in aspects such as grammar and vocabulary (Lipski, 2012). In addition to linguistic diversity, there are culture-related aspects that are unique to each country and region, which are typically underrepresented in low-resource communities. Such aspects are present in the training data used by today's pre-trained language models, albeit typically dominated by high-resource languages.

This work compiles a corpus of Uruguayan texts and presents models trained using this data. As far as we are concerned, these are the first data and general-purpose models tailored to conducting Natural Language Processing research with Uruguayan-specific texts. The dataset features 900,000 documents obtained from four Uruguayan news outlets with 400 million tokens in total in 6 GiB of uncompressed data. The data has been meticulously filtered and cleaned for quality purposes.

In the current context of NLP and AI, access to computational resources has become increasingly more difficult, especially in Global South countries. In particular, this type of language model is significantly resource-intensive to train. Considering this, besides creating a model specifically tailored to Uruguayan text, our motivation is also to create a model that is smaller and, hence, less computationally intensive to train and use than the available ones. Instead of fine-tuning an already pre-trained larger model, in this work, we train our model from scratch to tailor its size to make it appropriate for limited-resource settings.

Another motivation for developing specific resources for processing local texts, particularly news texts, is the growing interest in their automatic analysis by Uruguayan researchers in areas such as Sociology, Economics, and Communications. We believe it is necessary to have a language model that represents this type of text as well as possible.

We show the quality of the data by training BERT-based models on it through ablations on Uruguayan-related tasks and also by

Proceedings of the Second TDLE Workshop @LREC-COLING-2024, pages 53–60

comparing them with other pre-trained models such as BETO (Cañete et al., 2020), and XLM-RoBERTa (Conneau et al., 2020). We also perform a qualitative analysis of the knowledge captured by such models. The dataset and the pre-trained models are publicly available at https://huggingface.co/pln-udelar/rouberta-base-uy22-cased.

2. Related Work

Several works have compiled corpora in Spanish for research. Cañete (2019) compiled a 3-billion-word training corpus by combining multiple sources, including subtitles and news stories, an updated version of the one compiled by Cardellino (2019). Pérez et al. (2022) collected 622 million tweets in Spanish. Gutiérrez-Fandiño et al. (2022) built a massive corpus of 135 billion words from the Spanish Web Archive. Other works, such as (Wenzek et al., 2020; Conneau et al., 2020), have built multilingual datasets by leveraging efforts such as Common Crawl. As far as we are aware, our work is the first one to build a Spanish corpus dedicated to studying the Uruguayan variety and cultural references in the text.

Regarding pre-trained models, there have been efforts to build both multilingual models and Spanish-specific ones. Several multilingual models have originally been implemented from model architectures that had been used to train models for English first, such as Multilingual BERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020), mT5 (Xue et al., 2021), and mBART (Liu et al., 2020b). More recently, other efforts have focused on building multilingual large language models, such as BLOOM (Scao et al., 2022), GPT-4 (OpenAI, 2023) and PaLM (Chowdhery et al., 2022). A line of work has focused on Spanish-specific models (Cañete et al., 2020; De la Rosa et al., 2022) and specific domains, such as RoBERTuito (Pérez et al., 2022), specifically for Spanish tweets. Unlike previous works, this paper presents models trained on Uruguay-specific Spanish data, which can capture its particular linguistic and cultural features.

The idea of training a domain-specific LM has been attempted in the past. Still, these generally start from large models or are trained with significantly more data, making them computationally intensive. Some existing domain-specific LMs are created by fine-tuning a general language model (e.g. BioBERT, Lee et al. (2019), FinBERT, Araci (2019); MatSciBERT, Gupta et al. (2022)), or are trained from scratch as our model, but using a larger corpus (e.g. SciBERT, Beltagy et al. (2019); RoBERTuito, Pérez et al. (2022)). In this work, our main goal was to obtain good performance with a model significantly smaller than that of the usual language models, trained on a relatively small data set, and with shorter training times due to limitations in computational resources. So, we are testing not only the usefulness of having a domain-specific model but also the performance of a small model trained with few resources. Verifying the usefulness of such a model is notoriously relevant for us since we usually work in both model training and inference in low-resource contexts.

3. A Uruguayan News Corpus

We scraped four of some of the most important media outlets in Uruguay: *El Observador*, *El País*, *Montevideo Portal*, and *La Diaria*. The first three were scraped from the Internet, while the latter provided us with their articles. For every article, we retrieved the main text (i.e., the article's body) and some potentially useful metadata such as the URL, date, category, title, keywords, and a front picture or cover (if any). After one month of scraping (carried out between November and December 2022), we collected more than 6 GiB of uncompressed data, with articles spanning from the early 2000s up to December 2022. We call our new corpus *UY22*. Table 1 shows the distribution of articles for each website.

We conducted a data quality assurance process based on stripping the HTML tags, trimming and removing duplicate whitespaces, the normalization of strange characters using the *Unidecode* Python library, the removal of emojis, and converting the links into the string "<link>". We also deleted any article with fewer than sixteen words. We split the texts by document and split them into sentences. We refer interested readers to a more in-depth explanation of the scraping and preprocessing phases of this corpus to this project repository[1]. We made the raw and clean versions publicly available (the latter is about 4 GiB uncompressed).

4. ROUBERTa: a Uruguayan LM

We employ a RoBERTA-base (Liu et al., 2019) architecture and train it on the clean version of our data using HuggingFace's Transformers library (Wolf et al., 2019). We use a BPE (Sennrich et al., 2016) tokenizer with a vocabulary size of 30,000 tokens. The model is trained for 30 days on ClusterUY (Nesmachnow and Iturriaga, 2019) with one NVIDIA P100 (12 GiB) for about 6 million steps using RoBERTa's training objective (Masked Language Modeling – cross-entropy loss on the prediction of a masked token, where each token has a 15% masking probability). We show in Figure 1 the training loss curve we obtained. We

[1] https://gitlab.fing.edu.uy/uy22/uy22

Name	Website	# Articles	# Words	From	To	Share
El Observador	elobservador.com.uy	314,821	150,007,925	2011	2022	37%
El País	elpais.com.uy	147,004	92,605,424	2013	2022	23%
Montevideo Portal	montevideo.com.uy	433,244	145,422,666	2000	2022	36%
La Diaria	ladiaria.com.uy	20,000	14,079,916	2009	2021	4%

Table 1: UY22 corpus statistics. The share of each website is computed based on the number of words.

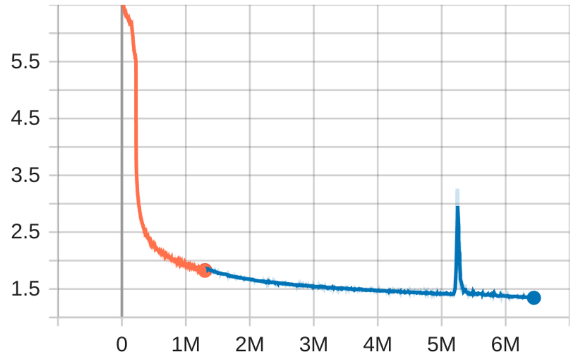

Figure 1: Loss curve for the cased variant of our model with an exponential moving average smoothing value of 0.6. The x-axis shows the number of training steps. The y-axis shows the loss. The color change shows when we restarted the training with a smaller max context length and a larger batch size.

trained for this number of steps due to our limited computational resources and the fact that the loss value was still converging. We chose to train the model from scratch instead of fine-tuning a more general pre-trained model for Spanish, seeking to obtain a model of an appropriate size for use with medium-end computers.

We note a loss spike between 5M and 6M training steps. This could be due to multiple reasons (Takase et al., 2023; Wortsman et al., 2024), including a large amount of consecutive bad-quality data (Soldaini et al., 2024), a large beta2 parameter when using Adam, or a very high learning rate for the batch size we employed. However, we still need to conduct further analyses to understand what is happening in this case.

Figure 2 shows the performance of the model on a Sentiment Analysis task (see Section 5 for details) at different moments during training. We employed a Dynamic Masked Language Modelling task, following Liu et al. (2019). Most hyperparameter values were similar to those used to train RoBERTa, including a max sequence length of 384. However, to cope with GPU memory limitations, we decided to stop it early during training and continue with a batch size of 32 and a max sequence length of 128 (plus two for the special tokens). The learning rate started from 1e-4 and was linearly decayed during training.

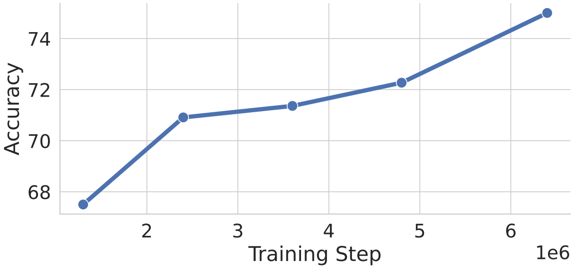

Figure 2: Performance of the cased variant of the model on a Sentiment Analysis task concerning the number of training steps.

The model is named *ROUBERTa* (after ROU-based RoBERTa, where ROU stands for “República Oriental del Uruguay” – the target country’s full name in its native language). We train both cased and uncased variants for our model, although, as we will see in Section 6, the cased variant generally has better results.

5. General Evaluation

This section and the following present the evaluation of our model. This first section will compare only the best-performing model against external baselines. At the same time, in Section 6, we present a deeper evaluation of some interesting cases with examples, and we show ablation tests to see how our design choices affected the model’s performance.

We evaluate our model on two in-domain tasks: Question Answering and Sentiment Analysis on Uruguayan news articles. We perform the experiments for our cased model and compare it with two strong baselines for Spanish: the BETO and XLM-Roberta models, using the cased versions of the models in all cases. We describe the two benchmarks hereafter.

Sentiment Analysis The first benchmark is a sentiment analysis dataset (Dufort y Álvarez et al., 2016), which is composed of a collection of short spans of text that contain an opinion (i.e., a statement by some actor about some topic) and its sentiment polarity in one of three classes (“POS”, “NEG” and “NEU”). The dataset contains 1261 examples and was split into 80% for training and 20% for

Model	Sentiment Analysis	Question Answering	
	Accuracy	EM	F1
XLM-RoBERTa	73.4	26.8	**36.4**
BETO	74.6	24.6	29.4
ROUBERTa	**75.0**	**28.1**	32.3

Table 2: Results of the main experiments.

testing. The model was fine-tuned for 6 epochs with a learning rate of 1e-5.

Question Answering The second benchmark is the QuALES question answering task (Rosá et al., 2022), which contains questions from Uruguayan news articles about the COVID-19 pandemic. All articles, questions, and answers are in Spanish. The QuALES dataset is rather small compared to other QA datasets, containing around 3,600 question-answer pairs (for comparison, SQuAD (Rajpurkar et al., 2016) has more than 100,000), and only 1,000 of those comprise the training set. The dataset format is similar to SQuAD's, which enabled us to experiment with a widely used strategy for Question Answering tasks, starting from a pre-trained BERT-based model and fine-tuning with SQuAD data. In this case, we swapped the SQuAD data with the QuALES data and used the following models as starting points: BETO (`bert-base-spanish-wwm-cased`), XLM-Roberta (`xlm-roberta-base`), and our ROUBERTa-base-cased. We fine-tuned the models for 3 epochs with a batch size of 16 and a max context length of 384. The learning rate was 2e-5, and the weight decay was 0.01. We employ the evaluation metrics Exact Match and F1.

Table 2 shows the results for both experiments. The ROUBERTa_cased model outperforms the other baselines for the sentiment analysis task and the exact match metric of the QA task while getting a solid second place considering the F1 metric of the QA task. In this second task, ROUBERTa_cased performs slightly better than BETO and outperforms XLM-RoBERTa by almost two points. However, in the QA task, ROUBERTa_cased only outperforms XLM-RoBERTa in the Exact Match metric and outperforms BETO in both analyzed metrics. Overall, we can say that our model achieves a more even performance, always within the top ranks among the three compared models. It is worth mentioning that XLM-RoBERTa was chosen for these experiments instead of comparing us with a Spanish version of RoBERTa, based on the results of the original QuALES competition (Rosá et al., 2022), in which the top performing systems used XLM-RoBERTa.

ID	Source	# Masked Sentences	XLM-RoBERTa	BETO	ROUBERTa
text01	La Diaria 06/21/2023	533	151	160	**219**
text02	La Diaria 06/21/2023	170	54	42	**70**
text03	Montevideo Portal 06/22/2023	235	63	75	**123**
text04	Montevideo Portal 06/08/2023	245	78	95	**114**
text05	Búsqueda 07/06/2023	335	98	107	**151**
text06	La Diaria 02/28/2024	304	77	66	**112**
text07	Montevideo Portal 02/20/2024	546	116	96	**193**
	Total	2368	637 (27%)	641 (27%)	**982** (**42**%)

Table 3: Evaluation based on the word-masking task. For each model, we show the number of masked words that were correctly predicted. The dates are in mm/dd/yyyy format.

6. Ablation Study

In this section, we perform an empirical justification of the decisions we have made to build our corpus and train our models.

6.1. Predicting Words in Unseen Texts

We are interested in inquiring if the trained model captures aspects of the Uruguayan culture. We evaluate our model and others on a masking task using five recent texts from three different Uruguayan media outlets. The objective is to evaluate whether the model captures country-specific information such as names of public people, locations, organizations, and the style of the local press. We analyze examples where the masked words contain such information, and, as we will show, our model tends to perform better than general models. In the following analysis, we include some examples to illustrate this behavior.

Note that the model has not seen any text employed in this evaluation since they belong to more recent news articles than the training data. Furthermore, one of the media outlets employed here, Búsqueda[2], is not part of the four media outlets used by our training data. Consequently, this evaluation measures the generalization capacity of our

[2] busqueda.com.uy

trained model on unseen in-domain data. The texts were selected based on different criteria. Some texts are about usual topics in the local press: text03 is about judicial issues, text05 is about insecurity, and texts 06 and 07 are about Carnival, a popular cultural activity in Uruguay. Other texts are about current topics that are not frequent in the country: text01, text02, and text04 are about a severe drought that caused issues with the drinking water distribution in 2023.

To carry out this evaluation, we proceed as follows. For each text t_i and each sentence s_{ij}, we generate different versions of the sentence by masking each word with more than four letters. Except for the masked word, each new sentence is the same as s_{ij}. By these means, we obtain an extended set of sentences for each text, $ExtSent_i$, where each original sentence s_{ij} has multiple versions, one for each masked word. Then, for each sentence in $ExtSent_i$, we obtain candidates for each masked word, using our model ROUBerta_cased, and three other strong models: BETO (Cañete et al., 2020), trained specifically for Spanish[3]; multilingual BERT (Devlin et al., 2019)[4], and the multilingual model XLM-RoBERTa (Conneau et al., 2020)[5]. Table 3 shows the results of this evaluation, except for multilingual BERT, which performed significantly worse than the rest of the models and was therefore not included in the table. As shown in the table, our model, trained exclusively with Uruguayan press texts, gives the best results for the five evaluated texts. It correctly predicts the masked word with a 42% top-1 accuracy, which is a high gap compared to the 27% accuracy obtained by the other models.

Analyzing the results, we observe some interesting examples. In texts about current topics not usually found in the press, such as the drought suffered this year, proper nouns related to our country are correctly predicted by our model, such as the name of the capital of Uruguay in the following example: *Es decir, el agua que sale por las canillas, sale con gusto salado, al menos en <mask> y el área metropolitana. || That is, the water that comes out of the taps, comes out tasting salty, at least in <mask> and the metropolitan area.*

Predictions
ROUBERTa: **Montevideo**
BETO: México
XLM-RoBERTa: Bogotá

On the other hand, the style of the texts also seems to have been captured by our model, as shown in the following example, where it correctly predicts a verb form very usual in the local press: *Luego de que radio Universal <mask> sobre la adjudicación de una vivienda bajo la modalidad de alquiler con opción a compra a una militante de Cabildo Abierto (CA) sin pasar por sorteo || After radio Universal <mask> about the awarding of a house under the rent-to-buy modality to a Cabildo Abierto (CA) militant without going through a raffle*

Predictions
ROUBERTa: **informara**
BETO: ##a
XLM-RoBERTa: informó

It can also be seen that our model incorporated the lexical preferences of the local press, as seen in the following example: *Así lo anunció la titular de la <mask>, Karina Rando, este jueves en conferencia de prensa. || This was announced by the head of the <mask>, Karina Rando, this Thursday at a press conference.*

Predictions
ROUBERTa: **cartera**
BETO: cadena
XLM-RoBERTa: entidad

Finally, for the Carnival theme, our model correctly predicts the word *murga*, which is a typical artistic expression of the Uruguayan carnival: *Desde que el Carnaval volvió tras la pandemia, solo una <mask> obtuvo el primer premio y fue Asaltantes con Patente. || Since Carnival returned after the pandemic, only one <mask> won first prize and that was Asaltantes con Patente.*

Predictions
ROUBERTa: **murga**
BETO: persona
XLM-RoBERTa: empresa

6.2. Is the Uruguayan Data Necessary?

To study the effect of using Uruguayan-specific data compared to a general-Spanish dataset, we trained a new RoBERTa (Liu et al., 2019) model with the corpus used for training BETO (Cañete et al., 2020; Cañete, 2019). RoBERTa models, ours, and the one trained with the BETO corpus were fine-tuned for the Sentiment Analysis task on Uruguayan news, following the steps described in Section 5. Table 4 shows the performance of both models on this task. We can see that our model achieves better results than the one trained with the BETO corpus, even when the latter is five times larger.

6.3. Whole-Word Masking

We consider the model's performance when using the whole-word masking technique introduced in BERT (Devlin et al., 2019) code repository. For this evaluation, we consider the same sentiment analysis. Table 5 shows the results. Not employing

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

[4] https://huggingface.co/bert-base-multilingual-cased

[5] https://huggingface.co/xlm-roberta-large

Training data	Size (GiB)	Acc.
BETO's (Cañete, 2019)	20	65.0
UY22 (ours)	4	**68.6**

Table 4: The model's performance on a Sentiment Analysis task when varying the training data. The uncased variant is employed.

Whole-word masking	Accuracy
Yes	35.0
No	**68.6**

Table 5: The model's performance on a Sentiment Analysis task when using the whole-word masking technique. The uncased variant is employed.

whole-word masking proved to be superior in our case, which is, on the one hand, inconsistent with BERT experiments but, on the other hand, consistent with what was reported by Dai et al. (2022).

6.4. Case Sensitivity

We study the effect of case sensitivity in the tokenization. These refer to the cased and uncased variants of the model. We present the results in Table 6. Similarly to other works, such as BERT (Devlin et al., 2019), the cased variant performs better than the uncased one.

7. Conclusion

In this work, we present a dataset specific to Uruguayan Spanish based on news articles and RoBERTa-based models pre-trained on it. We demonstrate the value of our new corpus and the pre-trained models through quantitative and qualitative evaluations employing Uruguayan-news-based tasks. We make both publicly available and hope to enable further research on Uruguayan Natural Language Processing. At the same time, we encourage other community members to replicate our efforts on other Spanish language varieties.

Our model performs better for the analyzed tasks, but most importantly, it did so using a smaller context length, a smaller corpus, and less GPU VRAM than usual. This shows that it is possible to achieve competitive metrics using fewer resources and smaller models. When comparing our results with the ones reported by (Agerri and Agirre, 2023), we observe our model was trained with a corpus significantly smaller than those considered in that paper and with a much smaller parameter count. Despite this consideration, our model achieves better results than others, particularly when compared to XLM-RoBERTa (except in F1 for the QA task presented in Section 5), which was the best model in the mentioned work. This is particularly relevant to researchers in this region, where we usually work in low-resource contexts for model training and subsequent use.

Variant	Accuracy
Uncased	68.6
Cased	**75.0**

Table 6: The model's performance on a Sentiment Analysis task when varying the case sensitivity.

The most important takeaway from this work is that we built a much smaller language model, trained on much less data and requiring much less computational power, and that still keeps up or outperforms other baselines for relevant tasks. This is essential in research labs with limited access to computational resources.

Ethics Statement

Even if we employed a small model, which requires considerably less power than larger models like RoBERTa, language model training typically requires significant energy consumption. However, the carbon footprint associated with our model's training was at least partially reduced given that we employed ClusterUY's infrastructure, which during the time of our experiments used more than 90% renewable energy sources[6]. Still, further analysis is needed to measure how big the impact is.

8. Bibliographical References

Rodrigo Agerri and Eneko Agirre. 2023. Lessons learned from the evaluation of spanish language models.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

[6] https://www.gub.uy/ministerio-industria-energia-mineria/comunicacion/noticias/uruguay-logra-90-energias-renovables-matriz-electrica-contexto-tres-anos

Cristian Cardellino. 2019. Spanish Billion Words Corpus and Embeddings.

José Cañete. 2019. Compilation of large spanish unannotated corpora.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. In *PML4DC at ICLR 2020*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Yong Dai, Linyang Li, Cong Zhou, Zhangyin Feng, Enbo Zhao, Xipeng Qiu, Piji Li, and Duyu Tang. 2022. "Is whole word masking always better for chinese bert?": Probing on chinese grammatical error correction.

Javier De la Rosa, Eduardo G Ponferrada, Manu Romero, Paulo Villegas, Pablo González de Prado Salas, and María Grandury. 2022. BERTIN: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68:13–23.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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. Association for Computational Linguistics.

Guillermo Dufort y Álvarez, Fabián Kremer, and Gabriel Mordecki. 2016. Determinación de la orientación semántica de las opiniones transmitidas en textos de prensa. Bachelor's thesis, Instituto de Computación, Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay.

Tanishq Gupta, Mohd Zaki, NM Anoop Krishnan, and Mausam. 2022. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquín Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano Oller, Carlos Rodríguez Penagos, Aitor Gonzalez-Agirre, and Marta Villegas Montserrat. 2022. MarIA: Spanish language models. *Procesamiento del Lenguaje Natural*, 68:39–60.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Robin Jia, Mike Lewis, and Luke Zettlemoyer. 2022. Question answering infused pre-training of general-purpose contextualized representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 711–728, Dublin, Ireland. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

John M Lipski. 2012. Geographical and social varieties of spanish: An overview. *The handbook of Hispanic linguistics*, pages 1–26.

Xiaodong Liu, Kevin Duh, Liyuan Liu, and Jianfeng Gao. 2020a. Very deep transformers for neural machine translation. *arXiv preprint arXiv:2008.07772*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020b. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sergio Nesmachnow and Santiago Iturriaga. 2019. Cluster-UY: Collaborative scientific high performance computing in Uruguay. In *Supercomputing*, pages 188–202, Cham. Springer International Publishing.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Juan Manuel Pérez, Damián Ariel Furman, Laura Alonso Alemany, and Franco M. Luque. 2022. RoBERTuito: a pre-trained language model for social media text in Spanish. In *Proceedings of the Language Resources and Evaluation Conference*, pages 7235–7243, Marseille, France. European Language Resources Association.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Aiala Rosá, Luis Chiruzzo, Lucía Bouza, Alina Dragonetti, Santiago Castro, Mathias Etcheverry, Santiago Góngora, Santiago Goycoechea, Juan Machado, Guillermo Moncecchi, et al. 2022. Overview of QuALES at IberLEF 2022: Question answering learning from examples in spanish. *Procesamiento del Lenguaje Natural*, 69:273–280.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. BLOOM: A 176B-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

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

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*.

Sho Takase and Shun Kiyono. 2021. Rethinking perturbations in encoder-decoders for fast training. *arXiv preprint arXiv:2104.01853*.

Sho Takase, Shun Kiyono, Sosuke Kobayashi, and Jun Suzuki. 2023. Spike no more: Stabilizing the pre-training of large language models. *arXiv preprint arXiv:2312.16903*.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. *arXiv preprint arXiv:2005.05635*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Mitchell Wortsman, Tim Dettmers, Luke Zettlemoyer, Ari Morcos, Ali Farhadi, and Ludwig Schmidt. 2024. Stable and low-precision training for large-scale vision-language models. *Advances in Neural Information Processing Systems*, 36.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin 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. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2021. Retrospective reader for machine reading comprehension. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14506–14514.