ALBETO and DistilBETO: Lightweight Spanish Language Models

José Cañete [1,3] Sebastián Donoso [1,3] Felipe Bravo-Marquez [1,3,4] Andrés Carvallo [2,3] Vladimir Araujo [2,3,4,5]

[1]University of Chile [2]Pontificia Universidad Catolica de Chile [3]National Center for Artificial Intelligence (CENIA) [4]Millennium Institute for Foundational Research on Data (IMFD) [5]KU Leuven

Introduction

- In recent years there have been considerable advances in pre-trained language models, where non-English language versions have also been made available.
- Many lightweight versions of these models (with reduced parameters) have also been released to speed up training and inference times.
- These lighter models (e.g., ALBERT, DistilBERT) for languages other than English are still scarce.
- We present ALBETO and DistilBETO, which are versions of ALBERT and DistilBERT pre-trained exclusively on Spanish corpora.
- We train several versions of ALBETO ranging from 5M to 223M parameters and one of DistilBETO with 67M parameters.
- When evaluating the models in different tasks, the results show that our lightweight models achieve competitive results to those of BETO (Spanish-BERT) despite having fewer parameters.

KEYWORDS: ALBERT, BERT, DistilBERT, Efficient Models, Language Models

Data and Models

Data

The data used to train all models was the same as that used to train BETO [4], which is an updated version of the dataset proposed by Cardellino [2]. This dataset has approximately 3 billion words which includes all Spanish Wikipedia and almost all the Spanish portion of the OPUS Project [14].

Spanish ALBERT (ALBETO)

ALBERT [8] is a more efficient BERT-style model in terms of parameters because it uses the weight-tied strategy, which means to share all parameters across layers of the model.

We introduce 5 ALBETO models: *tiny*, *base*, *large*, *xlarge* and *xxlarge*. These five models share a vocabulary of 31K lowercase tokens, that was constructed using SentencePiece [7] over the training dataset.

We trained each model using a single TPU v3-8.

Spanish DistilBERT (DistilBETO)

We trained the DistilBETO model using the distillation technique to transfer the knowledge of the BETO model to this new model following the work of DistilBERT [12]. DistilBETO was trained during 90k steps using a single GPU NVIDIA RTX 3090.

Models, sizes and task performance

Model	Parameters	Evaluation Average
BETO *uncased*	110M	77.48
BETO *cased*	110M	81.02
DistilBETO	67M	73.22
ALBETO *tiny*	5M	70.86
ALBETO *base*	12M	79.35
ALBETO *large*	18M	78.12
ALBETO *xlarge*	59M	80.20
ALBETO *xxlarge*	223M	81.34

Table 1. Comparison of different models in terms of size (number of parameters) and task performance (the average of the results in every task).

Evaluation Tasks

We evaluated all our models on 6 tasks, which are all part of the GLUES [4] benchmark.

1. **Document Classification** is the task of assigning an entire document to an appropriate category. For this task we are using the Spanish part of MLDoc corpus [13].
2. **Part of Speech** is a sequence labeling task that consists of tagging words in a text with their corresponding syntactic categories or part-of-speech. The dataset used for this task is the Spanish subset of Universal Dependencies (v1.4) Treebank [10].
3. **Named Entity Recognition** is a sequence labeling task, in which one tries to label entities in the text with their corresponding type, which can be names of people, organizations, places and miscellaneous items. For this task we are using the Spanish part of the Shared Task of CoNLL-2002 [11].
4. **Paraphrase Identification** consists of verifying whether two sentences are semantically equivalent or not. We are using the Spanish portion of PAWS-X [15] dataset.
5. **Natural Language Inference** is the task of determining whether a "hypothesis" is true (entailment), false (contradiction), or undetermined (neutral) given a "premise". For this task we are using the Spanish part of XNLI [5].
6. **Question Answering** consists of, given a context and a question about that context, highlighting the sequence of words within that context that answers the question. For this task we considered four different datasets: MLQA [9], TAR [3], XQuAD [1] and SQAC [6].

Results

In the following tables we present the results of each model in each evaluated task.

Model	POS	NER
BETO *uncased*	97.70	83.76
BETO *cased*	**98.84**	**88.24**
DistilBETO	97.50	81.19
ALBETO *tiny*	97.04	75.11
ALBETO *base*	98.08	83.35
ALBETO *large*	97.87	83.72
ALBETO *xlarge*	98.06	82.30
ALBETO *xxlarge*	98.35	84.36

Table 2. Comparison of ALBETO, DistilBETO and BETO models on the test set for the tasks of POS and NER, which are sequence tagging tasks and are evaluated using F1 score as metric.

Model	MLDoc	PAWS-X	XNLI
BETO *uncased*	96.38	84.25	77.76
BETO *cased*	96.65	89.80	81.98
DistilBETO	96.35	75.80	76.59
ALBETO *tiny*	95.82	80.20	73.43
ALBETO *base*	96.07	87.95	79.88
ALBETO *large*	92.22	86.05	78.94
ALBETO *xlarge*	95.70	89.05	81.68
ALBETO *xxlarge*	**96.85**	**89.85**	**82.42**

Table 3. Comparison of ALBETO, DistilBETO and BETO models on the test set for the tasks of MLDoc, PAWS-X and XNLI. These tasks are treated as sentence classification tasks and use the accuracy as evaluation metric.

Model	MLQA	SQAC	TAR, XQuAD
BETO *uncased*	64.12 / 40.83	72.22 / 53.45	74.81 / 54.62
BETO *cased*	67.65 / 43.38	78.65 / 60.94	77.81 / 56.97
DistilBETO	57.97 / 35.50	64.41 / 45.34	66.97 / 46.55
ALBETO *tiny*	51.84 / 28.28	59.28 / 39.16	66.43 / 45.71
ALBETO *base*	66.12 / 41.10	77.71 / 59.84	77.18 / 57.05
ALBETO *large*	65.56 / 40.98	76.36 / 56.54	76.72 / 56.21
ALBETO *xlarge*	68.26 / 43.76	78.64 / 59.26	**80.15 / 59.66**
ALBETO *xxlarge*	**70.17 / 45.99**	**81.49 / 62.67**	79.13 / 58.40

Table 4. Comparison of ALBETO, DistilBETO and BETO models on the task of QA. We show the results of the test set in each case. The task uses two metrics which are showed as F1 Score / Exact Match.

Conclusions

- We presented DistilBETO and five ALBETO models (*tiny*, *base*, *large*, *xlarge*, and *xxlarge*), comprising **six new pre-trained language models for Spanish** language.
- We also comprehensively evaluated each proposed model fine-tuned on a set of NLP tasks for Spanish. Our results indicate that the proposed models are competitive with the current models available for Spanish and are much more efficient in their number of parameters.
- We hope this work will expand the availability of pre-trained language models based on the Spanish language to gather a wider NLP community, including researchers, developers, and students.
- We envision several avenues of future research. First, we expect to evaluate these models on more tasks to increase the coverage of GLUES, which is our current evaluation benchmark [4]. Also, we also want to further analyze the fine-tuning and inference speed of these models. Finally, we plan to release more distilled models fine-tuned explicitly for many NLP tasks.

Acknowledgements

This work was supported in part by the National Center for Artificial Intelligence CENIA FB210017, Basal ANID and by the Patagón supercomputer of Universidad Austral de Chile (FONDEQUIP EQM180042). Felipe Bravo-Marquez was supported by ANID FONDECYT grant 11200290, U-Inicia VID Project UI-004/20 and ANID -Millennium Science Initiative Program - Code ICN17_002.

References

[1] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856, 2019.

[2] Cristian Cardellino. Spanish Billion Words Corpus and Embeddings, March 2016.

[3] Casimiro Pio Carrino, Marta R Costa-jussà, and José AR Fonollosa. Automatic spanish translation of the squad dataset for multilingual question answering. *arXiv preprint arXiv:1912.05200*, 2019.

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

[5] Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*, 2018.

[6] 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. Maria: Spanish language models. 2022.

[7] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018.

[8] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020.

[9] Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*, 2019.

[10] Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, ..., and Hanzhi Zhu. Universal dependencies 1.4, 2016. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[11] Erik F Sang. Introduction to the conll-2002 shared task: Language-independent named entity recognition. *arXiv preprint cs/0209010*, 2002.

[12] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[13] Holger Schwenk and Xian Li. A corpus for multilingual document classification in eight languages. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may 2018. European Language Resources Association (ELRA).

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

[15] Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. *arXiv preprint arXiv:1908.11828*, 2019.