SPANISH PRE-TRAINED BERT MODEL AND EVALUATION DATA

José Cañete,[*] **Gabriel Chaperon, Rodrigo Fuentes**
Department of Computer Science, Universidad de Chile
`{jcanete,gchapero,rfuentes}@dcc.uchile.cl`

Jou-Hui Ho, Hojin Kang
Department of Electrical Engineering, Universidad de Chile
`{jouhui.ho,ho.kang.k}@ug.uchile.cl`

Jorge Pérez
Department of Computer Science, Universidad de Chile &
Millennium Institute for Foundational Research on Data (IMFD), Chile
`jperez@dcc.uchile.cl`

ABSTRACT

The Spanish language is one of the top 5 spoken languages in the world. Nevertheless, finding resources to train or evaluate Spanish language models is not an easy task. In this paper we help bridge this gap by presenting a BERT-based language model pre-trained exclusively on Spanish data. As a second contribution, we also compiled several tasks specifically for the Spanish language in a single repository much in the spirit of the GLUE benchmark. By fine-tuning our pre-trained Spanish model we obtain better results compared to other BERT-based models pre-trained on multilingual corpora for most of the tasks, even achieving a new state-of-the-art on some of them. We have publicly released our model, the pre-training data and the compilation of the Spanish benchmarks.

1 INTRODUCTION

The field of natural language processing (NLP) has made an incredible progress in the last two years. Two of the most decisive features that have driven this improvement are the self-attention mechanism, in particular the Transformer architecture (Vaswani et al., 2017), and the introduction of unsupervised pre-training methods (Peters et al., 2018; Howard & Ruder, 2018; Devlin et al., 2018) that take advantage of huge amounts of unlabeled text corpora. Thus the leading strategy today for achieving good performance is to first train a Transformer-based model, say M, with a general language-modeling task over a huge unlabeled corpus and then, after this first training is over, "fine-tune" M by continue training it for a specific task using labeled data. Built upon these ideas, the BERT architecture –that stands for "Bidirectional Encoder Representations from Transformers"– (Devlin et al., 2018), and several improvements thereof (Liu et al., 2019; Lan et al., 2019; Yang et al., 2019b; Clark et al., 2019), changed the landscape of NLP during 2019.

BERT was initially released in two versions, one pre-trained over an English corpus and another over a Chinese corpus (Devlin et al., 2018). As a way of providing a resource for other languages besides English and Chinese, the authors also released a "multilingual" version of BERT (we call it mBERT from now on) pre-trained simultaneously over a corpus including more than 100 different languages. The mBERT model has shown an impresive performance when fine-tuned for language-specific tasks and has achieved state-of-the-art results on many cross-lingual benchmarks (Wu & Dredze, 2019; Pires et al., 2019). The good performance of mBERT has drawn the attention of many different NLP communities, and efforts have been made to produce BERT versions trained on monolingual data. This has lead to the release of BERT models in Russian (Kuratov & Arkhipov,

---

[*]Work partially performed while at Adereso.

2019), French (Martin et al., 2019; Le et al., 2019), Dutch (de Vries et al., 2019; Delobelle et al., 2020), Italian (Polignano et al., 2019), and Portugese (Souza et al., 2019).

In this paper we present the first BERT model pre-trained for the Spanish language. Despite Spanish being widely spoken (much more than the previously mentioned languages) finding resources to train or evaluate Spanish language models is not an easy task. For this reason we also compiled several Spanish-specific tasks in a single repository much in the spirit of the GLUE benchmark (Wang et al., 2019). By fine-tuning our pre-trained Spanish model we obtain better results compared to other BERT-based models pre-trained on multilingual corpora for most of the tasks, even achieving a new state-of-the-art on some of them. We have released our pre-trained model, the training corpus, and the compilation of benchmarks as free resources to be used by the Spanish NLP community[1]

In the rest of this paper we first present our Spanish-BERT model, then we describe the tasks that we have compiled into a benchmark that we call GLUES (GLUE for Spanish), and finally the results obtained by our model in some of the GLUES tasks. Before going into the details of our model and results we briefly review the related work.

2 RELATED WORK

Pre-trained language models using deep neural networks became very popular starting with ULM-Fit (Howard & Ruder, 2018). ULMFit is based on a standard recurrent neural network architecture and a language-modeling task (predicting the next token from the previous sequence of tokens). By using vast amounts of text, ULMFit is trained first for the language-modeling task trying that the model acquires *general knowledge* from a big corpus. The model is then fine-tuned in a supervised way to solve a specific task using labeled data. The empirical results showed that pre-training plus fine-tuning can considerably outperform training a model from scratch for the same supervised task.

A similar pre-training strategy was later used by Devlin et al. (2018) to propose the BERT model. Compared with ULM-Fit, in BERT the recurrent architecture is changed by self-attention (Vaswani et al., 2017) which allows the prediction of a token to be dependant on every other token in the same sequence. The task used for pre-trainig BERT, called *masked language modeling*, is based on corrupting an input sequence by arbitrary deleting some of the tokens and then training the model to reconstruct the original sequence (Devlin et al., 2018). Several variations of BERT in terms of the training method and the task used for pre-training have been proposed (Liu et al., 2019; Joshi et al., 2019; Yang et al., 2019b). There have also been efforts to make models more efficient in terms of number of parameters or training time (Sanh et al., 2019; Lan et al., 2019; Clark et al., 2019).

Wu & Dredze (2019) and Pires et al. (2019) studied Multilingual BERT models, that is, models pre-trained simultaneously on corpora from different languages. These works showed how a single model can learn from several languages, setting strong baselines for non-English tasks. XLM (Lample & Conneau, 2019) introduced a supervised objective which involved parallel multilingual data and XLM-RoBERTa (Conneau et al., 2019) got the multilingual models to the *big leagues* in terms of model size.

Several single-language BERT models came with results that usually got better performance than multilingual models as it is the case with CamemBERT (Martin et al., 2019) and FlauBERT (Le et al., 2019) for French, BERTje (de Vries et al., 2019) and RobBERT (Delobelle et al., 2020) for Dutch, FinBERT (Virtanen et al., 2019) for Finish, to name a few. Our work is similar to these but for the Spanish language. To the best of our knowledge, our paper presents the first open release of a Spanish BERT model and evaluation.

3 SPANISH-BERT MODEL, DATA AND TRAINING

We trained a model similar in size to a BERT-Base model (Devlin et al., 2018). Our model has 12 self-attention layers with 16 attention-heads each (Vaswani et al., 2017), using 1024 as hidden size. In total our model has 110M parameters. We trained two versions, one with cased data and one with uncased data, with a dataset that we next describe.

---

[1] https://github.com/dccuchile/beto

For training our model we collected text from different sources. We used all the data from Wikipedia and all of the sources of the OPUS Project (Tiedemann, 2012) that had text in Spanish. This sources includes United Nations and Government journals, TED Talks, Subtitles, News Stories and more. The total size of the corpora gathered was comparable with the corpora used in the original BERT. Our corpus for training has about 3 billion words and we release it for later use[2]. Our corpus can be considered as an updated version of the one compiled by Cardellino (2016).

For both versions of our model, cased and uncased, we constructed a vocabulary of 31K subwords using the byte pair encoding algorithm provided by the SentencePiece library (Kudo & Richardson, 2018). We added 1K place-holder tokens for later use which gave us a vocabulary of 32K tokens.

For training our BERT models we consider some details of training that have been successful in RoBERTa (Liu et al., 2019). In particular, we integrate the Dynamic Masking technique in our training which refers to using different masks for the same sentence in our corpus. The Dynamic Masking we used was 10x, which means that every sentence had 10 different masks. We also consider the Whole-Word Masking (WWM) technique from the updated version of BERT (Devlin et al., 2018). WWM ensures that when masking a specific token, if the token corresponds to a sub-word in a sentence, then all contiguous tokens conforming the same word are also masked. Also we trained on larger batches compared with the original BERT (but smaller than RoBERTa). We trained each model (cased and uncased) for 2M steps, with learning rate of 0.0001 and the first 10000 steps as warmup. Also the training was divided in two phases as described by You et al. (2019): we trained the first 900k steps with a batch size of 2048 and maximum sequence length of 128, and the rest of the training with batch size of 256 and maximum sequence length of 512. All the pre-training was done using Google's preemptible TPU v3-8.

4 GLUES

In this section we present the GLUES benchmark, a compilation of common NLP tasks in the Spanish language, following the idea of the original English GLUE benchmark (Wang et al., 2019). Through this benchmarks, we hope to help standardize future Spanish NLP efforts[3]. We next describe the tasks that we currently consider in GLUES.

Natural Language Inference: XNLI The Multi-Genre Natural Language Inference (MNLI) dataset (Williams et al., 2017) is a set of pair of sentences. The first sentence is called the premise, and the second the hipothesis. The task is to predict whether the premise entails the hypothesis (entailment), contradicts it (contradiction), or neither entails nor contradicts it (neutral). That is, the task is a 3-class classification. The Cross-Lingual Natural Language Inference (XNLI) corpus (Conneau et al., 2018) is an evaluation dataset that extends the MNLI by adding development and test set for 15 languages. In this setup we train using the Spanish machine translation of the MNLI dataset, and use the development and test set from the XNLI corpus. This task is evaluated by simple accuracy.

Paraphrasing: PAWS-X The PAWS-X (Yang et al., 2019a) is the multilingual version of the PAWS dataset (Zhang et al., 2019). The task consists in determining if two sentences are semantically equivalent or not. The dataset provides standard (translate) train, development and test set, and we used the Spanish portion. It is evaluated using simple accuracy.

Named Entity Recognition: CoNLL Named Entity Recognition consists in determining phrases in a sentence that contain the names of persons, organizations, locations, times and quantities. For this task we use the dataset by Tjong Kim Sang (2002) that focuses on persons, organizations, and locations with a fourth category of miscellaneous entities, all tagged using the BIO format (Ramshaw & Marcus, 1999). This dataset provides standard train, development and test sets and the performance is measured with F1 score.

Part-of-Speech Tagging: Universal Dependencies v1.4 A part of speech (POS) is a category of words with similar grammatical properties such as noun, verb, adjective, preposition and conjunc-

---

[2]https://github.com/josecannete/spanish-corpora
[3]See https://github.com/dccuchile/glues for a detailed description on how to obtain train-dev-test data for each task.

tion. For the POS tagging task we use the Spanish subset of the Universal Dependencies (v1.4) Treebank (Nivre et al., 2016). The version of the dataset is chosen following the works of Wu & Dredze (2019) and Kim et al. (2017). The dataset provides standard train, development and test sets. This task is evaluated by the accuracy of predicted POS tags.

Document Classification: MLDoc   The MLDoc dataset (Schwenk & Li, 2018) is a balanced subset of the Reuters corpus. This task consists in classifying the docuements into four categories, CCAT (Corporate/Industrial), ECAT (Economics), GCAT (Government/Social), and MCAT (Markets). This dataset provides multiple sizes for the train split (1k, 2k, 5k and 10k), plus standard development and test sets for eight languages. We chose to train using the largest available test split in Spanish. This task is evaluated using simple accuracy.

Dependency Parsing: Universal Dependencies v2.2   A dependency tree represents the grammatical structure of a sentence defining relationships in the form of labeled-edges from "dependent" to "head" words. The label of each edge represents the type of the relationship. The task of dependency parsing consists in constructing a dependency tree for a given sentence. For this task we use a subset of the Universal Dependencies v2.2 Treebank (Nivre et al., 2018), in particular, the concatenation of the AnCora and GSD Spanish portions of the dataset. This decision and the version we chose follow the work by Ahmad et al. (2019). The task is evaluated using the Unlabeled Attachment Score (UAS) and Labeled Attachment Score (LAS) (Kübler et al., 2009). UAS is the percentage of words that have been assigned the correct head, whereas LAS is the percentage of words that have been assigned both the correct head and the correct label for the relationship.

Question Answering: MLQA, TAR and XQuAD   Given a context and a question, the task of question answering (QA) consists in finding the sequence of contiguous words within the context that answers the question. This task is usually evaluated using two metrics averaged across all questions: exact match, that corresponds to the percentage of answers that match exactly, and F1 score, where answers are treated as bags of words. For our benchmark we consider three translated versions of the SQuAD v1.1 dataset (Rajpurkar et al., 2016), namely, MLQA (Lewis et al., 2019), XQuAD (Artetxe et al., 2019), and TAR (Carrino et al., 2019) that we next explain. MLQA provides translations for seven language with train data produced with machine translation from English, and development and test data translated by professionals (Lewis et al., 2019). XQuAD provides test set to evaluate cross-lingual models in 11 languages (Artetxe et al., 2019). Carrino et al. (2019) proposed the TAR method (Translate Align Retrieve) that can be used to produce automatic machine translated versions of QA datasets. They provided a TAR-translation from English to Spanish of SQuAD v1.1 (Carrino et al., 2019). In our benchmark we consider the train sets in MLQA and TAR, and the test sets in MLQA and XQuAD.

5   EVALUATION

5.1   FINE-TUNING

Since one of the goals of our work was to compare the performance of Spanish-BERT to mBERT (Wu & Dredze, 2019; Yang et al., 2019a), our experimental setup closely follows the one by Wu & Dredze (2019). Task specific output layers are incorporated following the original BERT work (Devlin et al., 2018).

For each task, no preprocessing is performed except tokenization from words into subwords using the learned vocabulary and WordPiece (Wu et al., 2016). We use the Adam optimizer (Kingma & Ba, 2014) for fine-tuning with standard parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$) and L2 weight decay of 0.01. We warm up the learning rate for the first 10% of steps, with a linear decay afterwards.

To allow for fine-tuning in a single GPU, we limit the length of each sentence to 128 tokens for all tasks. To accommodate for tasks that require word level classification we use the sliding window approach described by Wu & Dredze (2019). After the first window, the last 64 tokens are kept for the next window, and only 64 new tokens are fed into the model.

For the hyperparameter selection, we run experiments using different combinations of batch size, learning rate and number of epochs, following the values recommended by Devlin et al. (2018):

Model	XNLI	PAWS-X	NER	POS	MLDoc
Best mBERT	78.50^a	89.00^b	87.38^a	97.10^a	95.70^a
es-BERT uncased	80.15	**89.55**	82.67	98.44	**96.12***
es-BERT cased	**82.01**	89.05	**88.43**	**98.97***	95.60

Table 1: Comparison of Spanish BERT (es-BERT) with the best results obtained by multilingual BERT models where the fine tune was done only on the Spanish train data in every dataset. Superscript denotes results obtained by: (*a*) Wu & Dredze (2019) and (*b*) Yang et al. (2019a). The "*" indicates a new state-of-the-art resut in the respective Spanish benchmark.

Model	MLQA, MLQA	TAR, XQuAD	TAR, MLQA
Best mBERT	53.90 / 37.40^c	**77.60** / **61.80**^{d}	68.10 / **48.30**^{d}
es-BERT uncased	67.85 / **46.03**	77.52 / 55.46	68.04 / 45.00
es-BERT cased	**68.01** / 45.88	77.56 / 57.06	**69.15** / 45.63

Table 2: Comparison of Spanish BERT (es-BERT) with the best results obtained by multilingual BERT models in Question Answering. We only consider models where the fine tune was done on the Spanish train data in every dataset. Results are presented as F1 / ExactMatch. Every column header denotes the train set (left) and test set (right) used in everry case. Different superscript denotes results obtained by different authors: (*c*) Lewis et al. (2019) and (*d*) Carrino et al. (2019)

batch size {16, 32}; learning rate {5e-5, 3e-5, 2e-5}; number of epochs {2, 3, 4}. Extensive hyper-parameter search is left for future work.

5.2 RESULTS

Tables 1 and 2 show our results compared with the best mBERT result reported in the literature for the same setting. Spanish BERT outperforms most of the mBERT results, except for some QA settings (Table 2). One of the largest difference can be seen on the XNLI task, which is the task that has the largest training dataset in Spanish. We note that in two of the most standard Spanish datasets (POS and MLDoc) we obtained a new state of the art. We also note that there are some important differences in the performances in the QA datasets. The low quality of the machine translation in MLQA might be one possible reason. We observed that nearly half have of the 81K examples in MLQA has a mismatch between the answer and its starting position in the context.

It is important to notice that our models are *Spanish-only* and thus, cannot take advantage neither of the original English train set in every translated benchmark, nor from train data in other languages. Taking advantage of data in different languages is a capability that multilingual models have by design. Actually, there has been some recent work on big multilingual models that can obtain better results in the Spanish datasets when fine tuned with general, not necessarily Spanish data. This is the case of the XLM-RoBERTa model (Conneau et al., 2019) that with 560M parameters and consuming training data from different languages, obtained results of 85.6% for XNLI and 89% for NER in the Spanish test set. Both results are better to the one that we obtained by fine tuning with Spanish data only (Table 1). Similarly, Yang et al. (2019a) obtained 90.7% in the Spanish test-set of PAWS-X when fine tuning mBERT including the original English train set.

6 CONCLUSION

The advent of Transformer-based pre-trained language models has greatly improved the accessibility of the average user to high performing models. In this paper we successfully pre-train a Spanish-only model and open-source it together with the training corpus and evaluation benchmarks for the community to use.

The ease of use of a pre-trained NLP model implies that its use cases are much broader, given that practitioners from disciplines other than computer science could fine-tune them for their domain-

specific downstream tasks. By releasing our Spanish-BERT model we hope to encourage the research and applications of deep learning techniques in Spanish-speaking countries.

Another direction of the work is to improve Spanish NLP models in terms of size as well as memory and computation requirements. We are currently working on pre-training different ALBERT models (Lan et al., 2019) for Spanish, with number of parameters ranging from 5M to 220M. Our first results with the smaller models are encouraging and we plan to release these models as well.

REFERENCES

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. 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)*, pp. 2440–2452, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1253. URL https://www.aclweb.org/anthology/N19-1253.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*, 2019.

Cristian Cardellino. Spanish Billion Words Corpus and Embeddings, March 2016. URL https://crscardellino.github.io/SBWCE/.

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.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, 2019.

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*, 2019.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. Robbert: a dutch roberta-based language model. *arXiv preprint arXiv:2001.06286*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *arXiv preprint arXiv:1907.10529*, 2019.

---

[4]https://www.adere.so/

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. Cross-lingual transfer learning for POS tagging without cross-lingual resources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2832–2838, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1302. URL https://www.aclweb.org/anthology/D17-1302.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. Dependency parsing. *Synthesis lectures on human language technologies*, 1(1):1–127, 2009.

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.

Yuri Kuratov and Mikhail Arkhipov. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*, 2019.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. Flaubert: Unsupervised language model pre-training for french. *arXiv preprint arXiv:1912.05372*, 2019.

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

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

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*, 2019.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Yevgeni Berzak, Riyaz Ahmad Bhat, Eckhard Bick, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Fabricio Chalub, Çağrı Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Claudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Jan Hajič, Linh Hà Mỹ, Dag Haug, Barbora Hladká, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Jessica Kenney, Natalia Kotsyba, Simon Krek, Veronika Laippala, Lucia Lam, Phng Lê H`ông, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Keiko Sophie Mori, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Lng Nguy˜ên Thị, Huy`ên Nguy˜ên Thi Minh, Vitaly Nikolaev, Hanna Nurmi, Petya Osenova, Robert Östling, Lilja Øvrelid, Valeria Paiva, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Livy Real, Laura Rituma, Rudolf Rosa, Shadi Saleh, Baiba Saulīte, Sebastian Schuster, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó,

Mária Šimková, Kiril Simov, Aaron Smith, Carolyn Spadine, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Lars Wallin, Jing Xian Wang, Jonathan North Washington, Mats Wirén, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. Universal dependencies 1.4, 2016. URL `http://hdl.handle.net/11234/1-1827`. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, John Bauer, Sandra Bellato, Kepa Bengoetxea, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Rogier Blokland, Victoria Bobicev, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Carly Dickerson, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Tomaž Erjavec, Aline Etienne, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Radu Ion, Elena Irimia, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Tolga Kayadelen, Václava Kettnerová, Jesse Kirchner, Natalia Kotsyba, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê H`ông, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Niko Miekka, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Shinsuke Mori, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lng Nguy˜ên Thị, Huy`ên Nguy˜ên Thị Minh, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Adédayọ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Thierry Poibeau, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Michael Rießler, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Shoval Sadde, Shadi Saleh, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Muh Shohibussirri, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Yuta Takahashi, Takaaki Tanaka, Isabelle Tellier, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Veronika Vincze, Lars Wallin, Jonathan North Washington, Seyi Williams, Mats Wirén, Tsegay Woldemariam, Tak-sum Wong, Chunxiao Yan, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, Manying Zhang, and Hanzhi Zhu. Universal dependencies 2.2, 2018. URL `http://hdl.handle.net/11234/1-2837`. LINDAT/CLARIN digital library at the

Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *preprint arXiv 1802.05365*, 2018.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert? *CoRR*, abs/1906.01502, 2019. URL http://arxiv.org/abs/1906.01502.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481. CEUR, 2019. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074851349&partnerID=40&md5=7abed946e06f76b3825ae5e294ffac14.

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

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

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 (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-00-9.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*, 2019.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pp. 2214–2218, 2012.

Erik F. Tjong Kim Sang. 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)*, 2002. URL https://www.aclweb.org/anthology/W02-2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*, 2019.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. 2019. In the Proceedings of ICLR.

Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*, 2019.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

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

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pp. 5754–5764, 2019b.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. In *International Conference on Learning Representations*, 2019.

Yuan Zhang, Jason Baldridge, and Luheng He. Paws: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*, 2019.