SPANISH PRE-TRAINED BERT MODEL AND EVALUATION DATA

Anonymous authors
Paper under double-blind review

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 model on these tasks we achieve state-of-the-art results on several of them setting a new baselines for Spanish NLP. In particular our model outperforms other BERT-based models pre-trained on multilingual corpora. We released our model and the compilation of the Spanish benchmarks to be used by the NLP community.

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 for NLP (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 state-of-the-art 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, 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-like 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 Spanish-BERT model we achieve state-of-the-art results on several tasks. In particular our model outperforms mBERT on part-of-speech tagging, named-entity recognition, and natural-language inference. We release 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 shown 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. 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. 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.

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 training our BERT model we consider some details of training we saw 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

[1]The link to the resources is not provided for anonymity reasons.

[2]URL not provided for anonymity reasons

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 for 2M steps on each model, with batch size of 2048 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 idea of the original GLUE benchmark (Wang et al., 2019) for the english language. Through this benchmarks, we hope to guide and standardize future Spanish NLP efforts. We next describe the tasks that we currently consider in GLUES.

Natural Language Inference: XNLI The Cross-Lingual NLI Corpus (Conneau et al., 2018) is an evaluation dataset that extends the MNLI (Williams et al., 2017) dataset by adding development and test set for 15 languages. Given a premise sentence and a hypothesis sentences, the task is to predict whether the premise entails the hypothesis (entailment), contradicts the hypothesis (contradiction), or neither (neutral). That is, the task is a 3-class classification. 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. It is evaluated using simple accuracy.

Named Entity Recognition: CoNLL For this task we use the dataset by Tjong Kim Sang (2002). Named Entity Recognition consists in determining if each word in a sentence corresponds to an entity or not. Named entities are phrases that contain the names of persons, organizations, locations, times and quantities. This particular dataset focuses on the first three and adds a fourth category of miscellaneous entities. This dataset is tagged using the BIO schema. This dataset provides standard train, development and test sets and the performance in this task is measured with F1 score. For this task, Precision is the percentage of named entities found that are correct and Recall is the percentage of named entities present in the corpus that are found.

Part-of-Speech Tagging: Universal Dependencies v1.4 Part-of-speech (POS) tagging is the task of tagging a word in a text with its part of speech. A part of speech is a category of words with similar grammatical properties. Common Spanish parts of speech are noun, verb, adjective, adverb, pronoun, preposition, conjunction, etc. For this 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 (Schwenk & Li, 2018) dataset 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. We chose to train using the largest available test split. This task is evaluated using simple accuracy.

Dependency Parsing: Universal Dependencies v2.2 The task of dependency parsing consists in assigning a dependency tree to a given sentence. A dependency tree represents the grammatical structure of a sentence and defines the relationship between "head" words an "dependent" words which are associated to those heads. The relationship between the two words is expressed with an edge of the dependency tree and the type of relationship is represented by the label of said edge.

Model	XNLI	PAWS-X	NER	POS	MLDoc
Best mBERT results	78.50[a]	**90.70**[b]	87.38[a]	97.10[a]	95.70[a]
Spanish BERT uncased	80.15	89.55	82.67	98.44	**96.12**
Spanish BERT cased	**82.01**	89.05	**88.43**	**98.97**	95.60

Table 1: Comparing our Spanish-BERT with the best results achieved by alternative multilingual BERT models. Since GLUES is a spanish-only benchmark, no all previous works have evaluated mBERT in every GLUES task. Superscript a corresponds to the results obtained by Wu & Dredze (2019) and superscript b to the one obtained by Yang et al. (2019a).

For this task we use a subset of the Universal Dependencies v2.2 Treebank (Nivre et al., 2018). The spanish portion of this dataset consists of three subsets, Spanish_AnCora, Spanish_GSD and Spanish_PUD. We use the concatenation of the AnCora and GSD portions of the dataset. This decision and the version choice is made following the work from Ahmad et al. (2018).

This task is evaluated using the metrics UAS and LAS, which stand for Unlabeled Attachment Score and Labeled Attachment Score, respectively. 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 The MultiLingual Question Answering (MLQA) (Lewis et al., 2019) is a multilingual dataset for question answering derived from SQuAD (Rajpurkar et al., 2016). Given a context and a question, the task of question answering consists in finding the word or sequence of words within the context that answers the question.

This dataset set provides a train split translated from English, and new development and test sets for each language. This task is evaluated using Exact Match, that is, the percentage of answers that match exactly, and F1 score, where the prediction and ground truth are treated as bags of tokens, and compute their F1 score. Then this individual score is averaged across all questions. This task provides an evaluation script along the dataset.

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 from Wu & Dredze (2019). Task specific output layers are incorporated following the original work from Devlin et al. (2018).

For each task, no preprocessing is performed except tokenization from words into subwords using the learned vocabulary and WordPiece. We use Adam (Kingma & Ba, 2014) for fine-tuning with β_1 of 0.9, β_2 of 0.999 and L2 weight decay of 0.01. We warm up the learning rate over the first 10% of steps and then linearly decay the learning rate.

In order to be able to fine-tune in one 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 in 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 recommended values from Devlin et al. (2018): batch size {16, 32}; learning rate {5e-5, 3e-5, 2e-5}; number of epochs {2, 3, 4}.

5.2 Results

In Table 1 we present our results in comparison to other results obtained by independent works using mBERT. Spanish BERT outperforms most of the results previously obtained, except in PAWS-

X dataset, and our results also surpass those from XLM (Lample & Conneau, 2019) and UDPipe (Kondratyuk, 2019). The largest difference can be seen on the XNLI task, which is the task that has the largest training dataset. The best result in the PAWS-X task is obtained by using the training from multiple languages availables in the task and fine-tuning mBERT using this merged dataset. This is posible because mBERT is a multilingual model, but since Spanish BERT is a single language model we fine-tuned and evaluated using only the spanish portion of the dataset.

It is important to mention that our model wasn't able to surpass the state of the art in the NER task stablished by Straková et al. (2019). We don't include their result in our comparison because they designed an architecture specifically tailored for the NER task. Future work could involve replacing their embedding layer with Spanish-BERT to provide a more accurate comparison.

At the time of writing, we are still performing the last experiments on Dependency Parsing and Question Answering. Nevertheless, the current results show a strong advantage from our single language model over previous multi-language models in Spanish tasks.

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.

References

Wasi Uddin 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, 2018.

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

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.

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, 2014.

Daniel Kondratyuk. 75 languages, 1 model: Parsing universal dependencies universally. *arXiv preprint arXiv:1904.02099*, 2019.

Yuri Kuratov and Mikhail Arkhipov. Adaptation of deep bidirectional multilingual transformers for russian language, 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 Thị 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, Adédayọ̀ 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, 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, 2016.

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, 2019.

Jana Straková, Milan Straka, and Jan Hajič. Neural architectures for nested ner through linearization, 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.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. Paws-x: A cross-lingual adversarial dataset for paraphrase identification, 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.

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