FinEst BERT and CroSloEngual BERT

Less Is More in Multilingual Models

Matej Ulčar(✉) and Marko Robnik-Šikonja

Faculty of Computer and Information Science, University of Ljubljana,
Večna pot 113, Ljubljana, Slovenia
{matej.ulcar,marko.robnik}@fri.uni-lj.si

Abstract. Large pretrained masked language models have become state-of-the-art solutions for many NLP problems. The research has been mostly focused on English language, though. While massively multilingual models exist, studies have shown that monolingual models produce much better results. We train two trilingual BERT-like models, one for Finnish, Estonian, and English, the other for Croatian, Slovenian, and English. We evaluate their performance on several downstream tasks, NER, POS-tagging, and dependency parsing, using the multilingual BERT and XLM-R as baselines. The newly created FinEst BERT and CroSloEngual BERT improve the results on all tasks in most monolingual and cross-lingual situations.

Keywords: Contextual embeddings · BERT model · Less-resourced languages · NLP

1 Introduction

In natural language processing (NLP), a lot of research focuses on numeric word representations. Static pretrained word embeddings like word2vec [12] are recently replaced by dynamic, contextual embeddings, such as ELMo [14] and BERT [4]. These generate a word vector based on the context the word appears in, mostly using the sentence as the context.

Large pretrained masked language models like BERT [4] and its derivatives achieve state-of-the-art performance when fine-tuned for specific NLP tasks. The research into these models has been mostly limited to English and a few other well-resourced languages, such as Chinese Mandarin, French, German, and Spanish. However, two massively multilingual masked language models have been released: a multilingual BERT (mBERT) [4], trained on 104 languages, and newer even larger XLM-RoBERTa (XLM-R) [3], trained on 100 languages. While both, mBERT and XLM-R, achieve good results, it has been shown that monolingual models significantly outperform multilingual models [11,20]. Arkhipov et al. (2019) [2] trained a four language (Russian, Bulgarian, Polish, Czech) BERT model by bootstrapping mBERT. They reported improvements over mBERT on named entity recognition task.

P. Sojka et al. (Eds.): TSD 2020, LNAI 12284, pp. 104–111, 2020.
https://doi.org/10.1007/978-3-030-58323-1_11

In our work, we reduced the number of languages in multilingual models to three, two similar less-resourced languages from the same language family, and English. The main reasons for this choice are to better represent each language, and keep sensible sub-word vocabulary, as shown by Virtanen et al. (2019) [20]. We decided against production of monolingual models, because we are interested in using the models in multilingual sense and for cross-lingual knowledge transfer. By including English in each of the two models, we expect to better transfer existing prediction models from English to involved less-resourced languages. Additional reason against purely monolingual models for less-resourced languages is the size of training corpora, i.e. BERT-like models use transformer architecture which is known to be data hungry.

We thus trained two multilingual BERT models: FinEst BERT was trained on Finnish, Estonian, and English, while CroSloEngual BERT was trained on Croatian, Slovenian, and English. In the paper, we present the creation and evaluation of these models, which required considerable computational resources, unavailable to most NLP researchers. We make the models which are valuable resources for the involved less-resourced languages publicly available[1].

2 Training Data and Preprocessing

BERT models require large quantities of monolingual data. In Sect. 2.1 we first describe the corpora used, followed by a short description of their preprocessing in Sect. 2.2.

2.1 Datasets

To obtain high-quality models, we used large monolingual corpora for each language, some of them unavailable to the general public. High-quality English language models already exist and English is not the main focus of this research, we therefore did not use all available English corpora in order to prevent English from overwhelming the other languages in our models. Some corpora are available online under permissive licences, others are available only for research purposes or have limited availability. The corpora used in training are a mix of news articles and general web crawl, which we preprocessed and deduplicated. Details about the training set sizes are presented in Table 1, while their description can be found in works on the involved less-resourced languages, e,g., [18].

2.2 Preprocessing

Before using the corpora, we deduplicated them for each language separately, using the Onion (ONe Instance ONly) tool[2]. We applied the tool on sentence

[1] CroSloEngual BERT: http://hdl.handle.net/11356/1317
FinEst BERT: http://urn.fi/urn:nbn:fi:lb-2020061201.

[2] http://corpus.tools/wiki/Onion.

Table 1. The training corpora sizes in number of tokens and the ratios for each language.

Model	CroSloEngual	FinEst
Croatian	31%	0%
Slovenian	23%	0%
English	47%	63%
Estonian	0%	13%
Finnish	0%	25%
Tokens	$5.9 \cdot 10^9$	$3.7 \cdot 10^9$

Table 2. The sizes of corpora subsets in millions of tokens used to create wordpiece vocabularies.

Language	FinEst	CroSloEngual
Croatian	/	27
Slovenian	/	28
English	157	23
Estonian	75	/
Finnish	97	/

level for those corpora that did have sentences shuffled, and on paragraph level for the rest. As parameters, we used 9-grams with duplicate content threshold of 0.9.

BERT models are trained on subword (wordpiece) tokens. We created a wordpiece vocabulary using bert-vocab-builder tool[3], which is built upon tensor2tensor library [19]. We did not process the whole corpora in creating the wordpiece vocabulary, but only a smaller subset. To balance the language representation in vocabulary, we used samples from each language. The sizes of corpora subsets are shown in Table 2. The created wordpiece vocabularies contain 74,986 tokens for FinEst and 49,601 tokens for CroSloEngual model.

3 Architecture and Training

We trained two BERT multilingual models. FinEst BERT was trained on Finnish, Estonian, and English corpora, with altogether 3.7 billion tokens. CroSloEngual BERT was trained on Croatian, Slovenian, and English corpora with together 5.9 billion tokens.

Both models use bert-base architecture [4], which is a 12-layer bidirectional transformer encoder with the hidden layer size of 768 and altogether 110 million parameters. We used the whole word masking for the masked language model training task. Both models are cased, i.e. the case information was preserved. We followed the hyper-parameters settings of Devlin et al. (2018) [4], except for the batch size and total number of steps. We trained the models for approximately 40 epochs with maximum sequence length of 128 tokens, followed by approximately 4 epochs with maximum sequence length of 512 tokens. The exact number of steps was calculated using the expression $s = \frac{N_{tok} \cdot E}{b \cdot \lambda}$, where s is the number of steps the models were trained for, N_{tok} is the number of tokens in the train corpora, E is the desired number of epochs (in our case 40 and 4), b is the batch size, and λ is the maximum sequence length.

We trained FinEst BERT on a single Google Cloud TPU v3 for a total of 1.24 million steps where the first 1.13 million steps used the batch size of 1024

[3] https://github.com/kwonmha/bert-vocab-builder.

and sequence length 128, and the last 113 thousand steps used the batch size 256 and sequence length 512. Similarly, CroSloEngual BERT was trained on a single Google Cloud TPU v2 for a total of 3.96 million steps, where the first 3.6 million steps used the batch size of 512 and sequence length 128, and the last 360 thousand steps were trained with the batch size 128 and sequence length 512. Training took approximately 2 weeks for FinEst BERT and approximately 3 weeks for CroSloEngual BERT.

4 Evaluation

We evaluated the two new BERT models on sensible languages and three downstream evaluation tasks available for the four involved less-resourced languages: named entity recognition (NER), part-of-speech tagging (POS), and dependency parsing (DP). We compared both models with BERT-base-multilingual-cased model (mBERT). On the NER task we compared also XLM-RoBERTa (XLM-R) and Finnish BERT (FinBERT).

4.1 Named Entity Recognition

NER is a sequence labeling task, which tries to correctly identify and classify each token from an unstructured text into one of the predefined named entity (NE) classes, or as not NE. The publicly available NER datasets for the involved languages that we used have only three NE classes in common. To allow a more direct comparison between languages, we reduced them to the four labels in common: *person*, *location*, *organization*, and *other*. All tokens, which are not NE or belong to any other NE class were labeled as *other*.

For Croatian and Slovenian, we used NER data from hr500k [10] and ssj500k [8], respectively. Not all sentences in Slovenian ssj500k are annotated, so we excluded those that are not annotated. The English dataset comes from the CoNLL 2013 shared task [17]. For Finnish we used the Finnish News Corpus for NER [15], and as the Estonian dataset we used the Nimeüksuste korpus [9].

The implementation uses the Huggingface's Transformer library v2.8, and our code is based on its NER example[4]. We fine-tuned each of our BERT models with an added token classification head for 3 epochs on the NER data. We compared the results with mBERT, XLM-R and FinBERT models, which we fine-tuned with exactly the same parameters on the same data. We used maximum sequence length of 512 and batch size of 6 for all models and languages.

We evaluated the models in a monolingual setting (training and testing on the same language), and cross-lingual setting (training on one language, testing on another). We present the results as macro average F_1 scores of the three NE classes, excluding *other* label. Results are shown in Table 3.

In monolingual setting, the differences in performance of tested models on English data is negligible. In other languages, our models outperform both the

[4] https://github.com/huggingface/transformers/tree/v2.8.0/examples/ner.

Table 3. The results of NER evaluation task. The scores are macro average F_1 scores of the three NE classes. NER models were fine-tuned from mBERT(mB), CroSloEngual BERT (CSE), FinEst BERT (FE), XLM-RoBERTa (XR), and FinBERT (FB).

Train	Test	mB	CSE	XR
Croatian	Croatian	0.790	0.884	0.817
Slovenian	Slovenian	0.897	0.920	0.914
English	English	0.939	0.944	0.937
Croatian	English	0.807	0.868	0.773
English	Croatian	0.602	0.799	0.641
Slovenian	English	0.745	0.845	0.747
English	Slovenian	0.708	0.833	0.739
Croatian	Slovenian	0.810	0.891	0.855
Slovenian	Croatian	0.765	0.849	0.786

Train	Test	mB	FE	XR	FB
Finnish	Finnish	0.933	0.957	0.930	0.954
Estonian	Estonian	0.898	0.927	0.908	0.876
English	English	0.939	0.945	0.937	0.922
Finnish	English	0.688	0.812	0.722	0.573
English	Finnish	0.764	0.900	0.823	0.817
Estonian	English	0.774	0.816	0.755	0.641
English	Estonian	0.783	0.832	0.794	0.523
Finnish	Estonian	0.798	0.880	0.825	0.529
Estonian	Finnish	0.819	0.914	0.869	0.823

mBERT and XLM-R, the difference is especially large in Croatian. FinEst BERT performs on par with FinBERT on Finnish. In cross-lingual setting, both FinEst and CroSloEngual BERT show a significant improvement over both mBERT and XLM-R. This leads us to believe that multilingual BERT models with fewer languages are more suitable for cross-lingual knowledge transfer.

4.2 Part-of-Speech Tagging and Dependency Parsing

Next, we evaluated the created BERT models on two more syntactic classification tasks: POS-tagging and DP. In the POS-tagging task, we predict the grammatical category of each token (verb, adjective, punctuation, adverb, noun, etc). DP models predict the tree structure, representing the syntactic relations between words in a given sentence.

We trained classifiers on universal dependencies (UD) treebank datasets, using universal part-of-speech (UPOS) tag set. For Croatian, we used the dataset of Agic and Ljubesic (2015) [1]; for English, we used A Gold Standard Dependency Corpus [16], and for Estonian we used Estonian Dependency Treebank [13], converted to UD. The Finnish treebank used is based on the Turku Dependency Treebank [6]. Slovenian treebank [5] is based on the ssj500k corpus [8].

We used Udify tool [7] to train both POS tagger and DP classifiers at the same time. We fine-tuned each BERT model for 80 epochs on the treebank data, keeping the tool parameters at default values, except for "warmup_steps" and "start_step" values, which we changed to the number of training batches in one epoch.

We present the results of POS tagging as UPOS accuracy in Table 4. In the monolingual setting, the differences in performance between different BERT models are small for this task. FinEst and CroSloEngual BERTs perform slightly better than mBERT on all languages, except Croatian, where mBERT and CroSloEngual BERT are equal. On Finnish, FinBERT (acc = 0.984) slightly outperforms FinEst BERT (acc = 0.981). The differences are more pronounced in cross-lingual setting. When training on Slovenian, Finnish, or Estonian and

Table 4. The performance on the UD POS-tagging task, using UPOS accuracy for CroSloEngual BERT (CSE), FinEst BERT, and mBERT.

Train	Test	mBERT	CSE
Croatian	Croatian	0.983	0.983
English	English	0.969	0.972
Slovenian	Slovenian	0.987	0.991
English	Croatian	0.876	0.869
English	Slovenian	0.857	0.859
Croatian	English	0.750	0.756
Croatian	Slovenian	0.917	0.934
Slovenian	English	0.686	0.723
Slovenian	Croatian	0.920	0.935

Train	Test	mBERT	FinEst
English	English	0.969	0.970
Estonian	Estonian	0.972	0.978
Finnish	Finnish	0.970	0.981
English	Estonian	0.852	0.878
English	Finnish	0.847	0.872
Estonian	English	0.688	0.808
Estonian	Finnish	0.872	0.913
Finnish	English	0.535	0.701
Finnish	Estonian	0.888	0.919

testing on English, CroSloEngual and FinEst BERT significantly outperform mBERT. The exception is training on English and testing on Croatian, where mBERT outperforms CroSloEngual BERT.

We present the results of DP task with two metrics, the unlabeled attachement score (UAS) and labeled attachment score (LAS). In the monolingual setting, CroSloEngual BERT shows improvement over mBERT on all three languages (Table 5) with the highest improvement on Slovenian and only a marginal improvement on English. FinEst BERT outperforms mBERT on Estonian and Finnish, with the biggest margin being on the Finnish data, while the two models perform equally on English data. FinBERT again outperforms FinEst on Finnish, scoring UAS = 0.946 and LAS = 0.930.

In the cross-lingual setting, the results are similar to those seen on the POS tagging task. Major improvements of FinEst and CroSloEngual BERT over mBERT are observed in English-Estonian, English-Finnish and English-Slovenian pairs, minor improvements in Estonian-Finnish and Croatian-Slovenian pairs, while on English-Croatian pair mBERT outperformed CroSlo-Engual BERT.

Table 5. The results on the DP task presented with UAS and LAS scores for CroSlo-Engual BERT, FinEst BERT, and mBERT.

		mBERT		CroSloEngual	
Train	Test	UAS	LAS	UAS	LAS
Croatian	Croatian	0.930	0.891	0.940	0.903
English	English	0.917	0.894	0.922	0.899
Slovenian	Slovenian	0.938	0.922	0.957	0.947
English	Croatian	0.824	0.724	0.822	0.725
English	Slovenian	0.830	0.719	0.848	0.736
Croatian	English	0.759	0.627	0.782	0.657
Croatian	Slovenian	0.880	0.802	0.912	0.840
Slovenian	English	0.741	0.578	0.794	0.648
Slovenian	Croatian	0.861	0.773	0.891	0.810

		mBERT		FinEst	
Train	Test	UAS	LAS	UAS	LAS
English	English	0.917	0.894	0.918	0.895
Estonian	Estonian	0.880	0.848	0.909	0.882
Finnish	Finnish	0.898	0.867	0.933	0.915
English	Estonian	0.697	0.531	0.768	0.591
English	Finnish	0.706	0.561	0.781	0.624
Estonian	English	0.633	0.492	0.726	0.567
Estonian	Finnish	0.784	0.695	0.864	0.801
Finnish	English	0.543	0.433	0.684	0.558
Finnish	Estonian	0.782	0.691	0.852	0.778

5 Conclusion

We built two large pretrained trilingual BERT-based masked language models, Croatian-Slovenian-English and Finnish-Estonian-English. We showed that the new CroSloEngual and FinEst BERTs perform substantially better than massively multilingual mBERT on the NER task in both monolingual and cross-lingual setting. The results on POS tagging and DP tasks show considerable improvement of the proposed models for several monolingual and cross-lingual pairs, while they are never worse than mBERT.

In future, we plan to investigate different combinations and proportions of less-resourced languages in creation of pretrained BERT-like models, and use the newly trained BERT models on the problems of news media industry.

Acknowledgments. The work was partially supported by the Slovenian Research Agency (ARRS) core research programme P6-0411. This paper is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). Research was supported with Cloud TPUs from Google's TensorFlow Research Cloud (TFRC).

References

1. Agić, Ž., Ljubešić, N.: Universal dependencies for Croatian (that work for Serbian, too). In: The 5th Workshop on Balto-Slavic Natural Language Processing, pp. 1–8 (2015)
2. Arkhipov, M., Trofimova, M., Kuratov, Y., Sorokin, A.: Tuning multilingual transformers for language-specific named entity recognition. In: Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, pp. 89–93. Association for Computational Linguistics, Florence, August 2019. https://doi.org/10.18653/v1/W19-3712
3. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116 (2019)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Dobrovoljc, K., Erjavec, T., Krek, S.: The universal dependencies treebank for Slovenian. In: Proceeding of the 6th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017) (2017)
6. Haverinen, K., et al.: Building the essential resources for Finnish: the Turku dependency treebank. LREC **48**, 493–531 (2013)
7. Kondratyuk, D., Straka, M.: 75 languages, 1 model: parsing universal dependencies universally. In: Proceedings of the 2019 EMNLP-IJCNLP, pp. 2779–2795 (2019)
8. Krek, S., et al.: Training corpus ssj500k 2.2 (2019). Slovenian language resource repository CLARIN.SI
9. Laur, S.: Nimeüksuste korpus. Center of Estonian Language Resources (2013)
10. Ljubešić, N., Klubička, F., Agić, Ž., Jazbec, I.P.: New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian. In: Proceedings of the LREC 2016 (2016)

11. Martin, L., et al.: CamemBERT: a tasty French language model. arXiv preprint arXiv:1911.03894 (2019)
12. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. arXiv preprint 1309.4168 (2013)
13. Muischnek, K., Müürisep, K., Puolakainen, T.: Estonian dependency treebank: from constraint grammar tagset to universal dependencies. In: Proceedings of LREC 2016 (2016)
14. Peters, M.E., et al.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018)
15. Ruokolainen, T., Kauppinen, P., Silfverberg, M., Lindén, K.: A Finnish news corpus for named entity recognition. Lang. Res. Eval. **54**(1), 247–272 (2020)
16. Silveira, N., et al.: A gold standard dependency corpus for English. In: Proceedings of LREC-2014 (2014)
17. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In: Daelemans, W., Osborne, M. (eds.) Proceedings of CoNLL-2003, Edmonton, Canada, pp. 142–147 (2003)
18. Ulčar, M., Robnik-Šikonja, M.: High quality ELMo embeddings for seven less-resourced languages. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 4731–4738. European Language Resources Association, Marseille, May 2020
19. Vaswani, A., et al.: Tensor2tensor for neural machine translation. In: Proceedings of the AMT, pp. 193–199 (2018)
20. Virtanen, A., et al.: Multilingual is not enough: BERT for Finnish. arXiv preprint arXiv:1912.07076 (2019)