Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

doi:10.3233/FAIA200597

Evaluating Multilingual BERT for Estonian

Claudia KITTASK [1] and Kirill MILINTSEVICH and Kairit SIRTS
Institute of Computer Science, University of Tartu, Estonia

Abstract. Recently, large pre-trained language models, such as BERT, have reached state-of-the-art performance in many natural language processing tasks, but for many languages, including Estonian, BERT models are not yet available. However, there exist several multilingual BERT models that can handle multiple languages simultaneously and that have been trained also on Estonian data. In this paper, we evaluate four multilingual models—multilingual BERT, multilingual distilled BERT, XLM and XLM-RoBERTa—on several NLP tasks including POS and morphological tagging, NER and text classification. Our aim is to establish a comparison between these multilingual BERT models and the existing baseline neural models for these tasks. Our results show that multilingual BERT models can generalise well on different Estonian NLP tasks outperforming all baselines models for POS and morphological tagging and text classification, and reaching the comparable level with the best baseline for NER, with XLM-RoBERTa achieving the highest results compared with other multilingual models.

Keywords. multilingual BERT, NER, POS tagging, text classification, Estonian

1. Introduction

Large pretrained language models, also called contextual word embeddings, such as ELMo [1] or BERT [2] have been shown to improve many natural language processing tasks. Training large contextual language models is complex both in terms of the required computational resources as well as the training process and thus, the number of languages for which the pretrained models are available is still limited.

Although according to [3], language-specific BERT models are currently available for 19 languages, many more languages are supported via multi-lingual models. The aim of the multilingual models is to reduce the necessity to train language-specific models for each language separately. Experiments on various tasks, such as named entity recognition (NER) [4] or parsing pipeline tasks [5], have shown that multilingual contextual models can help to improve the performance over the baseline models not based on contextual word embeddings.

There are several multilingual models available that also include Estonian language. For instance, multilingual BERT (mBERT) [2] has been trained jointly on Wikipedia data on 104 languages, including Estonian. Estonian is also included in the cross-lingual language model (XLM-100) [6], which was trained on 100 Wikipedia languages, and

[1]Corresponding Author: Claudia Kittask; E-mail: claudiakittask@gmail.com

cross-lingual RoBERTa (XLM-RoBERTa) [4], which was trained on much larger CommonCrawl corpora and also includes 100 languages. Finally, DistilBERT [7] is a smaller version of the BERT model obtained from the BERT models via knowledge distillation, which is a compression technique where the compact model is trained to reproduce the behaviour of the larger model. The multilingual DistilBERT (DistilmBERT) has been distilled from the mBERT model featuring the same 104 Wikipedia languages.

The aim of the current work is to evaluate the existing multilingual BERT models on several NLP tasks on Estonian. In particular, we will apply the BERT models on NER, POS and morphological tagging, and text classification tasks. We compare four multilingual models—mBERT, XLM-100, XLM-RoBERTa and DistilmBERT—to find out which one of those performs the best on our Estonian tasks. We compare the results of the multilingual BERT models with task-specific baselines and show that multilingual BERT models improve the performance of the Estonian POS and morphological tagging and text classification tasks and achieve comparable results for named entity recognition. Overall, XLM-RoBERTa achieves the best results compared with other multilingual BERT models used.

2. Related Work

Although most research on multilingual BERT models has been concerned about zero-shot cross-lingual transfer [8], we are more interested in those previous works that, similar to us, evaluate multilingual BERT models in comparison to monolingual (non-English) baselines. We next review some examples of such work.

Virtanen et al. [9] evaluated multilingual BERT alongside with the monolingual Finnish BERT on several NLP tasks. In their work, multilingual BERT models outperformed monolingual baselines for text classification and NER tasks, while for POS-tagging and dependency parsing, the multilingual BERT models fell behind the previously proposed methods, most of which were utilizing monolingual contextual ELMo embeddings [1]. Baumann [10] evaluated multilingual BERT models on German NER task and found that while the multilingual BERT models outperformed two non-contextual LSTM-CRF-based baselines, it performed worse than a model utilizing monolingual contextual character-based string embeddings [11]. Kuratov et al. [12] applied multilingual BERT models on several tasks in Russian. They found that multilingual BERT outperformed non-contextual baselines for paraphrase identification and question answering and fell below a baseline for sentiment classification.

The pattern in all these works is similar: the multilingual BERT models perform better than non-neural or non-contextual neural baselines, but the multilingual BERT model is typically outperformed by approaches based on language-specific monolingual contextual comparison systems. We cannot test the second part of this observation as currently no monolingual language-specific BERT model exists for Estonian. However, we will show that the first part of this observation generally also holds for Estonian, i.e. the multilingual BERT models outperform non-contextual baselines for most of the experimental tasks used in this paper.

3. Experimental Tasks

This section describes the experimental tasks. We give also overview of the used data and the baseline models.

3.1. POS and Morphological Tagging

For POS and morphological tagging, we use the Estonian treebank from the Universal Dependencies (UD) v2.5 collection that contains annotations of lemmas, part of speech, universal morphological features, dependency heads and universal dependency labels. We train models to predict both universal POS (UPOS) and language-specific POS (XPOS) tags as well as morphological tags. We use the pre-defined train/dev/test splits for training and evaluation. Table 1 shows the statistics about the treebank splits.

Table 1. Statistics for the Estonian UD corpus

	Train	**Dev**	**Test**
Sentences	31,012	3,128	6,348
Tokens	344,646	42,722	48,491

As baselines, we report the results of Stanza [13] and UDPipe [14] obtained on the same Estonian UD v2.5 test set.

3.2. Article Type and Sentiment Classification

For text classification, we use the Estonian Valence corpus [15], which consists of 4088 paragraphs obtained from Postimees daily. The corpus has been annotated with sentiment as well as with rubric labels. The statistics of this dataset are given in Table 2. We split the data into training, testing and development set using 70/20/10 split preserving the ratios of different labels in the splits. All duplicates were removed from the corpus. In total, there were 17 duplicate paragraphs. We followed the suit of Pajupuu et al. [15] and removed the paragraphs annotated as ambiguous from the corpus. These paragraphs were shown to considerably lower the accuracy of the classification.

Table 2. Statistics of the Estonian Valence corpus

	Negative	**Ambiguous**	**Positive**	**Neutral**	**Total**
Opinion	429	242	162	139	972
Estonia	152	41	93	133	419
Life	138	47	207	128	520
Comments-Life	347	40	79	41	507
Comments-Estonia	368	27	50	56	501
Crime	170	12	11	16	209
Culture	57	40	86	79	262
Sports	76	81	152	76	385
Abroad	190	22	42	59	313
Total	1,927	552	882	727	4,088

For baseline, we trained supervised fastText classifiers [16] with pretrained fastText Wiki embeddings. The best hyperparameter values were found using the built-in fastText hyperparameter optimization.

3.3. Named Entity Recognition

The available Estonian NER corpus was created by Tkachenko et al. [17]. The corpus annotations cover three types of named entities: locations, organizations and persons. It contains 572 news stories published in local online newspapers Postimees and Delfi covering local and international news on a range of different topics. We split the data into training, testing and development set using 80/10/10 splits while preserving the document boundaries. Table 3 shows statistics of the splits.

Table 3. Statistics of the Estonian NER corpus

	Sentences	**Tokens**	**PER**	**LOC**	**ORG**	**Total**
Train	9,965	155,981	6,174	4,749	4,784	15,707
Dev	2,429	32,890	1,115	918	742	2,775
Test	1,908	28,370	1,201	644	619	2,464

As baselines, we report the performance of the CRF model [17] and the bilinear LSTM sequence tagger that was adapted from the Stanza POS tagger [13]. The tagger was trained on the NER annotations instead of POS tags, and the input was enriched with both POS tags and morphological features, i.e. the input to the NER model was the concatenation of the word, and its POS and morphological tag embeddings. The POS and morphological tags were predicted with the pre-trained Stanza POS tagger. The entity level performance is evaluated using the conlleval script from CoNLL-2000 shared task.

4. Experimental Setup

We conduct experiments with four different multilingual BERT models: multilingual cased BERT-base (mBERT), multilingual cased DistilBERT (DistilmBERT), cased XLM-100 and cross-lingual RoBERTa (XLM-RoBERTa). All these models are available via Hugging Face transformers library[2]. Each model is available with sequence lengths of 128 and 512 and we experiment with both. Table 4 shows some details of the models.

Table 4. Details of multilingual BERT models (all cased)

	Languages	**Vocab size**	**Parameters**
mBERT	104	119K	172M
XLM-100	100	200K	570M
DistilmBERT	100	119K	66M
XLM-RoBERTa	100	250K	270M

[2] `https://huggingface.co/transformers/`

To evaluate the performance of the multilingual BERT models on downstream tasks, we fine-tune all four BERT models for the NLP tasks described in Section 3. In addition to training the task-specific classification layer, we also fine-tune all BERT model parameters as well. For data processing and training, we used the scripts publicly available in the Hugging Face transformers repository. We tune the learning rate of the AdamW optimizer and batch size for each multilingual model and task on the development set using grid search. The learning rate was searched from the set of (5e-5, 3e-5, 1e-5, 5e-6, 3e-6). Batch size was chosen from the set of (8, 16). We find the best model for each learning rate and batch size combination by using early stopping with patience of 10 epochs on the development set.

5. Results

In subsequent sections, we present the experimental results on all multilingual BERT models for POS and morphological tagging, text classification and named entity recognition tasks.

5.1. POS and Morphological Tagging

The results for POS and morphological tagging are summarized in Table 5. In general, all tested multilingual BERT models are equally good and perform better than the Stanza and UDPipe baselines. DistilmBERT was the only multilingual model that did not exceed the baseline models results. On the other hand, the XLM-RoBERTa stands out with a small but consistent improvement over all other results displayed. Results also show that the sequence length of the model does not affect the performance in any way. The performance on XPOS is better than on UPOS. This is probably caused by the difference in the POS tag annotation schemes.

5.2. Text Classification

The sentiment and rubric classification task results are shown in Table 6. Multilingual models can easily outperform baseline fastText model. Similarly to POS and morphological tagging tasks, XLM-RoBERTa achieved the highest and DistilmBERT the lowest results overall. Even though there are more labels the in rubric classification task, it is still easier for the models to correctly classify than the sentiment classification task. Compar-

Table 5. POS and morphological tagging accuracy on Estonian UD test set.

Model	**UPOS**	**XPOS**	**Morph**	**UPOS**	**XPOS**	**Morph**
		Seq = 128			**Seq = 512**	
mBERT	97.42	98.06	96.24	97.43	98.13	96.13
DistilmBERT	97.22	97.75	95.40	97.12	97.78	95.63
XLM-100	97.60	98.19	**96.57**	97.59	98.06	96.54
XLM-RoBERTa	**97.78**	**98.36**	96.53	**97.80**	**98.40**	**96.69**
Stanza [13]	97.19	98.04	95.77			
UDPipe [14]	95.7	96.8	93.5			

Table 6. Rubric and sentiment classification accuracy

Model	Rubric	Sentiment	Rubric	Sentiment
	Seq = 128		Seq = 512	
mBERT	75.67	70.23	74.94	69.52
DistilmBERT	74.57	65.95	74.93	66.95
XLM-100	76.78	73.50	77.15	71.51
XLM-RoBERTa	**80.34**	**74.50**	**78.62**	**76.07**
fastText	71.01	66.76		

ison between the models with different sequence lengths is inconclusive—in some cases, the models with longer sequence are better, but not always.

5.3. Named Entity Recognition

The Table 7 (left) summarizes the NER results. We find that the task-specific StanfordNLP model is superior to all the multilingual BERT models, while XLM-100 and XLM-RoBERTa perform the best compared with other multilingual models. CRF based model was easily outperformed by all multilingual models except for DistilmBERT.

While performing these experiments, each sentence was treated as one sequence. This may have not optimally used the maximum sequence length available, especially in models with sequence length 512. As most sentences in our NER corpus do not reach the maximum length, we hypothesize that using longer sequences with the models of sequence length 512 would add more context for the model and thus improve the results. For that, we concatenate sentences from the same document to reach to the maximum 512 wordpiece sequence. The right-most section of the Table 7 shows the results of the experiments with longer input sequences. The numbers in the table show that concatenating the input sequences does not boost the scores. Compared with the regular results based on single sentences, only XLM-RoBERTa was able to utilize the maximum sequence length while the scores of other models decreased. The performance of the XLM-100 model suffered the most and obtained even lower results than DistilmBERT, which so far has gotten the lowest results in all tasks.

One possible reason why the multilingual BERT models were not able to improve over the Stanford tagger based NER model is that the Stanford baseline model makes use of the POS and morphological information while the BERT models do not. Adding

Table 7. NER tagging results. The right-hand part of the table shows the results with the models of sequence length 512, with the input sentences concatenated into sequences of maximum length

Model	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall	F1
	Seq = 128			Seq = 512			Seq = Concatenated		
mBERT	85.88	87.09	86.51	**88.47**	88.28	88.37	86.42	89.64	88.01
DistilmBERT	84.03	86.98	85.48	85.30	86.49	85.89	83.18	87.38	85.23
XLM-100	**88.16**	88.11	88.14	87.86	89.52	88.68	73.27	80.48	76.71
XLM-RoBERTa	87.55	**91.19**	**89.34**	87.50	**90.76**	**89.10**	**87.69**	**92.70**	**90.12**
CRF	87.97	88.03	87.99						
StanfordNLP	**90.55**	**91.07**	**90.80**						

Table 8. NER F1 scores with additional POS and morphological information.

	PRE-BERT			POST-BERT			Regular
	POS+Morph	POS	Morph	POS+Morph	POS	Morph	-
mBERT	82.58	83.80	86.41	87.10	**88.59**	87.13	85.39
distilmBERT	70.30	79.39	82.16	81.84	83.51	84.97	**85.48**
XLM-100	80.26	82.48	87.36	81.25	86.76	86.42	**88.14**
XLM-RoBERTa	89.71	**89.86**	89.43	89.52	86.76	87.62	89.34

POS and/or morphological information, the BERT model has the potential to improve their results, as especially POS information can be crucial for detecting proper names that make up a large number of named entities.

We experimented with two different approaches for adding POS and morphological information to the BERT-based models. The first approach (PRE-BERT) only changes the input of the models. Here, the POS and morphological information is input directly into the BERT model by adding the embeddings of POS and morphological tags to the default input embeddings by summing all embedding vectors. The second approach (POST-BERT) requires slight changes in the sequence classification model. Here, the embeddings of POS and morphological tags are concatenated to the output vector obtained from the BERT model and the concatenated representation is then input to the classification layer. We expect the POST-BERT method to perform better because in this approach, the POS and morphological information is fed to the model closer to the classification layer and thus has the more direct influence on the classification decision. The advantage of the PRE-BERT approach, on the other hand, is its simplicity as it does not require any changes in the model architecture. For training with both approaches, we used the POS and morphological information supplied with the NER corpus. The POS and morphological tags for the test part were obtained with the open-source Estonian morphological analyzer Vabamorf [18] that uses the same annotation scheme as supplied in the NER corpus.

Table 8 shows that the results of adding POS and/or morphological features is mixed. While mBERT achieves a large improvement and XLM-RoBERTa, a marginal increase in performance, the scores of other two models actually decrease quite a bit. Overall, as expected, the POST-BERT approach, where the extra features are concatenated to the output vector of BERT, is better than the PRE-BERT approach. The exception is again the XLM-RoBERTa model that with the PRE-BERT method achieves the best NER results of all multilingual models. However, this best score is still about one percentage point worse than the Stanford tagger baseline. From the three settings adding only POS or morphological features seems the best. To conclude, adding either POS or morphological features can be helpful for the mBERT and XLM-RoBERTa models, other two models were not able to use the extra features to increase the scores.

6. Conclusions

In this work, we compared multilingual BERT and BERT-like models with non-contextual baseline models on several downstream NLP tasks. For most tasks, multilingual models outperformed the previously proposed task-specific models, XLM-

RoBERTa achieving the highest scores on all the experimental tasks, while DistilmBERT performed the worst overall. Based on these results, we can recommend using the XLM-RoBERTa as a basis for neural NLP models for Estonian. Considering the results from previous works comparing multilingual BERT with language-specific BERT models [3,9], further performance gains can be obtained from training monolignual BERT for Estonian, in particular following the RoBERTa guidelines [19].

References

[1] Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep Contextualized Word Representations. In: Proceedings of NAACL; 2018. p. 2227–2237.

[2] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of NAACL; 2019. p. 4171–4186.

[3] Nozza D, Bianchi F, Hovy D. What the [MASK]? Making Sense of Language-Specific BERT Models. arXiv preprint arXiv:200302912. 2020.

[4] Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, et al. Unsupervised Cross-lingual Representation Learning at Scale. In: Proceedings of ACL; 2020. p. 8440–8451. Available from: https://www.aclweb.org/anthology/2020.acl-main.747.

[5] Kondratyuk D, Straka M. 75 Languages, 1 Model: Parsing Universal Dependencies Universally. In: Proceedings of EMNLP-IJCNLP; 2019. p. 2779–2795.

[6] Conneau A, Lample G. Cross-lingual Language Model Pretraining. In: NIPS; 2019. p. 7059–7069. Available from: http://papers.nips.cc/paper/8928-cross-lingual-language-model-pretraining.pdf.

[7] Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. arXiv preprint arXiv:191001108. 2019.

[8] Pires T, Schlinger E, Garrette D. How Multilingual is Multilingual BERT? In: Proceedings of ACL; 2019. p. 4996–5001.

[9] Virtanen A, Kanerva J, Ilo R, Luoma J, Luotolahti J, Salakoski T, et al. Multilingual is not enough: BERT for Finnish. arXiv preprint arXiv:191207076. 2019.

[10] Baumann A. Multilingual Language Models for Named Entity Recognition in German and English. In: Proceedings of RANLP SRW 2019; 2019. p. 21–27.

[11] Akbik A, Blythe D, Vollgraf R. Contextual String Embeddings for Sequence Labeling. In: Proceedings of COLING; 2018. p. 1638–1649.

[12] Kuratov Y, Arkhipov M. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. arXiv preprint arXiv:190507213. 2019.

[13] Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In: Proceedings of ACL System Demonstrations; 2020. .

[14] Straka M, Straková J. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In: Proceedings of the CoNLL 2017 Shared Task; 2017. p. 88–99.

[15] Pajupuu H, Altrov R, Pajupuu J. Identifying Polarity in Different Text Types. Folklore. 2016;64.

[16] Joulin A, Grave É, Bojanowski P, Mikolov T. Bag of Tricks for Efficient Text Classification. In: Proceedings of EACL; 2017. p. 427–431.

[17] Tkachenko A, Petmanson T, Laur S. Named Entity Recognition in Estonian. In: Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing; 2013. p. 78–83. Available from: https://www.aclweb.org/anthology/W13-2412.

[18] Kaalep HJ. An Estonian Morphological Analyser and the Impact of a Corpus on its Development. Computers and the Humanities. 1997;31(2):115–133.

[19] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:190711692. 2019.