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

doi:10.3233/FAIA200610

LVBERT: Transformer-Based Model for Latvian Language Understanding

Artūrs ZNOTIŅŠ [1] and Guntis BARZDIŅŠ
Institute of Mathematics and Computer Science, University of Latvia, Latvia

Abstract. This paper presents LVBERT – the first publicly available monolingual language model pre-trained for Latvian. We show that LVBERT improves the state-of-the-art for three Latvian NLP tasks including Part-of-Speech tagging, Named Entity Recognition and Universal Dependency parsing. We release LVBERT to facilitate future research and downstream applications for Latvian NLP.

Keywords. Transformers, BERT, language models, Latvian

1. Introduction

Pre-trained contextualized text representation models, especially BERT – the Bidirectional Encoder Representations from Transformers [1], have become very popular and helped to achieve state-of-the-art performances in multiple Natural Language Processing (NLP) tasks [2]. Previously, the most common text representations were based on word embeddings that aimed to represent words by capturing their distributed syntactic and semantic properties [3], [4]. However, these word embeddings do not incorporate information about the context in which the words appear. This issue was addressed by BERT and other pre-trained language models. The success of BERT and its variants has largely been limited to the English language. For other languages, one could use existing pre-trained multilingual BERT-based models [1], [5] and optionally fine-tune them, or retrain a language-specific model from scratch [6], [7]. The latter approach has been proven to be superior [8].

Our contributions are as follows:

- We present a methodology to pre-train the BERT model on a Latvian corpus.
- We evaluate LVBERT and show its superiority on three NLP tasks: Part-of-Speech (POS) tagging, Named Entity Recognition (NER) and Universal Dependency (UD) parsing.
- We make LVBERT model publicly available[2].

[1]Corresponding Author: Artūrs Znotiņš; Institute of Mathematics and Computer Science, University of Latvia, Riga, Latvia; E-mail: arturs.znotins@lumii.lv.
[2]https://github.com/LUMII-AILab/LVBERT

Table 1. Number of sentences and tokens in the pre-training dataset from each source

Source	#Sentences (M)	#Tokens (M)
Balanced Corpus	0.7	12
Wikipedia	1.3	25
Comments	5	80
News	20	380
Total	27	500

2. LVBERT

2.1. Model

In our experiments, we used the original implementation of BERT on TensorFlow with the whole-word masking and the next sentence prediction objectives. We used $\text{BERT}_{\text{BASE}}$ configuration with 12 layers, 768 hidden units, 12 heads, 128 sequence length, 128 mini-batch size and 32,000 token vocabulary. The model was trained on a single TPUv2 for 10,000,000 steps that took about 10 days.

2.2. Pre-training Dataset

The original BERT was trained on 3.3B tokens extracted from English Wikipedia and the Book Corpus [9]. Latvian Wikipedia dump is relatively small compared to English. To increase lexical diversity, we included texts from the Latvian Balanced corpus LVK2018 [10], Wikipedia, comments and articles from various news portals (see Table 1). Dataset contains 500M tokens.

2.3. Sub-Word Unit Segmentation

Sub-word tokenization is one of the problems of the multilingual BERT model that uses 110k shared sub-word token vocabulary. Because Latvian is under-represented in the training dataset, tokenization into sub-word units is very fragmented, especially for less frequent words. We trained SentencePiece model [11] on the pre-training dataset to produce a vocabulary of 32,000 tokens that was then converted to WordPiece format used by BERT. For sentence tokenization, we used LVTagger [12]. mBERT's sub-words tend to be shorter and less interpretable, for example:

```
mBERT:  So #fi #ja a #ši ie #sl #ē #dz #ās sa #vā m #ā #jā .
LVBERT: Sofija     a #ši ieslēdz #ās       savā   mājā     .
```

3. Evaluation

We evaluated LVBERT on three Latvian NLP tasks: POS, NER, UD. We compared LVBERT model results with the multilingual BERT model (mBERT) results and the current state-of-the-art on each task. We also fine-tuned multilingual BERT model (mBERT-adapted) on our pre-training dataset and evaluated it to assess usefulness of additional target language data. All model results were averaged over three runs.

Table 2. Named entity dataset statistics

	Train	Dev	Test
GPE	1600	218	207
entity	168	18	29
event	214	22	22
location	538	54	84
money	29	3	12
organization	1354	237	251
person	2466	320	306
product	231	31	31
time	967	144	115

3.1. Part-of-Speech Tagging

For POS tagging, we used bidirectional LSTM architecture and compared results to the current state-of-the-art Latvian morphological tagger [13]. We only evaluated the POS tag accuracy ignoring full morphological tag.

3.2. Named Entity Recognition

For training and evaluating NER, we used a recently published multi-layer text corpus for Latvian [14]. Named entity layer includes annotation of nine entity types: person, organization, geopolitical entity (GPE), location, product, event, time (relative or absolute date, time, or duration), money, and unclassified entity. In this work, only the outer level entities are considered, ignoring hierarchical annotation of named entities. We use the same train/development/test data split as for Universal Dependency layer to preserve corpus distribution of genres and to prevent document overlap between splits (see table 2).

We used a standard neural architecture consisting of bidirectional LSTM with a sequential conditional random fields layer above it. IOB2 (Inside, Outside, Beginning) tagging scheme was used to model named entities that span several tokens. The current best Latvian NER model based on GloVe word embeddings [15] was re-evaluated on the same dataset and compared to the BERT-based models.

3.3. Universal Dependency Parsing

For dependency parsing, we used a model based on biaffine classifiers on top of a bidirectional LSTM [16], specifically, AllenNLP[3] implementation, and compared results to the current state-of-the-art [15] re-evaluated on the latest Universal Dependency release.

4. Results

The main results of our experiments are presented in Table 3. LVBERT models achieved the best results in all three tasks. The most significant improvement was visible in the UD task. Multiligual BERT model under-performed in the NER and POS tagging tasks

[3] https://allennlp.org/

Table 3. Performance of LVBERT on Latvian NLP tasks compared to multilingual BERT and previous state-of-the-art

Task	Metric	Previous Best	mBERT	mBERT-adapted	LVBERT
POS	Accuracy	97.9	96.6	98.0	98.1
NER	F1-score	78.4	79.2	81.9	82.6
UD	LAS	80.6	85.7	88.1	89.9

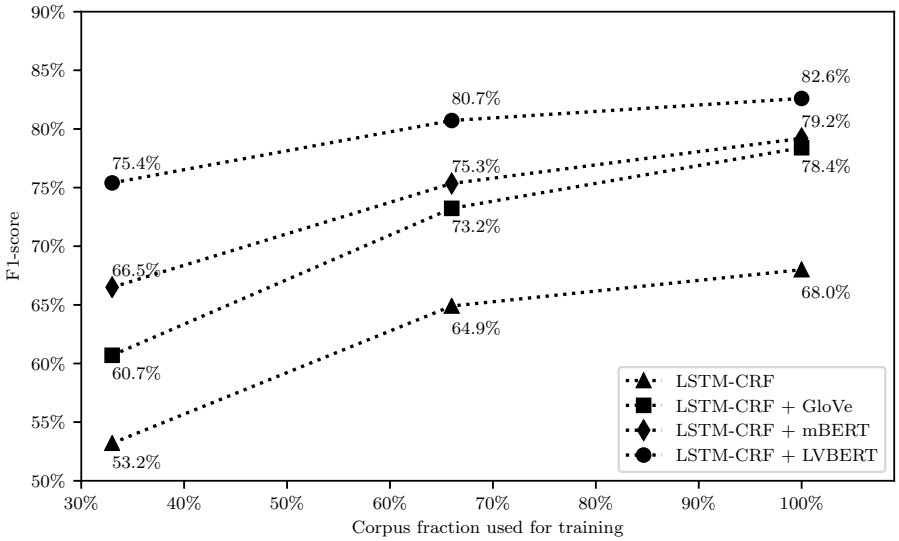

Figure 1. NER learning curve

compared to relatively simple word embedding based models showcasing its shortcomings for less resourced languages. To fully utilize the potential of BERT, the multilingual model should be at least fine-tuned on the specific language texts. Fine-tuned BERT model performed surprisingly well when compared to LVBERT given its vocabulary disadvantage. NER learning curve (see Figure 1) shows that pre-trained contextualized text representation models can more fully utilize limited amount of training data compared to simple word embeddings, especially if just a relatively small part of the annotated training corpus is used.

5. Conclusion

This paper showcases that even a relatively small language specific BERT model can significantly improve results over non-contextual representations and also multilingual BERT model. LVBERT sets a new state-of-the-art for several Latvian NLP tasks. By publicly releasing LVBERT model, we hope that it will serve as a new baseline for these tasks and that it will facilitate future research and downstream applications for Latvian NLP.

Acknowledgements

This research is funded by the Latvian Council of Science, project "Latvian Language Understanding and Generation in Human-Computer Interaction", project No. lzp-2018/2-0216.

References

[1] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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). Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171–4186. Available from: https://www.aclweb.org/anthology/N19-1423.

[2] Howard J, Ruder S. Universal language model fine-tuning for text classification. arXiv preprint arXiv:180106146. 2018.

[3] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems; 2013. p. 3111–3119.

[4] Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP); 2014. p. 1532–1543.

[5] Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, et al. Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:191102116. 2019.

[6] de Vries W, van Cranenburgh A, Bisazza A, Caselli T, van Noord G, Nissim M. BERTje: A Dutch BERT Model. arXiv preprint arXiv:191209582. 2019.

[7] Martin L, Muller B, Suárez PJO, Dupont Y, Romary L, de la Clergerie ÉV, et al. CamemBERT: a Tasty French Language Model. arXiv preprint arXiv:191103894. 2019.

[8] 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.

[9] Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE international conference on computer vision; 2015. p. 19–27.

[10] Levane-Petrova K. Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss, tā nozīme gramatikas pētījumos. Language: Meaning and Form. 2019;10:131–146. The Balanced Corpus of Modern Latvian, its role in grammar studies. Available from: https://www.apgads.lu.lv/fileadmin/user_upload/lu_portal/apgads/PDF/Valoda-nozime-forma/VNF-10/vnf_10-12_Levane_Petrova.pdf.

[11] Kudo T, Richardson J. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:180806226. 2018.

[12] Paikens P, Rituma L, Pretkalniņa L. Morphological analysis with limited resources: Latvian example. In: Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); 2013. p. 267–277.

[13] Paikens P. Deep Neural Learning Approaches for Latvian Morphological Tagging. In: Baltic HLT; 2016. p. 160–166.

[14] Gruzitis N, Pretkalnina L, Saulite B, Rituma L, Nespore-Berzkalne G, Znotins A, et al. Creation of a Balanced State-of-the-Art Multilayer Corpus for NLU. In: Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC); 2018. p. 4506–4513. Available from: http://www.lrec-conf.org/proceedings/lrec2018/pdf/935.pdf.

[15] Znotins A, Cirule E. NLP-PIPE: Latvian NLP Tool Pipeline. In: Human Language Technologies - The Baltic Perspective. vol. 307. IOS Press; 2018. p. 183–189. Available from: http://ebooks.iospress.nl/volumearticle/50320.

[16] Dozat T, Manning CD. Deep biaffine attention for neural dependency parsing. arXiv preprint arXiv:161101734. 2016.