Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)
© 2020 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA200603

Large Language Models for Latvian Named Entity Recognition

Rinalds VĪKSNA [a,b,1], Inguna SKADIŅA [a,b]
[a] *Tilde*
[b] *Faculty of Computing, University of Latvia, Latvia*

Abstract. Transformer-based language models pre-trained on large corpora have demonstrated good results on multiple natural language processing tasks for widely used languages including named entity recognition (NER). In this paper, we investigate the role of the BERT models in the NER task for Latvian. We introduce the BERT model pre-trained on the Latvian language data. We demonstrate that the Latvian BERT model, pre-trained on large Latvian corpora, achieves better results (81.91 F1-measure on average vs 78.37 on M-BERT for a dataset with nine named entity types, and 79.72 vs 78.83 on another dataset with seven types) than multilingual BERT and outperforms previously developed Latvian NER systems.

Keywords. Named entity recognition, NER, Latvian language, BERT

1. Introduction

Recently developed pre-trained language representation models have demonstrated significant improvement in many natural language processing tasks. The most popular are ELMo [1], BERT [2] and RoBERTa [3]. The BERT model has shown the state-of-the art performance for tasks of named entity recognition (NER), question answering, classification, and others. The multilingual BERT model (M-BERT)[2], pre-trained on Wikipedia texts in 104 languages, has demonstrated good results in zero-shot cross-lingual model transfer, where a model trained on one language (presumably, with large annotated corpora) is evaluated on another language [4]. However, some recent publications show that the monolingual BERT model could achieve significantly better results compared to the multilingual [5], [6], [7].

One of the tasks where pre-trained language representation models have been successfully applied is named entity recognition. Traditionally, NER is understood as identification of the text spans containing named entities and classifying them into predefined categories (e.g. person names (John, Barack Obama, etc.), organizations (BMW, IBM, etc.), locations (Riga, Washington, etc.) and other). NER serves as a basis for many natural language understanding tasks such as semantic annotation, question answering, ontology population, and opinion mining [8].

[1] Corresponding Author: Rinalds Vīksna; Tilde, Vienibas gatve 75a, LV-1004, Riga, Latvia; E-mail: rinalds.viksna@tilde.lv.

[2] https://github.com/google-research/bert.

Most of the research on the role of pre-trained language representation models for NER task has been performed on resource-rich languages, e.g., English, German, and Chinese [9], [10], while less attention has been paid to the less-resourced and morphologically rich languages like Persian [11], Finnish [6], and Portuguese [7].

In this paper, we examine the role of the BERT models in the NER task for the Latvian language. We introduce the Latvian BERT model and compare its performance to M-BERT in the NER task. Several configurations of BERT have been pre-trained on different corpora, adapted, and evaluated on two different datasets. Developed NER models have been compared with previously developed solutions (where possible). We demonstrate that pre-trained BERT models significantly improve the overall performance of NER and NER using the Latvian BERT model outperforms M-BERT.

2. Related Work

The first named entity recognizer for Latvian and Lithuanian TildeNER uses the StanfordNER random field classifier for training and tagging [12]. It also features a bootstrapping module. Evaluation on the test set using 7 class tagset (organization, person name, location, product, date, time, money) yielded F-measure of 60.19 for Latvian and 65.12 for Lithuanian. TildeNER was also compared to StanfordNER using a comparable corpus of 10 documents, where TildeNER achieved an F-measure of 56.46 while detecting location entities, 61.63 for person entities, and 65.71 detecting organizations.

The Latvian NLP tool pipeline [13] contains a NER module that tags 7 named entity classes: person, organization, geopolitical entity, location, product, time and event. It was trained on the Latvian Multilayer Corpus for NLU [14]. Data[3] [15] were serialized using a modified CoNLL-2003[4] data format, which supports hierarchical named entity annotation. Using a bidirectional LSTM neural network with CRF layer and word embeddings, the model achieves a 74.0 F1 measure on average. The model demonstrates good results on person entities (85.2 F1), while performing poorly on events (20.0 F1) and locations (45.1 F1).

Arkhipov et.al. [16] used multilingual BERT to initialize pre-training of the BERT model for Russian, Bulgarian, Czech, and Polish. The pre-trained model was extended with CRF layer to recognize five classes of entities: persons, locations, organizations, events, and products. Using additionally pre-trained SlavicBERT model together with additional CRF layer, they achieved an F1 score of 87.3 for Russian, 93.2 for Polish, 93.9 for Czech, and 87.2 for Bulgarian.

Souza et.al [7] also used a multilingual BERT model with a CRF layer to detect named entities in Portuguese. Authors compared two transfer learning approaches: feature-based and fine-tuning based. Feature-based approach uses Bi-LSTM layer and linear layer, and BERT is used to obtain word embeddings. In the fine-tuning approach, the classifier is a linear layer and all weights are updated during training. The best model uses fine-tuning approach with CRF and has achieved an F1 score of 74.15 against 70.33 F1 score obtained using a baseline Bi-LSTM.

[3] https://github.com/LUMII-AILab/FullStack/tree/master/NamedEntities.

[4] https://www.clips.uantwerpen.be/conll2003/ner/.

Virtanen et al. [6] pre-trained Finnish monolingual BERT model (FinBERT) on 234M sentences (about 3.3B tokens) crawled from the web and news. The FinBERT was evaluated on the NER task against uncased FinBERT and multilingual BERT models using the FiNER dataset, which includes nested named entities. NER model was built using the FinBERT as a base with a dense layer on top. This model achieved an F1 score of 92.4 on in-domain data and 81.47 on out of domain test set, while the multilingual BERT achieved F1 scores of 90.29 and 76.15, respectively.

3. Latvian BERT Model

3.1. Data Collection and Processing

For pre-training the Latvian BERT, unlabeled data were acquired from different sources: the EUbookshop[5], JRC-Acquis[6], Latvian Wikipedia, and various European and Latvian websites. Crawled data were then cleaned: boilerplate content and HTML tags were removed, texts converted into UTF-8 encoding, and language identification performed (documents with less than 80 % of Latvian content were removed). Documents containing long sequences of short segments or numbers were removed as well.

After cleaning, the text corpus used for BERT pre-training contained 124 million sentences or 1.6 billion tokens. In comparison, the English BERT model was pre-trained on 3.3 billion words from BookCorpus (800 million words) and English Wikipedia (2,500 million words), the Portuguese BERT was pre-trained on 2.6 billion tokens and the Finnish BERT was pre-trained on 3.3 billion tokens.

3.2. Pre-training

From the collected corpus, the byte pair encoding (BPE) vocabulary [17] was created using the sentencepiece [18] and converted to the wordpiece format used by BERT. BPE vocabulary was generated using a cased version of the corpus and its size was set to 30,000 word-pieces. BERT scripts were used to create pre-training examples with a sequence length of 128, while other parameters were set to match the original BERT model [2]. The model was pre-trained for 4M steps using learning rate 5e-5 and 10,000 warmup steps.

We also pre-trained the multilingual BERT model for 1M steps using Latvian data to evaluate the usefulness of additional pre-training.

4. NER Systems

Three NER systems using different BERT models have been trained and evaluated:

- "Multi-base": NER model that uses the multilingual BERT model;
- "Multi-updated": NER model that uses a multilingual model additionally pre-trained with Latvian data;

[5] http://opus.nlpl.eu/EUbookshop.php.

[6] http://opus.nlpl.eu/JRC-Acquis.php.

- "lv-base": NER model trained using Latvian BERT.

4.1. Datasets for NER Training

Only two rather small datasets are available and were used in our experiments: proprietary TildeNER dataset and the named entity annotated layer of the publicly available Latvian Multilayer Corpus (the AILab dataset)[7].

Table 1. TildeNER dataset statistics

NE type	NE count		
	Manually created data	**Bootstrapped data**	**TOTAL**
DATE	1,590	791	2,381
LOCATION	2,611	1,759	4,370
MONEY	289	671	960
ORGANIZATION	1,649	638	2,287
PERSON	1,037	1,282	2,391
PRODUCT	866	233	1,099
TIME	353	125	478
TOTAL	**8,395**	**5,499**	**13,966**

TildeNER dataset [12] consists of two parts (Table 1). The initial dataset was manually created for training and evaluation of the Latvian NER system. Annotation was performed by 2 annotators, while the third annotator resolved the disagreement. Additional annotated data were bootstrapped during the development process and verified by a human annotator.

Table 2. Entity count in Multilayer Corpus for NLU dataset

NE type	Entity count
PERSON	3,104
GPE	2,031
ORGANIZATION	1,847
TIME	1,227
PRODUCT	293
LOCATION	677
EVENT	259
ENTITY	215
MONEY	44
TOTAL	**9,697**

[7] https://github.com/LUMII-AILab/FullStack/tree/master/NamedEntities

The second dataset is from the Balanced State-of-the-Art Multilayer Corpus for NLU [15]. It contains 3,947 paragraphs with 9,697 outer and 944 inner entities. In this work, we use only outer entities (Table 2).

For NER model training, the corpus was transformed into a CONLL-2003 format. To enable comparison with the NER model developed by Znotiņš and Cīrule [13], classes "Money" and "Entity" were labeled as "O" – Other.

4.2. Training

We use a dense+crf layer on top of the BERT for classification (Figure 1). Words are split into word pieces using BERT tokenizer, and Latvian wordpiece vocabulary. When BERT tokenizer splits words into subword tokens, we label only the first subword of each word according to BIO (identifies the Beginning, the Inside, and the Outside of text segment) labeling scheme, while other subwords get label "X". This additional label "X" is removed later, as output is words. The NER model is trained for 12 epochs, using sequence length 128, train batch size 4, and learning rate 2e-5.

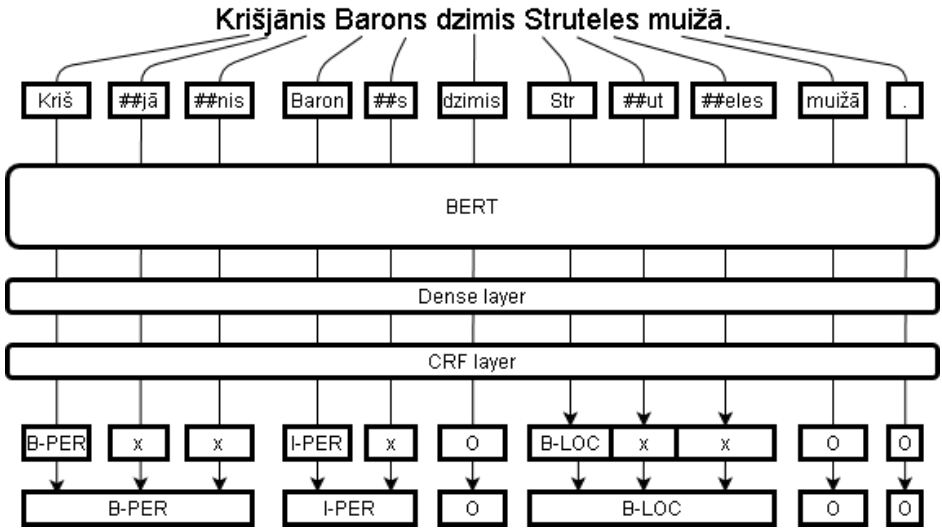

Figure 1. NER model architecture

5. Results and Evaluation

At first, we trained the NER model using data from the AILab dataset. An earlier version of this dataset was also used to train the NER model for NLP-PIPE [13][8]. This and other Latvian NER systems were evaluated by F1-measure using the CoNNL-2003 evaluation script[9]. Table 3 summarizes evaluation results, demonstrating that NER

[8] This NER model recognizes 7 out of 9 entity classes presented in the dataset.

[9] https://www.clips.uantwerpen.be/conll2002/ner/bin/conlleval.txt

model trained with the Latvian BERT (lv-base) outperforms the model, that was trained using multilingual BERT (Multi-base). The model that was trained using additionally pre-trained multilingual BERT (Multi-updated) performed poorly and thus was not used in further experiments. Probably, the learning rate 2e-5 was too high and additional pre-training did harm to the model. All models perform poorly in detecting classes that have little training data (less than 300 examples), i.e., product, event, entity, and money.

Table 3. Evaluation results (F1) on the AILab dataset

NE type	Multi-updated	Multi-base	lv-base	NLP-PIPE [13]
GPE	84.77	88.18	**89.66**	79.0
ENTITY	30.77	35.59	**44.68**	
EVENT	51.43	48.6	**59.46**	20.0
LOCATION	56.86	61.59	**67.79**	45.1
MONEY	**15.38**	11.11	0	
ORGANIZATION	71.9	77.91	**81.7**	78.5
PERSON	89.82	91.99	**94.91**	85.2
PRODUCT	45.61	**65.17**	64	40.0
TIME	64.37	63.14	65.64	**71.7**
F1	**75.20**	**78.37**	**81.91**	**74**

During experiments, we noticed that sometimes GPE and location categories are very similar and overlapping entities, and some classes are very small. Therefore, we decided to test model performance for only 4 classes. GPE and location were merged into LOCATION; person and organization classes were kept, and all the rest were merged in MISC class. The evaluation results in Table 4 show that with 4 classes, the Latvian BERT performs even better, achieving on average F1 score of 84.82.

Table 4. NER evaluation results (F1) on the AILab dataset (4 classes)

NE type	Multi-base	lv-base
LOCATION	86.23	90.49
MISC	63.11	65.93
ORGANIZATION	76.9	80.48
PERSON	91.63	95.34
F1	**81.1**	**84.82**

As next, we trained NER systems with the TildeNER dataset. As it is demonstrated in Table 5, for this dataset, the Latvian BERT achieves better F1 measure in total, but multilingual BERT also performs quite well. Although our results are not directly

comparable with TildeNER[10], a huge gap of F1-score, when detecting locations and persons, is observed.

Table 5. Evaluation results (F1) on the TildeNER dataset (7 classes)

NE type	Multi-base	lv-base	TildeNER
DATE	70.74	79.07	
LOCATION	90.09	90.03	56.46
MONEY	81.54	85.5	
ORGANIZATION	70.81	70.98	65.71
PERSON	86.88	90.96	61.63
PRODUCT	67.25	56.34	
TIME	71.7	79.31	
F1	**78.83**	**79.72**	

BERT-based NER systems identify products poorly: they detect multiple products, which are not "products". Examples include button combinations ("Command+i", "Ctrl"), computer user interface parts ("Dock", "Applications", "Start"). The detection of organizations also suffers, because organization names are often complex multiword expressions, and often some of the tokens are marked as organizations wrongly. Examples which are marked wrongly as organizations include "President of Latvia" (this one counts an error for location as well), "International Bonds", "International Coordination Committees", and others.

6. Conclusion and Next Steps

In this paper, we examined the impact of large pre-trained BERT language models on named entity recognition in the case of a morphologically rich less-resourced language, specifically, Latvian. We demonstrated that large pre-trained BERT language models have a significant impact on the quality of NERs: the Latvian BERT model, pre-trained on large Latvian corpora, achieves better results (81.91 F1-measure on average vs. 78.37 on multi-BERT for the AILab dataset with nine NE types, and 79.72 vs. 78.83 on the TildeNER dataset with seven types) than multilingual BERT and outperforms previously developed NER systems for Latvian that were created using different architectures.

Acknowledgements

This research has been supported by the European Regional Development Fund within the joint project of SIA TILDE and University of Latvia "Multilingual Artificial Intelligence Based Human Computer Interaction" No. 1.1.1.1/18/A/148.

[10] Results reported for TildeNER [12] are obtained on different dataset and presented only for location, person, organization in comparative evaluation with Stanford NER classifier.

References

[1] Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep Contextualized Word Representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics; 2018. p. 2227–37. Available from: http://aclweb.org/anthology/N18-1202.

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

[3] Wang Y., Sun Y., Ma Z., Gao L., Xu Y. and Sun T. Application of Pre-training Models in Named Entity Recognition; 2020. Available from: http://arxiv.org/abs/2002.08902.

[4] Pires T, Schlinger E, Garrette D. How Multilingual is Multilingual BERT? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics [Internet]. Stroudsburg, PA, USA: Association for Computational Linguistics; 2019. p. 4996–5001.

[5] Martin L, Muller B, Javier P, Su O, Romary L, Villemonte E, et al. CamemBERT : a Tasty French Language Model; 2019. Available from: https://arxiv.org/abs/1911.03894

[6] Virtanen A, Kanerva J, Ilo R, Luoma J, Luotolahti J, Salakoski T, et al. Multilingual is not enough: BERT for Finnish. 2019 Dec 15; Available from: http://arxiv.org/abs/1912.07076.

[7] Souza F, Nogueira R, Lotufo R. Portuguese Named Entity Recognition using BERT-CRF. 2019; Available from: http://arxiv.org/abs/1909.10649

[8] Marrero M, Urbano J, Sánchez-cuadrado S, Morato J, Gómez-berbís JM. Named Entity Recognition : Fallacies , Challenges and Opportunities. Computer Standards & Interfaces 5, 2013, p. 482-489

[9] Yadav, V., Bethard, S.; A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. Proceedings of the 27th International Conference on Computational Linguistics, August 20-26, 2018; Santa Fe, New Mexico, USA, pages 2145–2158.

[10] Li J, Sun A, Han J, Li C. A Survey on Deep Learning for Named Entity Recognition. IEEE Transactions on Knowledge and Data Engineering. 2020;1–1.

[11] Taher E, Hoseini SA, Shamsfard M. Beheshti-NER: Persian Named Entity Recognition Using BERT. 2020 Mar 19; Available from: http://arxiv.org/abs/2003.08875

[12] Pinnis M. Latvian and Lithuanian Named Entity Recognition with TildeNER. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12); 2012. p. 1258-1265.

[13] Znotins A, Cirule E. NLP-PIPE: Latvian NLP tool pipeline. Frontiers in Artificial Intelligence and Applications. 2018;307:183–9.

[14] Chinchor N. MUC-7 Named Entity Task Definition; 1998. https://www.aclweb.org/anthology/M98-1028.

[15] Gruzitis N., Pretkalnina L., Saulite B., Rituma L., Nespore-Berzkalne G., Znotins A. and Paikens P. Creation of a balanced state-of-the-art multilayer corpus for NLU. LREC 2018 - 11th International Conference on Language Resources and Evaluation. 2019; p. 4506–4513.

[16] 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. Stroudsburg, PA, USA: Association for Computational Linguistics; 2019. p. 89–93.

[17] Wu Y, Schuster M, Chen Z, Le Q V., Norouzi M, Macherey W, et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. 2016 Sep 26;1–23. Available from: http://arxiv.org/abs/1609.08144

[18] Kudo T, Richardson J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. EMNLP 2018 - Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Proceedings. 2018;66–71.