Tuning Multilingual Transformers for Named Entity Recognition on Slavic Languages

Mikhail Arkhipov[*,1] **Maria Trofimova**[*,2] **Yuri Kuratov**[*,3] **Alexey Sorokin**[*,◇,4]

[*]Neural Networks and Deep Learning Laboratory, Moscow Institute of Physics and Technology
[◇]Faculty of Mathematics and Mechanics, Moscow State University
[1]arkhipov.mu@mipt.ru
[2]mary.vikhreva@gmail.com
[3]yurii.kuratov@phystech.edu
[4]alexey.sorokin@list.ru

Abstract

Our paper addresses the problem of multilingual named entity recognition on the material of 4 languages: Russian, Bulgarian, Czech and Polish. We solve this task using the BERT model. We use a hundred languages multilingual model as base for transfer to the mentioned Slavic languages. Unsupervised pre-training of the BERT model on these 4 languages allows to significantly outperform baseline neural approaches and multilingual BERT. Additional improvement is achieved by extending BERT with a word-level CRF layer. Our system was submitted to BSNLP 2019 Shared Task on Multilingual Named Entity Recognition and took the 1st place in 3 competition metrics out of 4 we participated in. We open-sourced NER models and BERT model pre-trained on the four Slavic languages.

1 Introduction

Named Entity Recognition (further, NER) is a task of recognizing named entities in running text, as well as detecting their type. For example, in the sentence *Asia Bibi is from Pakistan*, the following NER classes can be detected: [*Asia Bibi*]$_{\text{PER}}$ *is from* [*Pakistan*]$_{\text{LOC}}$. The commonly used BIO-annotation for this sentence is shown in Figure 1.

The recognizer of named entities can be trained on a single target task dataset as any other sequence tagging model. However, it often benefits from additional data from a different source, either labeled or unlabeled, which is known as transfer learning. To enrich the model one can either train it on several tasks simultaneously (Collobert et al., 2011), which makes its word representations more flexible and robust, or pretrain on large amounts of unlabeled data to utilize unlimited sources available in the Web and then fine-tune them on a specific task (Dai and Le, 2015; Howard and Ruder, 2018).

One of the most powerful unsupervised models is BERT (Devlin et al., 2018), which is a multi-layer Transformer trained on the objective of masked words recovery and on the task of next sentence prediction (known also as Natural Language Inference (NLI) task). The original model was trained on vast amounts of data for more than 104 languages which makes its representations useful for almost any task. Our contribution is three-fold: first, multilingual BERT embeddings with a dense layer on the top clearly beat BiLSTM-CRF over FastText embeddings trained on the four target languages. Second, language-specific BERT, trained only on the target languages from Wikipedia and news dump, significantly outperforms the multilingual BERT. Third, we adapt a CRF layer as a a top module over the outputs of the BERT-based model and demonstrate that it improves performance even further.

2 Model Architecture

Our model extends the recently introduced BERT (Devlin et al., 2018) model. BERT itself is a multilayer transformer (Vaswani et al., 2017) which takes as input a sequence of subtokens, obtained using WordPiece tokenization (Wu et al., 2016), and produces a sequence of context-based embeddings of these subtokens. When a word-level task, such as NER, is being solved, the embeddings of word-initial subtokens are passed through a dense layer with softmax activation to produce a probability distribution over output labels. We refer the reader to the original paper, see also Figure 2.

We modify BERT by adding a CRF layer instead of the dense one, which was commonly used in other works on neural sequence labeling (Lample et al., 2016) to ensure output consistency. It also transforms a sequence of word-initial subtoken embeddings to a sequence of probability dis-

Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, pages 89–93,
Florence, Italy, 2 August 2019. ©2019 Association for Computational Linguistics

Asia	Bibi	is	from	Pakistan	.
B-PER	I-PER	O	O	B-LOC	O

Figure 1: An example of BIO-annotation for tokens.

tributions, however, each prediction depends not only on the current input, but also from the previous one.

3 Transfer from Multilingual Language Model

There are two basic options for building multilingual system: to train a separate model for each language or to use a single multilingual model for all languages. We follow the second approach since it enriches the model with the data from related languages, which was shown to be beneficial in recent studies (Mulcaire et al., 2018).

The original BERT embedder itself is essentially multilingual since it was trained on 104 languages with largest Wikipedias[1]. However, for our four Slavic languages (Polish, Czech, Russian, and Bulgarian) we do not need the full inventory of multilingual subtokens. Moreover, the original WordPiece tokenization may lack Slavic-specific ngrams, which makes the input sequence longer and the training process more problematic and computationally expensive.

Hence we retrain the Slavic BERT on stratified Wikipedia data for Czech, Polish and Bulgarian and News data for Russian. Our main innovation is the training procedure: training BERT from scratch is extremely expensive computationally so we initialize our model with the multilingual one. We rebuild the vocabulary of subword tokens using subword-nmt[2]. When a single Slavic subtoken may consist of multiple multilingual subtokens, we initilalize it as an average of their vectors, resembling (Bojanowski et al., 2016). All weights of transformer layers are initialized using the multilingual weights.

4 Experiment Details

4.1 Target Task and Dataset

The 2019 edition of the Balto-Slavic Natural Language Processing (BSNLP) (Piskorski et al., 2019) shared task aims at recognizing mentions of named entities in web documents in Slavic languages. The input text collection consists of sets of news articles from online media, each collection revolving around a certain entity or an event. The corpus was obtained by crawling the web and parsing the HTML of relevant documents. The 2019 edition of the shared task covers 4 languages (Bulgarian, Czech, Polish, Russian) and focuses on recognition of five types of named entities including persons (PER), locations (LOC), organizations (ORG), events (EVT) and products (PRO).

The dataset consists of pairs of files: news text and a file with mentions of entities with corresponding tags. There are two groups of documents in the train part of the dataset. Namely, news about Asia Bibi and Brexit. Brexit part is substantially bigger, therefore, we used it for training and Asia Bibi for validation.

4.2 Pre- and Post-processing

We use NLTK (Loper and Bird, 2002) sentence tokenizers for Bulgarian, Polish, and Czech. Due to the absence of Bulgarian sentence tokenizer we apply the English NLTK one instead. For Russian language we use DeepMIPT sentence tokenizer[3]. We replace all UTF separators and space characters with regular spaces. Due to mismatch of BSNLP 2019 data format and common format for tagging tasks we first convert the dataset to BIO format to obtain training data. After getting predictions in BIO format we transform them back to the labeling scheme proposed by Shared Task organizers. This step probably causes extra errors, so we partially correct them using post-processing.

We found that sometimes the model predicts a single opening quote without closing one. So we filter out all single quotation marks in the predicted entities. At the prediction stage we perform inference for a sliding window of two sentences with overlaps to reduce sentence tokenization errors.

The Shared Task also included the entity normalization subtask: for example, the phrase

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

[2] https://github.com/rsennrich/subword-nmt

[3] https://github.com/deepmipt/ru_sentence_tokenizer

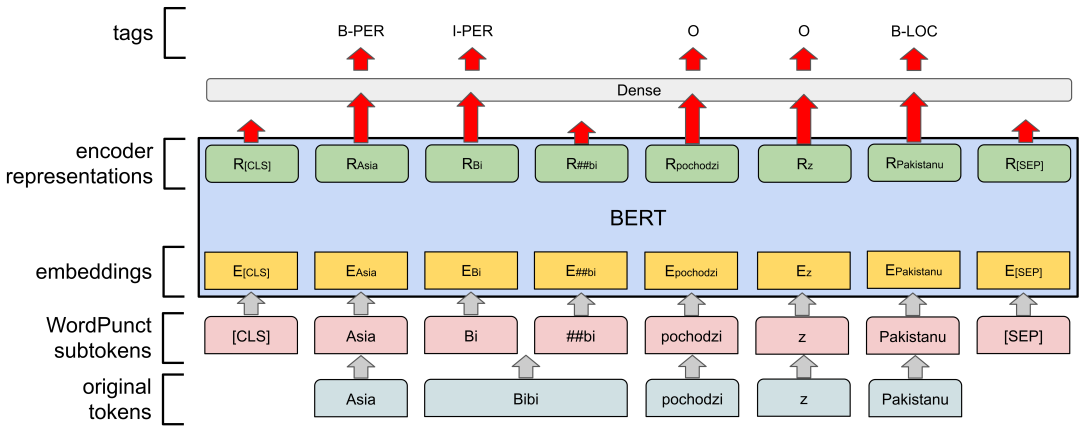

Figure 2: In the figure, E_s and R_s represent the input embedding and the contextual representation of subtoken s, $[CLS]$ is the special symbol to get full input representation, and $[SEP]$ is the special symbol to separate non-consecutive token sequences.

"Верховным судом Пакистана" (Supreme+Ins Court+Ins of Pakistan+Gen) should be "Верховный суд Пакистана". We used the UDPipe 2.3 (Straka et al., 2016) lemmatizers whose output was corrected using language-specific rules. For example, "Пакистана" (Pakistan+Gen) should not be lemmatized because in Russian noun modifiers remain in Genitive.

4.3 Model Parameters

See below parameters of transferring multilingual BERT from to Slavic languages. The training took 9 days with DGX-1 comprising of eight P-100 16Gb GPUs. We train BERT in two stages: train full BERT on sequences with 128 subtokens length and then train only positional embeddings on 512 length sequences. We found that both initialization from multilingual BERT and reassembling of embeddings speed up convergence of the model.

- **Batch size**: `256`
- **Learning rate**: `2e-5`
- **Iterations of full BERT training**: `1M`
- **Iterations of positional embeddings training**: `300k`

Parameters of all BERT-based NER models are:

- **Batch size**: `16`
- **BERT layers learning rate**: `1e-5`
- **Top layers learning rate**: `3e-4`
- **Optimizer**: `AdamOptimizer`
- **Epochs**: `3`

In contrast to original BERT paper (Devlin et al., 2018), we use different learning rates for the task-specific top layers and BERT layers when training BERT-based NER models. We found that this modification leads to faster convergence and higher scores.

We evaluate the model every 10 batches on the whole validation set and chose the one that performed best on it. Despite this strategy being very time consuming, we found it crucial to get extra couple of points. For all experiments we used the span F_1 score for validation.

Our best model used CRF layer and performed moving averages of variables by employing an exponential decay to model parameters.

5 Results

We evaluated Slavic BERT NER model on the BSNLP 2019 Shared Task dataset. The model is compared with two baselines: Bi-LSTM-CRF (Lample et al., 2016) and NER model based on multilingual BERT. For Bi-LSTM-CRF we use FastText word embeddings trained on the same data as Slavic BERT.

Table 1 presents the scores of our model on development set (Asia Bibi documents) when training on Brexit documents. We report a standard span-level F1-score based on the CONLL-2003 evaluation script (Sang and De Meulder, 2003) and three official evaluation metrics(Piskorski et al., 2019)[4]: Relaxed Partial Matching (RPM), Relaxed Exact Matching (REM), and Strict

[4] http://bsnlp.cs.helsinki.fi/BSNLP-NER-Evaluator-19.0.1.zip

Matching (SM). Our system showed top performance in multilingual setting for all mentioned metrics except RPM.

Even without CRF the multilingual BERT model significantly outperforms Bi-LSTM-CRF model. Adding a CRF layer strongly increases performance both for multilingual and Slavic BERT models. Slavic BERT is the top performing model. The error rate of Slavic BERT-CRF is more than one third less than the one of Multilingual BERT baseline.

We experimented with transfer learning from other NER corpora. We used three corpora as source for transfer: Russian NER corpus (Mozharova and Loukachevitch, 2016), Bulgarian BulTreeBank (Simov et al., 2004; Georgiev et al., 2009), and BSNLP 2017 Shared Task dataset(Piskorski et al., 2017)[6] with Czech, Russian, and Polish data. For pre-training we use stratified sample from the concatenated dataset. The set of tags for the task-specific layer includes all tags that occur in at least one dataset. After pre-training we replace the task-specific layer with the one suited for the BSNLP 2019 dataset and train until convergence. We find this approach to be beneficial for models without CRF, however, the CRF-enhanced model without NER pretraining demonstrates slightly higher scores.

Table 2 presents a detailed evaluation report across 4 languages for the top performing Slavic BERT-CRF model. Note that the languages with Latin script (Polish and Czech) demonstrate higher scores than Cyrillic-based ones (Russian and Bulgarian). Low scores for Russian might be caused by the dataset imbalance, since it covers only 7.7% of the whole BSNLP dataset, however, Bulgarian includes 39% but shows even lower quality, especially in terms of recall. We have two explanations: first, incorrect sentence tokenization since we used English sentence tokenizer for Bulgarian (this may explain the skew towards precision). Second, Russian and Bulgarian are much less related than Czech and Polish so they obtain less gain from having additional multilingual data.

5.1 Releasing the Models

We release the best BERT based NER model along with the BERT model pre-trained on the four competition languages[7]. We provide the code for the inference of our NER model as well as for using the pretrained BERT. The BERT model is fully compatible with original BERT repository.

6 Conclusion

We have established that BERT models pre-trained on task-specific languages and initialized using the multilingual model, significantly outperform multilingual baselines on the task of Named Entity Recognition. We also demonstrate that adding a word-level CRF layer on the top improves the quality of both extended models. We hope our approach will be useful to fine-tune language-specific BERTs not only for Named Entity Recognition but for other NLP tasks as well.

Acknowledgements

The research was conducted under support of National Technological Initiative Foundation and Sberbank of Russia. Project identifier 0000000007417F630002.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In *Advances in neural information processing systems*, pages 3079–3087.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Georgi Georgiev, Preslav Nakov, Kuzman Ganchev, Petya Osenova, and Kiril Ivanov Simov. 2009. Feature-rich named entity recognition for Bulgarian using conditional random fields. In *RANLP*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

[6] http://bsnlp-2017.cs.helsinki.fi/shared_task.html

[7] https://github.com/deepmipt/Slavic-BERT-NER

Model	Span F_1	RPM	REM	SM
Bi-LSTM-CRF (Lample et al., 2016)	75.8	73.9	72.1	72.3
Multilingual BERT [5]	79.6	77.8	76.1	77.2
Multilingual BERT-CRF	81.4	80.9	79.2	79.6
Slavic BERT	83.5	83.8	82.0	82.2
Slavic BERT-CRF	87.9	85.7 (90.9)	84.3 (86.4)	84.1 (85.7)

Table 1: Metrics for BSNLP on validation set (Asia Bibi documents). Metrics on the test set are in the brackets.

Language	P	R	F_1
cs	93.6	94.7	93.9
ru	88.2	86.6	87.3
bg	90.3	84.3	87.2
pl	93.4	93.1	93.2

Table 2: Precision (P), Recall (R), and F_1 RPM scores of Slavic BERT-CRF model for Czech (cs), Russian (ru), Bulgarian (bg) and Polish (pl) languages.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. *arXiv preprint cs/0205028*.

Valerie Mozharova and Natalia Loukachevitch. 2016. Two-stage approach in Russian named entity recognition. In *2016 International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT)*, pages 1–6. IEEE.

Phoebe Mulcaire, Swabha Swayamdipta, and Noah Smith. 2018. Polyglot semantic role labeling. *arXiv preprint arXiv:1805.11598*.

Jakub Piskorski, Laska Laskova, Michał Marcińczuk, Lidia Pivovarova, Pavel Přibáň, Josef Steinberger, and Roman Yangarber. 2019. The second cross-lingual challenge on recognition, classification, lemmatization, and linking of named entities across Slavic languages. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, Florence, Italy. Association for Computational Linguistics.

Jakub Piskorski, Lidia Pivovarova, Jan Šnajder, Josef Steinberger, and Roman Yangarber. 2017. The first cross-lingual challenge on recognition, normalization, and matching of named entities in Slavic languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 76–85, Valencia, Spain. Association for Computational Linguistics.

Erik F Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Kiril Simov, Petya Osenova, Sia Kolkovska, Elisaveta Balabanova, and Dimitar Doikoff. 2004. A language resources infrastructure for Bulgarian. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Milan Straka, Jan Hajic, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *LREC*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.