# BSNLP Shared Task 2021 submission: Multilingual Slavic Named Entity Recognition

**Rinalds Vīksna**
Tilde
Faculty of Computing,
University of Latvia
rinalds.viksna@tilde.lv

**Inguna Skadiņa**
Tilde
Faculty of Computing,
University of Latvia
inguna.skadina@tilde.lv

## Abstract

Named entity recognition, in particular for morphological rich languages, is challenging task due to the richness of inflected forms and ambiguity. This challenge is being addressed by SlavNER Shared Task. In this paper we describe system submitted to this task. Our system uses pre-trained multilingual BERT Language Model and is fine-tuned for six Slavic languages of this task on texts distributed by organizers. Our multilingual NER model achieves 83.7 F1 score on all corpora, with best result for Polish (88.8) and worst for Russian (79.1). Entity linking module achieved F1 score of 48.8 as evaluated by bsnlp2021 organizers.

## 1 Introduction

Named entity recognition, in particular for morphological rich languages, is challenging task due to the richness of inflected forms and their ambiguity. Evaluation results are usually lower for these languages, when compared to morphologically simpler languages. For instance, for Finnish language Virtanen et al. (2019) reports the F1 score of 92.4 on in-domain data and 81.47 on out of domain data, for Latvian state-of-the-art NER system (Znotiņš and Barzdins, 2020) achieves the F1 score 82.6, while for English LUKE model (Yamada et al., 2020) achieves F1 score of 94.3 on CoNLL-2003 dataset.

In this paper we present our submission to the SlavNER Shared Task on the analysis of Named Entities in multilingual Web documents in Slavic languages. Our submission implements modular architecture, consisting of three modules that correspond to the tasks of the Shared task - named entity recognition, entity normalization and multilingual entity linking.

Results of the previous challenges on named entity recognition show that fine-tuning of a large language model leads to the best overall result. Using this approach in previous Shared task the best result was achieved by Arkhipov et al. (2019), allowing to reach 87.2 F1 score for Bulgarian, 87.3 F1 for Russian, 93.2 F1 for Polish and 93.9 F1 score for Czech.

For Named Entity linking task we use a dynamic knowledge base, which is built at run-time using identified entity mentions and their embeddings, similar to Yamada et al. (2016). Our model uses pre-trained LaBSE model (Feng et al., 2020) to obtain aligned embeddings in different languages. We achieve average F1 score of 48.8 (us election 2020 dataset F1 score 51.98 and covid-19 dataset F1 score 42.3).

## 2 Data Preparation

We use data provided by Shared Task organizers to train the named entity recognition (NER) component. The data consists of raw text documents with some metadata (language, source, title, date, file-id) and annotation documents. Each annotation file contains a file-id linking it to the respective text document and list of Named Entities present in document.

In order to train NER, we transformed data into conll2003 like format. At first, raw documents were split into sentences using nltk library (Bird et al., 2009). Language specific nltk models were used for sentence segmentation where they were available and Russian model was applied when language specific models were not available.

Each token in sentence is labeled as either belonging to one of the named entity classes used in this task or labeled with label "O". Although in this dataset documents are categorized into 5 topics - "asia-bibi", "brexit", "nord-stream", "ryanair" and "other", we train single model for all topics.

As an additional data for training, we ex-



*Proceedings of the 8th BSNLP Workshop on Balto-Slavic Natural Language Processing*, pages 93–97

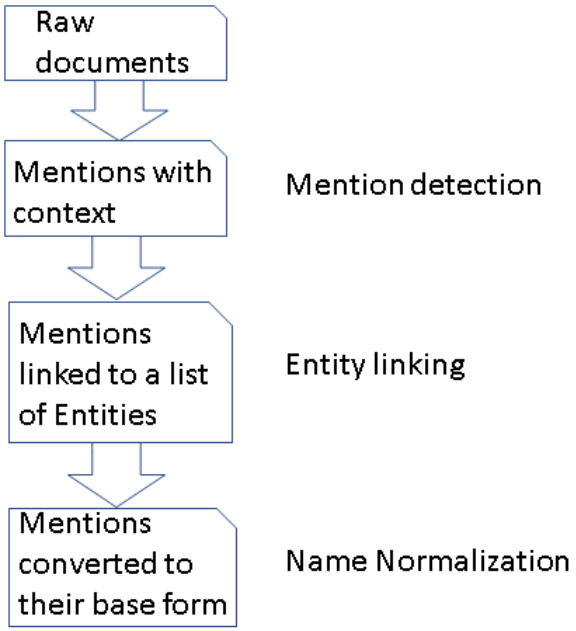


Figure 1: Overall System Architecture

plored various data sets in languages covered by bsnlp2021, however, in our opinion none of publicly available data sets match entity types required in this task. We found a data set in Latvian (Znotiņš, 2015), which includes the same entities as this task (person, location, organization, event and product). Since Latvian is also a morphologically rich inflected language, we decided to train a NER system using this data in addition to the data provided by shared task organizers.

## 3 Architecture and Modules

The architecture of our system is modular, consisting of three modules (Figure 1). At first, NER component identifies candidate entity mentions. Then, entity mentions are linked to already found entity mentions or added as new entity to a list. Finally, a base form for a given entity mention is obtained.

### 3.1 Mention Detection

We consider mention detection as sequence labeling problem aiming at labeling each token of the sequence. In our implementation we modify BERT (Devlin et al., 2019) model by adding dense layer and CRF layer on top of BERT model for Named Entity detection. We use multilingual BERT[1] provided by Google to fine-tune a single model for all six languages (Bulgarian,Czech,Polish,Russian, Slovene and Ukrainian) covered by the shared task.

[1]https://github.com/google-research/bert/blob/master/multilingual.md

| | bg | cs | pl | ru | sl | uk |
|---|---|---|---|---|---|---|
| EVT | 96.7 | 98.5 | 96.1 | 84.9 | 94.0 | 89.3 |
| LOC | 98.6 | 96.7 | 98.4 | 88.2 | 94.3 | 95.9 |
| ORG | 96.5 | 92.5 | 94.4 | 95.4 | 83.5 | 95.0 |
| PER | 97.3 | 96.1 | 96.6 | 95.2 | 92.0 | 96.9 |
| PRO | 85.6 | 94.9 | 89.2 | 61.5 | 66.6 | 66.6 |
| ALL | 96.8 | 95.4 | 96.0 | 87.3 | 89.4 | 95.3 |

Table 1: Internal evaluation results.

Provided data were split into 90% training dataset and 10% test set. We train single NER model using data from all 6 languages and evaluate for each language separately. Results of this internal evaluation are summarized in Table 1. This named entity mention recognition model achieved average F1 score of 93.

We expected that the test data for shared task will be crawled from Web and thus input may be noisy and include lowercase, uppercase and mixed case named entity mentions. Knowing that state-of-the-art NER models demonstrate good performance on grammatically correct datasets, while performing poorly on noisy data, in particular, data containing capitalization errors (Mayhew et al., 2020), we augmented training data with their noisy copies. To minimize impact of noisy data on NER performance, we augment training data using method described by Bodapati et al. (2019), i.e., using upper-cased and lower-cased text variants. We prepared four datasets for training:

- original data (TLD1),
- original data augmented with casing variants (TLD2),
- original data + Latvian (TLD3),
- original data + Latvian and augmented with casing variants (TLD4).

We trained four corresponding systems which were used in final evaluation (Table 2). The best performing system was TLD1. System TLD3 was trained using additional Latvian corpus, which allowed to detect Event and Person entities better, however this was not enough to reach better overall result. Systems TLD2 and TLD4 were trained on datasets augmented with their lower-cased and upper-cased versions. The augmentation with noisy data lead to performance decrease by 2.6 F1 points for both systems, apparently because there are few casing errors in the test data. Detection of Product and Event

| | TLD1 | TLD2 | TLD3 | TLD4 |
|---|---|---|---|---|
| PER | 92.0 | 91.0 | **92.3** | 91.4 |
| LOC | **92.9** | 90.9 | 92.6 | 91.1 |
| ORG | **78.7** | 75.1 | 77.8 | 75.3 |
| PRO | **59.3** | 51.0 | 58.1 | 48.2 |
| EVT | 26.2 | 24.1 | **31.6** | 20.6 |
| All | **83.7** | 81.1 | 83.5 | 80.9 |

Table 2: Four systems evaluated on shared task test data (Relaxed partial matching)

| | bg | cs | pl | ru | sl | uk |
|---|---|---|---|---|---|---|
| PER | 89.3 | 95.7 | 93.9 | 87.6 | 95.5 | 96.5 |
| LOC | 94.5 | 93.5 | 96.4 | 91.2 | 91.4 | 95.1 |
| ORG | 79.7 | 86.6 | 83.7 | 72.7 | 81.7 | 81.0 |
| PRO | 61.1 | 66.8 | 75.7 | 51.4 | 64.3 | 52.6 |
| EVT | 23.6 | 18.0 | 37.5 | 21.2 | 41.8 | 09.8 |
| All | 83.9 | 85.7 | 88.8 | 79.1 | 87.1 | 82.7 |

Table 3: Evaluation results for TLD1 system on shared task test data (Relaxed partial matching)

entities is poor for all systems, as they mostly failed to detect unseen events and products (e.g., covid, sputnik, coronavirus, inauguration, election).

Table 3 provides more detailed evaluation results of TLD1 system. The system performs better on Slovene, Polish and Czech texts which have Latin script, while for Bulgarian, Russian and Ukrainian which use Cyrillic script results are lower, still acceptable. For all languages Event type is poorly identified. It could be explained by entities from medicine domain (e.g., covid-19) which were not part of the training data and thus were the most challenging for our recognizer. Relatively poor results for Product and Event detection in Russian and Ukrainian can be partially explained by the fact that evaluation script rejected entities without quotation marks (e.g. Sputnik V is considered wrong, since "Sputnik V" is expected).

### 3.2 Entity Linking

The goal of the entity linking task is to associate entity mentions found in a text with corresponding entries in a Knowledge Base (KB) (Zheng et al., 2010). Entity linking consists of three sub-tasks: candidate generation, candidate ranking and unlinkable mention prediction (Shen et al., 2015). When performed without a knowledge base, entity linking reduces to entity coreference resolution, where entity mentions across one or multiple documents are clustered into multiple clusters, each representing specific entity, based on the entity mention and context. For entity linking we use mention-ranking model (Rahman and Ng, 2009) to decide whether or not an active mention is coreferent with a candidate antecedent.

| | Recall | Precision | F score |
|---|---|---|---|
| PER | 38.491 | 88.832 | 53.710 |
| LOC | 64.654 | 79.818 | 71.440 |
| ORG | 22.625 | 51.409 | 31.421 |
| PRO | 12.924 | 34.134 | 18.749 |
| EVT | 00.516 | 40.675 | 01.019 |
| All | 35.563 | 77.994 | 48.851 |

Table 4: Entity linking results evaluated on SlavNER test data (Document level, system TLD3)

For each entity mention we obtain embedding using language-agnostic BERT sentence embedding LaBSE. Each candidate mention is compared with entities already in a linked entities list by calculating cosine similarity score. We use entity type information as a hard consistency check (which filters out mentions which do not have the same type) (Khosla and Rose, 2020).

We use two similarity thresholds as hyperparameters: one for early stopping if cosine similarity is over 0.95 and second for unlinkable entity detection, set at 0.6. Entity having similarity score higher than early stopping value, is considered to be the same entity as candidate antecedent and no further comparison is needed. Entity with lower similarity score than unlinkable threshold for any antecedent, is considered as new and added to a list of entities found. For entities with similarity scores between these two hyperparameters, the most similar entity is selected and linked as correct entity.

The best results in entity linking task achieved TLD3 system. Evaluation results for this system are summarized in Table 4. Since this task depends on the results of mention detection task, results for Product and Event classes are poor. We can observe reasonable or even good precision, while recall is very poor for almost all entity types.

### 3.3 Entity Normalization

For entity normalization we apply Stanza (Qi et al., 2020) language specific lemmatizers. Since Stanza performs lemmatization on word level each word in multi-word named entity is lemmatized separately. Such approach was useful for person lemmatization, however failed for other categories of named

entities, in particular long organization names.

Evaluation results for the normalization task are summarized in Table 5. In almost all cases the system trained on data provided by shared task organizers (TLD1) achieved the highest F-score. For most of languages results are between 43 (Czech) and 52 (Ukrainian), except Bulgarian with only 15 F-Score. The reasons for such low results are multiple: first, errors in detection entity mentions automatically translate into missing normalized forms for normalization; second, multi-word entities are normalized by converting each word to its base form and third, stanza models used for have varying performance in different languages.

| | TLD1 | TLD2 | TLD3 | TLD4 |
|---|---|---|---|---|
| bg | 15.76 | 14.80 | **15.85** | 13.71 |
| cs | **43.36** | 40.56 | 42.74 | 40.03 |
| pl | **48.40** | 45.94 | 47.45 | 45.96 |
| ru | **44.12** | 42.20 | 43.64 | 42.28 |
| sl | **32.07** | 30.31 | 31.57 | 29.92 |
| uk | **52.10** | 49.78 | 50.71 | 50.25 |

Table 5: Evaluation results (F-score) for normalization task

## 4 Conclusion

In this paper we proposed modular architecture that allows to find Named Entities in six Slavic languages and links identified entities to the same entity in other documents in different languages. Each module can be separately updated to improve system performance.

As the next step, constituent modules of this system could be improved, for example with entity normalization rules for multi-word entities or implementing longer context for obtaining entity embeddings for linking.

## Acknowledgments

The research has been supported by the European Regional Development Fund within the research project "Multilingual Artificial Intelligence Based Human Computer Interaction" No. 1.1.1.1/18/A/148.

## References

Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly, Beijing.

Sravan Bodapati, Hyokun Yun, and Yaser Al-Onaizan. 2019. Robustness to capitalization errors in named entity recognition. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 237–242, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding.

Sopan Khosla and Carolyn Rose. 2020. Using type information to improve entity coreference resolution. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 20–31, Online. Association for Computational Linguistics.

Stephen Mayhew, Nitish Gupta, and Dan Roth. 2020. Robust named entity recognition with truecasing pretraining. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8480–8487. AAAI Press.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Altaf Rahman and Vincent Ng. 2009. Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 968–977, Singapore. Association for Computational Linguistics.

W. Shen, J. Wang, and J. Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint learning of the embedding of words and entities for named entity disambiguation. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 250–259, Berlin, Germany. Association for Computational Linguistics.

Zhicheng Zheng, Fangtao Li, Minlie Huang, and Xiaoyan Zhu. 2010. Learning to link entities with knowledge base. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 483–491, Los Angeles, California. Association for Computational Linguistics.

Artūrs Znotiņš. 2015. NLP-PIPE: Latvian NLP tool pipeline. CLARIN-LV digital library at MII, Universiy of Latvia.

Artūrs Znotiņš and Guntis Barzdins. 2020. *LVBERT: Transformer-Based Model for Latvian Language Understanding*, pages 111–115. IOS Press Ebooks.