Multilingual Named Entity Recognition and Matching Using BERT and Dedupe for Slavic Languages

Marko Prelevikj
Laboratory for Data Technologies,
University of Ljubljana,
Ljubljana, Slovenia
marko.prelevikj@gmail.com

Slavko Žitnik
Laboratory for Data Technologies,
University of Ljubljana,
Ljubljana, Slovenia
slavko.zitnik@fri.uni-lj.si

Abstract

This paper describes the University of Ljubljana (UL FRI) Group's submissions to the shared task at the Balto-Slavic Natural Language Processing (BSNLP) 2021 Workshop. We experiment with multiple BERT-based models, pre-trained in multi-lingual, Croatian-Slovene-English and Slovene-only data. We perform training iteratively and on the concatenated data of previously available NER datasets. For the normalization task we use Stanza lemmatizer, while for entity matching we implemented a baseline using the Dedupe library. The performance of evaluations suggests that multi-source settings outperform less-resourced approaches. The best NER models achieve 0.91 F-score on Slovene training data splits while the best official submission achieved F-scores of 0.84 and 0.78 for relaxed partial matching and strict settings, respectively. In multi-lingual NER setting we achieve F-scores of 0.82 and 0.74.

1 Introduction

Natural Language Processing (NLP) has recently become more popular, especially due to advancements in deep learning that helped to achieve superior results for various NLP tasks. One of the main issues - adaptation of models to low-resourced languages has been nicely addressed by multi-lingual and large pretrained models. Languages such as English, German, Spanish are very well developed, with a lot of resources, while on the other hand, under-developed languages, such as the Slavic languages, have less annotated resources. Still, research may show that highly inflected languages contain more semantic information that could be exploited using NLP in the future.

In this paper we explore existing resources for Slavic languages, more particularly, Slovene. We use novel BERT-based approaches (Devlin et al., 2018) and pretrained models, such as the CroSloEngual BERT (Ulčar and Robnik-Šikonja, 2020). We especially focused into how well they perform for the NER task which takes part in BSNLP 2021, and how we can improve its performance with pre-existing data sets, such as ssj500k (Krek et al., 2019) and BSNLP 2017 datasets.

Additionally, we provide a simple strategy for training the Dedupe method (Gregg and Eder, 2019) as part of the Entity Matching task. The strategy is based on choosing positive and negative examples for Dedupe, which manages to outperform the baseline of assigning a single cluster to all entities on a language level. Finally, we utilize NLP pipelines to perform PoS tagging of the documents and bring the named entities in a normalized form.

The paper is organized as follows. We first describe the Shared Task (Section 2), then present our method variations for named entity recognition and matching (Section 3). We conclude the paper with a subset of results and discussion (Section 4).

2 Task Description

2.1 Tasks

The BSNLP 2021 challenge consists of the three tasks: (a) Named Entity Detection and Classification, also known as Named Entity Recognition (NER), (b) Name Lemmatization (NL), and (c) Entity Matching (EM). The NER task focuses on detection and classification of named entities (NEs) in 5 classes: persons (PER), organizations (ORG), locations (LOC), events (EVT), and products (PRO). The purpose of the NL task is to compute the normal form of the detected named entities. Finally, the EM task connects all NE mentions that refer to the same entity with a unique identifier across the data set. The identifier should be the same on document-level, language-level, and cross-

Proceedings of the 8th BSNLP Workshop on Balto-Slavic Natural Language Processing, pages 80–85

language level.

2.2 Data

The provided data sets consist of two sets of data: (a) the raw documents retrieved from the Web, and (b) the annotations for each document, respectively. Further details about the data can be found on the BSNLP 2021 shared task official web page[1].

The data sets focus on various topics, such as Donald Trump, European Commission, Nord Stream, etc. Topics cover a variety of named entities, and thus it is worth exploring and comparing how the mentions are occurring across different languages.

2.3 Our Work

We have made an effort to participate in all of the described tasks. We detect and extract the NEs from the raw documents, normalize them, and try to link the mentions which correspond to the same entity. However, our focus lies within the scope of the NER task. More specifically, we are interested in the performance of our models on the Slovene part of the data sets, as our group specializes on improving the tools for digitization of the Slovene language.

3 Methods

In this section we describe the methodology we used to obtain our submission results. To do so, we built a pipeline which first pre-processes the provided data in a format, appropriate for training. Pre-processed data is then used to train NER models, normalize the NE occurrences, and find matches among them. In the last step we merge all the obtained results to create a BSNLP-compatible output.

3.1 Data Preprocessing and Entity Normalization

The provided data consists of two distinct parts: the raw documents, and the annotations for the corresponding documents. The first part of our pipeline merges both parts together. To do so, we tokenize the contents of the raw documents and detect the annotations of the NE within the tokens. The NE annotations are following the widely used IOB2 format (Ratnaparkhi, 1998). The tokenizers split the documents into sentences with corresponding words (tokens). So, in the end we obtain the tokens, the sentence they are a part of, and the ground truth annotation of the NEs.

We use the CLASSLA[2](Ljubešić, 2020) NLP pipeline, which we applied to the Slovene and Bulgarian parts of the provided data sets. For the rest of the languages we use Stanza library (Qi et al., 2020). These tokenizers are related: CLASSLA is a fork of the Stanza project, which specializes in Balkan languages. The tokenizers also provide processors for lemmatization of the tokens, and Part of Speech (PoS) taggers. Since our focus is the NER task, we used the lemmatization provided from the tokenizers as our NL submission. The PoS information is later used for the EM task.

3.2 Named Entity Recognition

As previously mentioned, we focused our efforts in exploring the performance of NER models for the Slovene language. Furthermore, we were interested in the performance of contemporary BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) based models. We further trained the pre-trained models for the NER downstream task.

From the large set of BERT-based models we used (a) the BERT-Base Multilingual-Cased[3] (M-Cased) and (b) BERT-Base Multilingual-Uncased[4] (M-Uncased) models which are both pre-trained by the original BERT authors (Devlin et al., 2018) on multiple languages. Additionally, we use (c) the CroSloEngual BERT[5] (Ulčar and Robnik-Šikonja, 2020), which is pre-trained on the Croatian, Slovene, and English languages and (d) the recently published SloBERTa 1.0[6] and (e) SloBERTa 2.0[7] models, which are pre-trained for the Slovene language only.

We use the following hyper-parameters for all of the models:

- epochs: 5,
- batch size: 32,
- maximum sentence size: 128 tokens,
- AdamW Optimizer learning rate: $3e-5$, and
- AdamW Optimzer epsilon: $1e-8$.

[1] http://bsnlp.cs.helsinki.fi/shared-task.html (Accessed: March 6, 2021).

[2] https://github.com/clarinsi/classla-stanfordnlp

[3] https://huggingface.co/bert-base-multilingual-cased

[4] https://huggingface.co/bert-base-multilingual-uncased

[5] http://hdl.handle.net/11356/1330

[6] http://hdl.handle.net/11356/1387

[7] http://hdl.handle.net/11356/1397

We trained multiple bundles of models, which we further describe in Appendix A. We use the standard F_1 based evaluation metrics, proposed in the BSNLP shared task guidelines: Relaxed Partial (RP), Relaxed Exact (RE), and Strict (S). Additionally, we use the Seqeval's (Nakayama, 2018) micro F_1 score to perform evaluation of the models as we are training them.

3.3 Entity Matching

For the entity matching task, we used the Dedupe (Gregg and Eder, 2019) library, which links the NEs based on provided positive and negative examples. The Dedupe takes into account additional attributes, apart from the NE mention. In our case, we use the lemma and the PoS tag obtained by the tokenizer.

Since the Dedupe approach is very memory greedy, we first group all the mentions by their first character. Then, we group all of the ground-truth mentions by their EM identifier, from which we sample $K = 3$ random chosen occurrences which are used as positive examples. The more demanding part is generating negative samples, which can grow exponentially when computing all of the possible combinations of non-matching NEs. To reduce this number, we toss a fair coin ($P(Heads) = P(Tails) = \frac{1}{2}$), in order to determine whether to keep the combination or not. Whenever the coin determines to keep the negative example, we use the same procedure for generating the negative examples as we do for the positive examples.

4 Results and Discussion

The training of the NER models is divided into multiple training bundles (see Appendix A). First we evaluated our models for all languages on a test set from each language - Table 1. The second training bundle focuses on the performance of the models on each (unseen) data set. The models were trained on three topics in a corpus and evaluated on the fourth one - the results are shown in Table 2. All these models were fine tuned based on M-Cased model as it is case-sensitive and it covers all of the languages from the BSNLP 2021 shared task.

The Slovene-based bundles try to identify the best model for the NER task. With the Slovene-All bundle we establish a baseline for the performance of the chosen models on the Slovene portion of the BSNLP 2021 data set. The results for the Slovene-

Lang	RP	RE	S	N
all	0.96	0.92	0.93	0.93
bg	0.97	0.95	0.95	0.95
cs	0.97	0.93	0.94	0.94
pl	0.96	0.92	0.93	0.93
ru	0.94	0.88	0.88	0.88
sl	0.92	0.86	0.89	0.89
uk	0.94	0.87	0.89	0.89
all	0.82	0.74	0.74	0.45
bg	0.85	0.80	0.78	0.22
cs	0.83	0.78	0.77	0.43
pl	0.86	0.81	0.82	0.49
ru	0.79	0.68	0.68	0.47
sl	0.84	0.76	0.78	0.42
uk	0.81	0.75	0.75	0.55

Table 1: Results for the Multilingual training bundle per language. All of the models were trained in FF mode. The language notation means that the corresponding model was trained and tested on that portion of the BSNLP 2021 data set. The table also contains the Normalization (N) score for the Name Lemmatization task. The bottom part shows official shared task results.

Set	RP	RE	S	N
N.Stream	0.95	0.92	0.93	0.93
Ryanair	0.92	0.88	0.88	0.88
Brexit	0.92	0.85	0.87	0.87
A.Bibi	0.89	0.81	0.82	0.82

Table 2: Results for the "Leave-One-Out" training bundle for all languages. We trained the M-Cased model in FF training mode for all data sets, except the one we tested. The table also contains the Normalization (N) score for the Name Lemmatization task.

Model	Mode	RP	RE	S
CroSloEngual	FC	**0.93**	0.87	0.90
SloBERTa 1.0	FF	0.93	**0.88**	0.90
SloBERTa 2.0	FC	0.93	0.87	0.89
SloBERTa 2.0	FF	0.93	0.87	0.89
CroSloEngual	FF	0.93	0.88	**0.90**
SloBERTa 1.0	FC	0.93	0.87	0.89
M-Cased	FF	0.93	0.87	0.89
M-Cased	FC	0.92	0.88	0.90
M-Uncased	FF	0.76	0.65	0.66
M-Uncased	FC	0.76	0.64	0.65

Table 3: Results for the Slovene-All training bundle.

Model	M	P	R	F_1
SloBERTa 1.0	FF	0.90	**0.88**	**0.89**
CroSloEngual	FC	0.90	0.87	0.88
SloBERTa 1.0	FC	0.90	0.86	0.88
M-Cased	FC	**0.91**	0.85	0.88
CroSloEngual	FF	0.89	0.86	0.88
M-Cased	FF	0.88	0.84	0.86
SloBERTa 2.0	FF	0.87	0.82	0.84
SloBERTa 2.0	FC	0.87	0.82	0.84
M-Uncased	FC	0.88	0.79	0.83
M-Uncased	FF	0.87	0.78	0.83

Table 4: Results for the Slovene Miscellaneous Only training bundle after applying the best SMO model. Training on the Slovene part of the BSNLP 2021 data set only on the EVT and PRO NE categories.

All bundle are presented in Table 3 which suggest that the CroSloEngual or the SloBERTa models are the best models to be used, which significantly outperform the M-Cased and M-Uncased models, which we take as a baseline in our case.

The most "complex" training bundle is the Slovene-Miscellaneous bundle is trained two-fold: first on 4-category data sets, and then uses an additional model to split the MISC class into PRO and EVT. For this purpose, we first explored the performance of the MISC classifier using the Slovene-Miscellaneous-Only training bundle. The results of the SMO bundle are shown in Table 4. Later we use only the best model from the SMO bundle to predict the MISC class of the models in the SM bundle. The results for the SM bundle are shown in Table 5.

The results from Table 5 also display the impact of additional data sets when training a Slovene model for the BSNLP data sets. The best-performing model (SloBERTa 1.0) is trained on all available BSNLP documents (a union of the 2017 and 2021 data sets), which suggests that more data can lead to better results. The rest of the trained models listed in Table 5 use the additionally introduced data set - ssj500k. Note that in this setting we show results for all 5 NE categories. We trained the models in two settings: (a) *Combination* setting takes all data at one and trains the model, while (b) *Iterative* first trains the model on one dataset, then continues training on another, etc. The models in Table 5 are first trained on 4 NE classes and then additional model is employed to further classify MISC classes into PRO and EVT (using best model from SMO bundle) to get the results.

By comparing the results from Table 3 and Table 5 we can notice that the two-step prediction (i.e. 4 NEs first and then additionally classifying PRO and EVT) of the NEs can be beneficial as it increases the strict matching F_1 score. Still, the increase does not seem to be significant, so in practice it can also be omitted to utilize less computing resources.

For the EM task, we obtained the results shown in Table 7. We have managed to improve over the 2 out of 3 metrics of the baseline. The baseline in our case is a single entity ID for all NE occurrences. We notice that with our approach we greatly improve the precision, but fail to identify many of the related entities, i.e., very low recall score.

For the participation in the Shared Task we submitted 5 system responses - one multi-language and others only for the Slovene language. In Table 6 we show results of the best performing system. Other systems achieved similar results with up to 1% difference, so there were no significant differences among them. The bottom part of Table 1 and Table 7 contain results of a M-Cased model, trained in a FF setting. We observe that the results of a multi-lingual model are much lower than our initial results. Also in the shared task data, results for the US election 2020 data set are in general better (i.e. 5-10%) than for Covid-19 dataset.

The code to reproduce the experiment results and re-train the models is publicly available in our public GitHub repository[8].

5 Conclusion

We have built a pipeline that addresses the BSNLP 2021 shared task. The pipeline normalizes the in-

[8] https://github.com/UL-FRI-Zitnik/BSNLP-2021-Shared-Task

Model	Training data sets					Metrics		
	ssj500k	2017	2021	2017 & 2021	Train	RP	RE	S
SloBERTa 1.0				X	Comb	0.94	0.89	0.91
CroSloEngual	X		X		Iter	0.93	0.88	0.91
SloBERTa 2.0	X		X		Iter	0.93	0.88	0.90
M-Cased	X			X	Iter	0.91	0.88	0.89
M-Uncased	X			X	Comb	0.78	0.68	0.69

Table 5: Results for the Slovene-Miscellaneous training bundle. The shown results are for the models trained in the FF learning mode. These results show the best training combinations per the RP metric of each used pre-trained model. The 2017, 2021 refer to BSNLP data sets.

SloBERTa 1.0 2017 & 2021 Comb FF	US Election 2020				Covid 2019				All Slovene data			
	RP	RE	S	N	RP	RE	S	N	RP	RE	S	N
	0.89	0.82	0.85	0.44	0.78	0.69	0.68	0.40	0.84	0.76	0.78	0.42

Table 6: Official results for the Slovene part of BSNLP 2021 Shared Task data. Apart from the model in the table, the following models were evaluated and performed similarly: (a) CroSloEngual ssj500k + 2021 Iter FC, (b) CroSloEngual ssj500k + 2021 Iter FF and (c) SloBERTa 2.0 ssj500k + 2021 Iter FF.

	Level	P	R	F_1
Baseline	Document	0.15	0.91	0.26
	Language	0.13	0.82	0.22
	Cross-Lang	0.11	0.81	0.19
Dedupe	Document	0.76	0.24	0.37
	Language	0.88	0.55	0.68
	Cross-Lang	0.83	0.09	0.17
Shared Task results	Document	0.72	0.20	0.32
	Language	0.76	0.14	0.24
	Cross-Lang	0.61	0.00	0.01

Table 7: Results for the EM task. The baseline consists of a single cluster for all of the entities.

puts, builds NER models, builds EM models, and persists the results. All the source code is publicly available. Furthermore, we have explored multiple options on how to improve the Slovene language models using more data and selecting specialized models for the targeted language. Finally, we present a simple strategy for Entity Matching which can be used for Language-Level linking of NEs.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Forest Gregg and Derek Eder. 2019. Dedupe. https://github.com/dedupeio/dedupe.

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Polona Gantar, Taja Kuzman, Jaka Čibej, Špela Arhar Holdt, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek, and Anja Zajc. 2019. Training corpus ssj500k 2.2. Slovenian language resource repository CLARIN.SI.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Nikola Ljubešić. 2020. The CLASSLA-StanfordNLP model for lemmatisation of non-standard slovenian 1.1. Slovenian language resource repository CLARIN.SI.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

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*.

Adwait Ratnaparkhi. 1998. Maximum entropy models for natural language ambiguity resolution.

Matej Ulčar and Marko Robnik-Šikonja. 2020. Finest bert and crosloengual bert: less is more in multilingual models.

A NER Training Strategy

Training of the NER models depends on a number of factors which include the base model used, the

data sets used for training and the target language for training. The provided data is split into 20% testing data, and 80% training data. The training data is further split into 10% (8% of all data) for validation, leaving 72% of all data for training.

Base Model We use 5 distinct BERT-based models for the training process. The majority of the models are focused towards the Slovene language.

Training Mode We differentiate between propagating the changes only in the fully-connected layer (FC learning), while the pre-trained model is intact, and fully fine-tuning the entire model in addition to training the fully-connected layer on top (FF learning).

Training Data Used The BSNLP shared task data is split per years when the challenge was held: 2017, 2019, and 2021. The data from 2017 has 4 classes (PER, ORG, LOC, and miscellaneous - MISC) which is inconsistent with the 2019 and 2021 data. The data from 2019 is a subset of the data from 2021, so we do not take the 2019 subset explicitly into account. Additionally, we enrich BSNLP's Slovene corpus with the ssj500k corpus (Krek et al., 2019), which is annotated consistently with the data from the BSNLP 2017 shared task.

Training Data Union Due to having multiple training sets, we also explore how they impact the performance of the trained models. In addition to training on a single data set, we train on unions of the data sets: sequentially training on multiple data sets, e.g., training first on ssj500k, and then BSNLP 2017, or combining the data sets first and learning on all data at once, e.g. training the model on the union of the ssj500k and BSNLP 2017 data sets.

A.1 Training bundles

We trained a significant amount of models in order to explore the effects of all aforementioned factors.

Multilingual (M) Trained only on the Multilingual Cased and Uncased models, in FF learning modes. The training and testing is performed on each language subset of the BSNLP 2021 data. We also train a model on all available languages.

Leave-One-Out (L1O) Training only on the Multilingual-Cased model in FF learning mode. In this mode we train the model in a leave-one-out-cross-validation fashion, where we rotate the excluded data set on which we evaluate the performance of the model.

Slovene-All (SA) Trained on all 5 models, in both learning modes. The training and testing data sets include only the Slovene portion of the BSNLP 2021 data set.

Slovene-Miscellaneous-Only (SMO) Trained on all 5 models, in both learning modes. The training and testing data sets include only the Slovene portion of the BSNLP 2021 data set which is reduced to only annotations for EVT and PRO classes. We build these models as an extension of the SM training bundle, in order to classify the MISC class into either EVT or PRO.

Slovene-Miscellaneous (SM) Training is performed on all 5 models, in both learning modes, including all possible combinations of the Slovene training data sets. The testing is performed on all available data sets having the EVT and PRO categories merged into MISC. Then we employ best model from the SMO bundle to further differentiate EVT and PRO classes from MISC predictions. In the end we built 100 distinct models, having 400 evaluations of the testing data sets.