Analysis of Transfer Learning for Named Entity Recognition in South-Slavic Languages

Nikola Ivačič [1]
Hanh Thi Hong Tran[1,2,3]
Boshko Koloski[1,2]
Senja Pollak[1]
[1]Jožef Stefan Institute,
[2]Jožef Stefan IPS,
1000 Ljubljana, Slovenia

Matthew Purver[1,4]
[3] University of La Rochelle
17000 La Rochelle, France
[4]School of Electronic Engineering
and Computer Science
Queen Mary University of London
London E1 4NS, UK

Abstract

This paper focuses on Named Entity Recognition for South-Slavic languages using pre-trained multilingual neural network models. We investigate whether the performance of the models for a target language can be improved by using data from closely related languages. The results show that this is not the case for the Slovene language, while for Croatian and Serbian, the results are better in selected cross-lingual settings. The most significant performance improvement is observed for the Serbian language, which has the smallest corpora, showing the potential of the method in less-resourced settings.

1 Introduction

Named Entity Recognition (NER) is one of the cornerstones of the NLP tasks and is widely used in many real-life applications, including in the news industry. In our study, we focus on South-Slavic languages and investigate whether the performance of the models for a target language can be improved by using data from closely related languages.

The research on NER has a long history. Already in the 90s, the research was performed by Grishman and Sundheim (1996), followed by Sang and De Meulder (2003); Segura-Bedmar et al. (2013), to mention a few of the early works. Early literature focused on rule-based models (Yu et al., 2020), which were based on a set of pre-defined patterns, and hand-crafted rules (e.g., LTG, NetOwl). These approaches were followed by the unsupervised methods (Collins and Singer, 1999; Nadeau et al., 2006), where no annotated data were required. The advent of machine learning algorithms opened a novel direction for NER tasks where feature engineering gained more traction (Krishnan and Manning, 2006; Mansouri et al., 2008; Liu et al., 2020). With recent advances in neural networks, NER was formulated as a sequence-labelling task and took advantage of the neural systems, especially Transformers, to minimize the effort of feature engineering (Lample et al., 2016; Tran et al., 2021). Ensemble systems that combine different machine learning (Ekbal and Saha, 2011; Saha and Ekbal, 2013) and neural representation (Tran et al., 2021) or architectures (Chiu and Nichols, 2016; Liu et al., 2018) were also under consideration. Besides rich-resourced languages (e.g., English), there is a shift to several less-resourced ones, including the Slavic family (see several organized shared tasks Piskorski et al. (2017, 2019, 2021)).

The availability of multilingual large language models and transfer learning strategies (Devlin et al., 2019) have simplified the cross-lingual transfer for a variety of NLP tasks. This opened new opportunities in the development of multilingual applications, especially in settings with limited resources. Cross-lingual learning allows for overcoming the problems with the lack of data, including in zero- and few-shot learning, where no or very small number of data for the target language is available. Moreover, getting the performance of a multilingual neural model as close as possible to the performance of a monolingual one can be very beneficial also in terms of simplicity and scalability, as a single model can be used instead of many monolingual ones. Last but not least, even if data for the target language is available, adding data in other languages can lead to an improvement in results.

Multilingual models have been used in a large number of tasks, including cross-lingual hate-speech detection (Pelicon et al., 2021b), zero-shot sentiment analysis (Pelicon et al., 2021a) as well as for NER (Arkhipov et al., 2019; Suppa and Jariabka, 2021). It was shown that the multilingual BERT transformer model outperforms the BiLSTM-CRF model for the NER task. The performance can be even further improved with a word-level CRF layer (Arkhipov et al., 2019). Nevertheless, it is also evident that XLM-Roberta outper-

Table 1: List of Used Corpora, which shows each corpus with an abbreviated name used in this paper, followed by the number of sentences, the number of tokens it contains, and lastly, its long name.

Corpus	Sentences	Tokens	Long Name
			Slovene
bsnlp	18106	400291	BSNLP 2017/21 (Piskorski et al., 2021)
500k	9483	193611	ssj500k 2.3 (Krek et al., 2021)
ewsd	2024	31233	ELEXIS-WSD 1.0 (Martelli et al., 2022)
scr	18139	391526	SentiCoref 1.0 (Žitnik, 2019)
			Croatian
bsnlp	820	18704	BSNLP 2017 and 2021 (Piskorski et al., 2021)
500k	24780	504227	hr500k 1.0 (Ljubešić et al., 2018)
			Serbian
set	3891	86726	SETimes.SR 1.0 (Batanović et al., 2018)
			Bosnian
wann	8917	199378	WikiANN / PAN-X (Rahimi et al., 2019)
			Macedonian
wann	16227	156467	WikiANN / PAN-X (Rahimi et al., 2019)

forms BERT (Suppa and Jariabka, 2021) in such tasks. The closest to our paper is the work by Prelevikj and Zitnik (2021), who showed that the monolingual NER model performance for the Slovene language is practically equal to that of a multilingual one.

In our paper, we focus on NER in Slovene, Croatian and Serbian and aim to answer the following question: does fine-tuning with related languages influence the performance of a multilingual model compared to fine-tuning only in the target language?

The rest of the paper is structured as follows. First, we present the corpora we used and how we preprocessed them, followed by their analysis. Next, we continue with presenting the methodology, where we first introduce the measures, models, hyper-parameters, and software used. Finally, we continue by evaluating the results and by presenting conclusions.

2 Data Description

In this section, we first present all the corpora used. Then, we continue with the description of the conversion of these datasets to the expected format and conclude with the corpora structure analysis.

We used the most common and established NER corpora for selected languages (see Table 1). The assumption and strategy for gathering corpora were also: "the more, the better."

We used NER tags in IOB2 (Ramshaw and Marcus, 1995) format from the CoNLL-2003 shared task (Tjong Kim Sang and De Meulder, 2003) as a common denominator for all corpora and experiments. Each corpus was first combined if split, then converted to a common format, reshuffled, and split to train/validation/test set in an 80/10/10 ratio.

We produced combined corpora by concatenating the sets without further reshuffling so that the experiments could be repeated.

Our study uses Slovene, Croatian, and Serbian as target languages. However, in addition to those, also Bosnian and Macedonian are considered as the source languages, as they are closely related.

Corpora used are presented in Table 1. Note that the ones for Slovene were obtained from BSNLP and parts of a newly published combined Training corpus SUK 1.0 (Arhar Holdt et al., 2022), which contained NER annotations (ssj500k, ELEXIS-WSD, and SentiCoref).

2.1 Data Conversion

The first obstacle was the different NER tags used in corpora. We decided to keep only the common tags: PER, LOC, and ORG. For example, the BSNLP corpus uses PRO and EVT tags, while the *wann* corpus lacks a MISC tag common to 500k training corpora. All non-common tags, including MISC, were replaced with O (outside IOB).

The second obstacle was the difference in format. BSNLP corpus, for instance, uses separate files for verbatim text and NER tags, with no positional reference between one another. We used CLASSLA (Ljubešić and Dobrovoljc, 2019) sentence segmentation and tokenization with a custom conversion script to solve this problem.

In addition, we removed a small amount (54) of very short sentences, as they were often noisy (e.g. conversion errors).

Next, we converted corpora from standard CoNLL format to CSV format with two fields:

- Sentence: whitespace separated sentence word tokens.
- NER: white space separated NER tags for each sentence word token.

Table 2: Example whitespace separated sentence word tokens with corresponding IOB2 NER tags.

Obtoženka	Asia	Bibi	zapustila	Pakistan
O	B-PER	I-PER	O	B-LOC

Finally, we split the corpus data into train, validation, and test sets.

2.2 Corpora analysis

Comparing the corpora showed the differences that could potentially be problematic for obtaining aligned model performance. Especially considering the NER tag ratios where the WikiANN automatically annotated corpora structure was standing out (see Table 3 and Figure 1). This is also one of the reasons why in our experiments, WikiANN corpora were only considered for additional training but not as target language gold standards.

Table 3: Analysis of Combined Corpora - shows each language's combined corpora number of tokens per sentence, followed by the number of NER tags per token. Finally, the PER, LOC, and ORG columns show the ratios with respect to all NER tags.

Lang.	tok./sent.	NER/tok.	PER%	LOC %	ORG %
sl	21.29	9.09%	31.70%	22.20%	34.13%
hr	20.43	7.41%	28.71%	20.55%	30.82%
sr	22.29	12.01%	29.96%	30.12%	32.35%
bs	**7.81**	**36.91%**	31.65%	29.67%	38.67%
mk	**9.64**	**28.07%**	34.89%	30.32%	34.79%

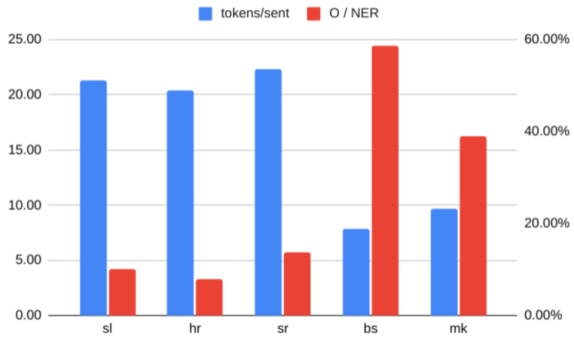

Figure 1: WikiANN corpus skew

Fortunately, we were unable to detect any inconsistencies regarding performance measurements.

3 Methodology

In the following section, we present the methodology used in our experiments to test our hypothesis that the NER classification F1-score increases when we fine-tune the pre-trained multilingual model with an additional, related language.

3.1 Method

The selected method was first to select the pre-trained embeddings, train the baseline model for each language and produce NER classification measurements. Baseline models were fine-tuned with only one - target language.

We experimented with two multilingual models, BERT multilingual base model (cased) (Devlin et al., 2018) and XLM-RoBERTa (base-sized model) (Conneau et al., 2019). However, pilot results showed better performance of XLM-RoBERTa, which was used in the final experiments presented in this paper.

Next, we combined additional language corpora, re-trained the model, and measured performance on the target language test set again. We focus only on three selected languages for evaluation, Slovene, Croatian and Serbian, but consider Bosnian and Macedonian as additional source languages.

We used the HuggingFace transformers Python library (Wolf et al., 2020) for all the experiments.

3.2 Parameters

For all the experiments, we used the following hyper-parameters:

- 256 max-length for tokenizer
- PyTorch's AdamW algorithm with 5e-5 learning rate
- batch size of 20
- 40 epochs (preliminary runs showed best F1-scores between epochs 15 and 35)
- F1-score for best model selection and training progression.

4 Evaluation

In the following section, we define the F1-score we used for evaluation. Then we present the experiment results: the evaluation of the pre-trained multilingual model, followed by the evaluation of fine-tuning for each language.

For all classification measurements, the Seqeval library (Nakayama, 2018) was used. Although the library uses CoNLL evaluation by default, we chose "strict" mode evaluation. When calculating measurements, the strict mode also considers the IOB2 tag's "beginning" and "inside" parts. Therefore the NER tags must match exactly.

4.1 Evaluation measure

For the evaluation of the classification models, we used the traditional F-measure or balanced F-score, which is the harmonic mean of precision and recall:

$$\text{F1-score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

The Precision and Recall are defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN}$$

given that:

- FP: a NER tag that is predicted but not present in the test.
- FN: a NER tag present in the test but missing in our prediction.
- TP: a NER tag that is correctly predicted.

The overall F1-score, used in the evaluation tables and figures, is a macro-averaged F1-score over all three NER tags. Macro-averaged F1-score is computed using the arithmetic mean of all the per-class F1 scores:

$$\text{Macro-averaged F1-score} = \frac{1}{n}\sum_{i=1}^{n} F1_i$$

where $F1_i$ is the F1-score for ith NER tag.

The average distance from the baseline was used as a measure to show the overall variability of different models tested with the same test set. We also report the maximum reduction in error rate achieved for each tag.

4.2 Results

Here, we present results for the three target languages.

4.2.1 Slovene

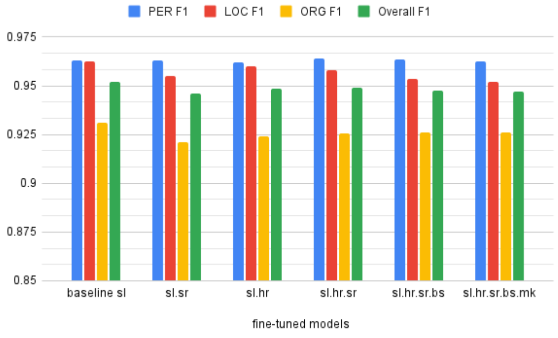

Figure 2: Slovene language test set model performance

The Slovene test set shows surprising model stability. This stability comes, assumingly, from larger corpora compared to the others. It might be that the quality of the corpora also plays a crucial role in this observation.

Table 4: Slovene language test set model performance

Model	PER F1	LOC F1	ORG F1	Overall F1
baseline sl	0.963	**0.963**	**0.931**	**0.952**
sl.sr	0.963	0.955	0.921	0.946
sl.hr	0.962	0.960	0.924	0.948
sl.hr.sr	**0.964**	0.958	0.925	0.949
sl.hr.sr.bs	**0.964**	0.953	0.926	0.948
sl.hr.sr.bs.mk	0.962	0.952	0.926	0.947
avg. dist.	0.00071	0.0070	0.0063	0.0043
error reduction	2.7%	-	-	-

If we observe the average distance from the baseline in the table's last row, we can see that it is only near 0.5%. For the PER tag, the error rate is reduced by a small amount (2.7%), but other tags are not improved.

4.2.2 Croatian

The Croatian language test set shows higher variability when tested with different models, most significantly on the ORG tag. It might be that the other corpora training is influencing variability. However, there is now some overall performance gain from the training: we can see that the average distance from the baseline is 0.5-1%, with reductions in error rates between 6 and 11%.

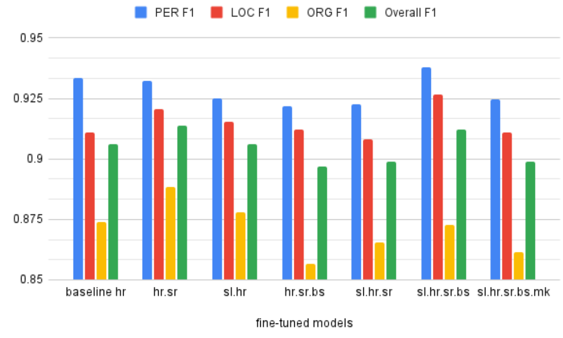

Figure 3: Croatian language test set model performance

Table 5: Croatian language test set model performance

Model	PER F1	LOC F1	ORG F1	Overall F1
baseline hr	0.934	0.911	0.874	0.906
hr.sr	0.932	0.921	**0.888**	**0.914**
sl.hr	0.925	0.915	0.878	0.906
hr.sr.bs	0.922	0.912	0.856	0.897
sl.hr.sr	0.923	0.908	0.865	0.899
sl.hr.sr.bs	**0.938**	**0.927**	0.873	0.912
sl.hr.sr.bs.mk	0.925	0.911	0.861	0.899
avg. dist.	0.0076	0.0055	0.0098	0.0062
error reduction	6.1%	18.0%	11.1%	8.5%

4.2.3 Serbian

The Serbian language test set showed the most significant increase in performance over the baseline. Its average distance in performance measurements from the baseline is from approximately 0.5% to 2.5%, with large reductions in error rate of 43%-68%. The main suspect for this phenomenon is the Serbian corpus size. It is the smallest included in this analysis, and therefore benefits most from additional cross-lingual training on other corpora.

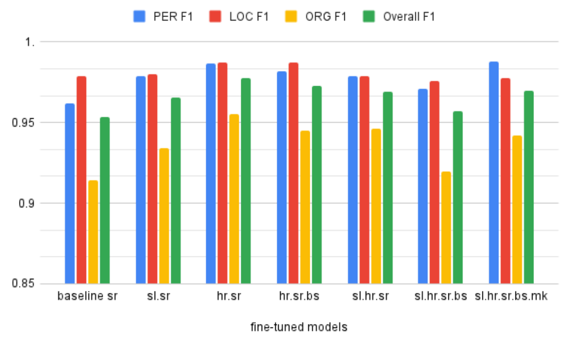

Figure 4: Serbian language test set model performance

Table 6: Serbian language test set model performance

Model	PER F1	LOC F1	ORG F1	Overall F1
baseline sr	0.962	0.979	0.914	0.954
sl.sr	0.979	0.980	0.934	0.965
hr.sr	0.987	**0.988**	**0.956**	**0.978**
hr.sr.bs	0.982	0.987	0.945	0.973
sl.hr.sr	0.979	0.979	0.946	0.969
sl.hr.sr.bs	0.971	0.976	0.920	0.957
sl.hr.sr.bs.mk	**0.988**	0.978	0.942	0.970
avg. dist.	0.019	0.0037	0.026	0.015
error reduction	68.4%	42.9%	48.8%	52.2%

5 Conclusion

We have shown that model performance can be influenced substantially by cross-lingual training with other language corpora, but that improvements only seem to occur if the target language has relatively small corpora. While for Slovene, the monolingual setting generally performs better, for Croatian and Serbian, the results are slightly better in selected cross-lingual settings. The most significant performance improvement is shown for the Serbian language, which has the smallest corpora. This indicates that fine-tuning with other closely related languages may benefit only the “low resource” languages.

Our initial hypothesis has not been fully upheld, but the result is still beneficial. First, when considering less-resourced settings, leveraging closely related languages is beneficial. Second, the performance does not degrade much if we fine-tune the model with additional language corpora from the same family. This is an important finding, as using a multilingual model in an application is a simpler solution than having several monolingual models.

In future work, we propose further investigating how performance changes when distantly related languages are used for fine-tuning the models. This will further benefit the usage in an industrial setting if the performance is not degraded, as having a single model that supports more languages with similar performance to monolingual training is more scalable and practical.

Acknowledgements

We acknowledge financial support from several sources: the Slovenian Research Agency via research core funding for the programme Knowledge Technologies (P2-0103) and the projects CANDAS (Computer-assisted multilingual news discourse analysis with contextual embeddings, J6-2581) and SOVRAG (Hate speech in contemporary conceptualizations of nationalism, racism, gender and migration, J5-3102); the UK EPSRC via the projects Sodestream (Streamlining Social Decision Making for Improved Internet Standards, EP/S033564/1) and ARCIDUCA (Annotating Reference and Coreference In Dialogue Using Conversational Agents in games, EP/W001632/1); and via the project RobaCOFI (Robust and adaptable comment filtering), which indirectly received funding from the European Union’s Horizon 2020 research and innovation action programme via the AI4Media Open Call #1 issued and executed under the AI4Media project (Grant Agreement no. 951911). This paper reflects only the authors’ views; the EC and the AI4Media project are not responsible for any use that may be made of the information contained herein.

References

Špela Arhar Holdt, Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Polona Gantar, Jaka Čibej, Eva Pori, Luka Terčon, Tina Munda, Slavko Žitnik, Nejc Robida, Neli Blagus, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Taja Kuzman, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek, and Anja

Zajc. 2022. Training corpus SUK 1.0. Slovenian language resource repository CLARIN.SI.

Mikhail Arkhipov, Maria Trofimova, Yurii 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.

Vuk Batanović, Nikola Ljubešić, Tanja Samardžić, and Tomaž Erjavec. 2018. Training corpus SETimes.SR 1.0. Slovenian language resource repository CLARIN.SI.

Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the association for computational linguistics*, 4:357–370.

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *1999 Joint SIGDAT conference on empirical methods in natural language processing and very large corpora*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

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.

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.

Asif Ekbal and Sriparna Saha. 2011. Weighted vote-based classifier ensemble for named entity recognition: a genetic algorithm-based approach. *ACM Transactions on Asian Language Information Processing (TALIP)*, 10(2):1–37.

Ralph Grishman and Beth M Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

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. 2021. Training corpus ssj500k 2.3. Slovenian language resource repository CLARIN.SI.

Vijay Krishnan and Christopher D Manning. 2006. An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics*, pages 1121–1128.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270.

Liyuan Liu, Jingbo Shang, Xiang Ren, Frank Xu, Huan Gui, Jian Peng, and Jiawei Han. 2018. Empower sequence labeling with task-aware neural language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Shifeng Liu, Yifang Sun, Bing Li, Wei Wang, and Xiang Zhao. 2020. Hamner: Headword amplified multi-span distantly supervised method for domain specific named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8401–8408.

Nikola Ljubešić, Željko Agić, Filip Klubička, Vuk Batanović, and Tomaž Erjavec. 2018. Training corpus hr500k 1.0. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić and Kaja Dobrovoljc. 2019. What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 29–34, Florence, Italy. Association for Computational Linguistics.

Alireza Mansouri, Lilly Suriani Affendey, and Ali Mamat. 2008. Named entity recognition approaches. *International Journal of Computer Science and Network Security*, 8(2):339–344.

Federico Martelli, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Sandford Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafael Ureña-Ruiz, José-Luis Sancho-Sánchez, Veronika Lipp, Tamás Váradi, András Győrffy, Simon László, Valeria Quochi, Monica Monachini, Francesca Frontini, Carole Tiberius, Rob Tempelaars, Rute Costa, Ana Salgado, Jaka Čibej, and Tina Munda. 2022. Parallel sense-annotated corpus ELEXIS-WSD 1.0. Slovenian language resource repository CLARIN.SI.

David Nadeau, Peter D Turney, and Stan Matwin. 2006. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. In *Advances in Artificial Intelligence: 19th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2006, Québec City, Québec, Canada, June 7-9, 2006. Proceedings 19*, pages 266–277. Springer.

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

Andraž Pelicon, Ravi Shekhar, Matej Martinc, Blaž Škrlj, Matthew Purver, and Senja Pollak. 2021a. Zero-shot cross-lingual content filtering: Offensive language and hate speech detection. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 30–34, Online. Association for Computational Linguistics.

Andraž Pelicon, Ravi Shekhar, Blaž Škrlj, Matthew Purver, and Senja Pollak. 2021b. Investigating cross-lingual training for offensive language detection. *PeerJ Computer Science*, 7:e559.

Jakub Piskorski, Bogdan Babych, Zara Kancheva, Olga Kanishcheva, Maria Lebedeva, Michał Marcińczuk, Preslav Nakov, Petya Osenova, Lidia Pivovarova, Senja Pollak, Pavel Přibáň, Ivaylo Radev, Marko Robnik-Sikonja, Vasyl Starko, Josef Steinberger, and Roman Yangarber. 2021. Slav-NER: the 3rd cross-lingual challenge on recognition, normalization, classification, and linking of named entities across Slavic languages. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 122–133, Kiyv, Ukraine. Association for Computational Linguistics.

Jakub Piskorski, Laska Laskova, Michał Marcińczuk, Lidia Pivovarova, Pavel Přibáň, Josef Steinberger, and Roman Yangarber. 2019. The second cross-lingual challenge on recognition, normalization, classification, and linking of named entities across Slavic languages. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 63–74, 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.

Marko Prelevikj and Slavko Zitnik. 2021. Multilingual named entity recognition and matching using BERT and dedupe for Slavic languages. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 80–85, Kiyv, Ukraine. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text chunking using transformation-based learning.

Sriparna Saha and Asif Ekbal. 2013. Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition. *Data & Knowledge Engineering*, 85:15–39.

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

Isabel Segura-Bedmar, Paloma Martínez Fernández, and María Herrero Zazo. 2013. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). Association for Computational Linguistics.

Marek Suppa and Ondrej Jariabka. 2021. Benchmarking pre-trained language models for multilingual ner: Traspas at the bsnlp2021 shared task. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 105–114.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, page 142–147, USA. Association for Computational Linguistics.

Thi Hong Hanh Tran, Antoine Doucet, Nicolas Sidere, Jose G Moreno, and Senja Pollak. 2021. Named entity recognition architecture combining contextual and global features. In *Towards Open and Trustworthy Digital Societies: 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021, Virtual Event, December 1–3, 2021, Proceedings*, pages 264–276. Springer.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. *arXiv preprint arXiv:2005.07150*.

Slavko Žitnik. 2019. Slovene corpus for aspect-based sentiment analysis - SentiCoref 1.0. Slovenian language resource repository CLARIN.SI.