Using a Frustratingly Easy Domain and Tagset Adaptation for Creating Slavic Named Entity Recognition Systems

Luis Adrián Cabrera-Diego
La Rochelle Université, L3i,
La Rochelle, 17031, France
luis.cabrera_diego@univ-lr.fr

Jose G. Moreno
Université Paul Sabatier, IRIT,
Toulouse, 31062, France
jose.moreno@irit.fr

Antoine Doucet
La Rochelle Université, L3i,
La Rochelle, 17031, France
antoine.doucet@univ-lr.fr

Abstract

We present a collection of Named Entity Recognition (NER) systems for six Slavic languages: Bulgarian, Czech, Polish, Slovenian, Russian and Ukrainian. These NER systems have been trained using different BERT models and a Frustratingly Easy Domain Adaptation (FEDA). FEDA allow us creating NER systems using multiple datasets without having to worry about whether the tagset (e.g. Location, Event, Miscellaneous, Time) in the source and target domains match, while increasing the amount of data available for training. Moreover, we boosted the prediction on named entities by marking uppercase words and predicting masked words. Participating in the 3rd Shared Task on SlavNER[1], our NER systems reached a strict micro F-score of up to 0.908. The results demonstrate good generalization, even in named entities with weak regularity, such as book titles, or entities that were never seen during the training.

1 Introduction

Named Entity Recognition (NER) is a fundamental task in domain of Natural Language Processing (NLP) that consists of extracting entities that semantically refer to aspects such as locations, people or organizations (Luoma et al., 2020). Since the creation of BERT (Devlin et al., 2019), multiple NER systems have brought the state of the art to new levels of performance. Nonetheless, there are many challenges that still need to be faced, especially in the case of less-resources languages.

In the *2nd Shared Task on SlavNER* (Piskorski et al., 2019), the top-two systems in the detection Named Entities (NEs), Tsygankova et al. (2019) and Arkhipov et al. (2019), managed to reach a relaxed partial micro F-score of 0.9, followed by two other systems with values slightly better than 0.8 (Moreno et al., 2019). For the *3rd Shared Task on SlavNER*, we consider that in order to improve the scores, in terms of the strict evaluation, and NEs related to products and events, it is necessary to include additional data that could improve the generalization of the models to any kind of topic.

While in the literature there are multiple techniques for training models over additional datasets, such as transfer learning and domain adaptation, using these techniques might pose additional questions. For example, to determine which layers to freeze, fine-tune or substitute. Furthermore, different datasets might use dissimilar tagsets, which might be incompatible (Nozza et al., 2021).

In this paper, we present the participation of laboratory *L3i* in the 3rd Shared Task on SlavNER. Specifically, we participate with multiple NER systems for Slavic languages using different BERT models and training over diverse datasets through a *Frustratingly Easy Domain Adaptation* (FEDA) algorithm (Daumé III, 2007; Kim et al., 2016).[2] The FEDA algorithm has for objective to learn common and domain-specific patterns between multiple datasets, while keeping separately patterns belonging only to the domain-specific data (Daumé III, 2007). Particularly, the use of FEDA allow us sharing the knowledge and patterns found in multiple datasets without having to worry about which different tagsets are used among them.

Apart from the FEDA algorithm, we explore some other techniques that might improve the performance of our NER system based on the ideas of Cabrera-Diego et al. (2021). Specifically, we analyze whether the marking and enrichment of uppercase tokens can improve the detection of NEs. As well, we use the prediction of masked tokens as a way to improve NER systems' generalization.

[1] bsnlp.cs.helsinki.fi/shared-task.html, last visited on 9 March 2021

[2] github.com/EMBEDDIA/NER_FEDA

Proceedings of the 8th BSNLP Workshop on Balto-Slavic Natural Language Processing, pages 98–104
April 20, 2021.

The rest of the paper is organized as follows. In Section 2, we introduce the background for the proposed work. This is followed by the methodology in Section 3. The data and the experimental settings are described in Section 4 and Section 5, respectively. In Section 6, we present the results obtained. Finally, the conclusions and future work are detailed in Section 7.

2 Background

Uppercase sentences: Although most of the NER corpora found in the literature provide texts following standard case rules, it is not infrequent to find datasets containing some sentences in which all the words are in uppercase, e.g. English CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003) or SSJ500k (Krek et al., 2019). In NLP systems based on BERT or similar, where Byte Pair Encoding (BPE) tokenizers are used, the presence of uppercase sentences might pose a greater challenge than standard case sentences. The reason is that an uppercase word have different BPE tokens with respect to its lower and title-case versions, and in consequence different dense representation (Powalski and Stanislawek, 2020; Sun et al., 2020).

Weak generalization: One of the most challenging aspects of NER systems is to deal with NEs that have a weak or zero regularity, such as names of movies, and NEs that were never seen during training (Lin et al., 2020b). Some methods found in the literature for improving generalization consists of learning manually defined triggers (Lin et al., 2020a), but also permuting NEs and reducing context such as in Lin et al. (2020b).

FEDA: Originally proposed by Daumé III (2007), the FEDA was firstly designed for sparse machine learning algorithms. Later, Kim et al. (2016), proposed a neural network version of this domain adaptation algorithm. While the former resides in duplicating input features, the latter consists of activating specific neural network layers.

3 Methodology

Consider $\mathcal{D} = \{D_1, D_2, \ldots, D_n | n > 1\}$ a collection of datasets from which we want to train a model. Furthermore, consider a classifier C a stack of two linear layers in which in between we set an activation layer ReLU and a dropout. The first linear layer has a size of 512, while the output h produced by C has a size of l, which is the number of different labels found in $\mathcal{D}$. Thus, the proposed model for doing the FEDA consists of adding on top of BERT $n+1$ classifiers such that we have $\mathcal{C} = \{C_0, C_1, C_2, \ldots, C_n\}$. The classifier C_0 represents a general classifier that will receive as input the sentences from all the datasets in $\mathcal{D}$, while $C_k \in \{\mathcal{C} | 0 < k \leq n\}$ represent a specialized classifier that will focus only on the sentences that belong to the dataset $D_k \in \{\mathcal{D} | 0 < k \leq n\}$. For each sentence belonging to a dataset D_k, we do the element-wise sum between h_0 and h_k, i.e. $H_k = h_0 + h_k$. Finally, H_k is introduced it into a CRF layer, which will determine the labels of each word in a sentence. Figure 1 depicts the proposed architecture.

For increasing the generalization of our NER systems, we explore the prediction of masked tokens during the training as proposed by Cabrera-Diego et al. (2021). Firstly, this method converts randomly selected tokens, within a sentence, into BERT's special token *[MASK]*. Then, the NER system has to predict correctly the sentence's NEs, despite the missing information, as well as predicting the masked tokens. The prediction of masked tokens is done by introducing BERT's output into a linear layer, which has the same size of the pre-trained vocabulary. During training, the loss produced by the prediction of masked tokens is added to the loss produced by the recognition of NEs; during testing, this layer is inactive.

Although Powalski and Stanislawek (2020) propose *UniCase*, an architecture for training a language model that learns the casing of a word separately to the tokenization, in this work, we use a simpler method that does not require to retrain a language model. Specifically, we use a marking and enrichment approach, where an uppercase word is tagged with two special BERT's tokens, defined by us as *[UP]* and *[up]*, and where we include additional case versions. For instance, the word "ROME" becomes "[UP] [ROM, ##E] [Rome] [r,##ome] [up]". It is important to indicate that the prediction of the NE type is done uniquely over the first token, which correspond to the special token [UP]. In other words, the output produced by BERT for the rest of the tokens is masked. The marking of the uppercase words is based on the ideas proposed by Cabrera-Diego et al. (2021).

4 Datasets

We use the data provided by the organizers for the 3rd Shared Task on SlavNER. However, for the

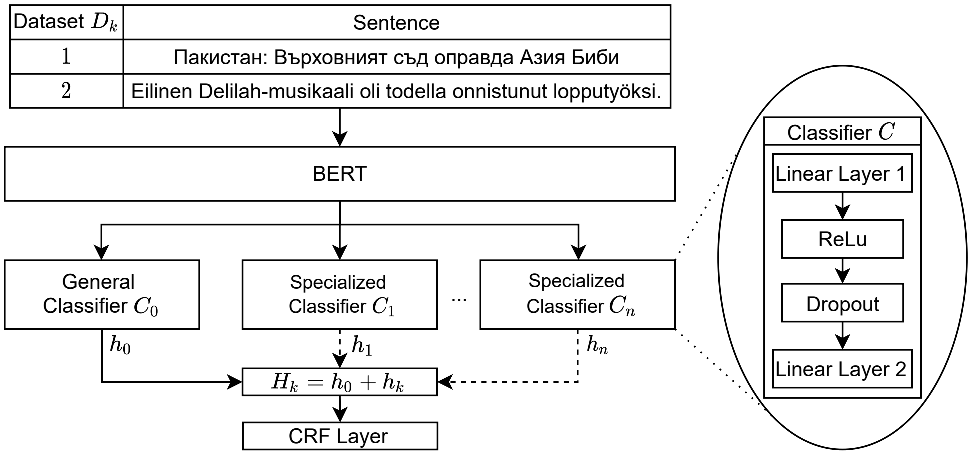

Figure 1: Our FEDA-based architecture for NER with BERT.

development of our internal models, we use the topics of *Nord Stream* and *Ryanair* as testing partition, while the rest as training and development. For the final models, all the data provided is split into training and development sets.

Besides the data provided by SlavNER's organizers, we use the follwing NER corpora:

SlavNER 2017 (Piskorski et al., 2017): Slavic Corpus annotated with 4 NE types: Location, Miscellaneous, Organization and Person.

Collection Named Entities 5 (CNE5) (Mozharova and Loukachevitch, 2016)[3]: Russian NER corpus manually annotated with five NE types: Geopolitical, Location, Media, Person and Organization.

Czech Named Entity Corpus 2.0 (CNEC) (Ševčíková et al., 2007): Czech corpus annotated with fine-grained NE. In this work, we have used 6 types of NE: Location, Organization, Media, Artifact, Person and Time.

FactRuEval[4]: Russian corpus annotated with three NE types: Location, Organization and Person.

Finnish NER (Luoma et al., 2020): Although Finnish is not a language to process in SlavNER, it has similar NE types to those used in the shared task: Date, Event, Location, Organization, Person, Product and Time. We use this dataset to enrich the NEs knowledge, specially on events and products.

National Corpus of Polish (NKJP)[5] (Przepiórkowski et al., 2012): Polish corpus tagged with five NE types: Person, Organization, Geopolitical, Location, Date and Time.

NER-UK[6]: Collection of 264 Ukrainian documents manually annotated with four types of NE: Location, Miscellaneous, Organization and Person.

Polish Corpus of Wrocław University of Technology (KPWr)[7] (Marcińczuk et al., 2016): Polish dataset annotated with nine super NE types, from these six were chosen: Event, Location, Organization, Person, Place and Product. Location and Place were merged as the former.

SSJ500k (Krek et al., 2019): Slovene corpus annotated with four types of NE: Location, Miscellaneous, Organization and Person.

Wikiann (Pan et al., 2017): It is a multilingual NER corpus based on Wikipedia articles; it was annotated automatically using three types of NEs: Location, Organization and Person. We use of the corpus partitions used by Rahimi et al. (2019).

We use for all the additional corpora their training, development and testing partitions; if these are not provided, we create them using a stratified approach to ensure a proportional number of NEs.

5 Experimental Setup

Regarding BERT, we use different pre-trained models: *CroSloEngual* (Ulčar and Robnik-Šikonja, 2020), *Polish BERT*[8], *RuBERT* (Kuratov and Arkhipov, 2019) and *Language-Agnostic BERT Sentence Embedding (LaBSE)* (Feng et al., 2020).

All the files coming from SlavNER are tokenized and, those used for training and development are annotated at token-level. For Bulgarian and Slovene, we tokenize the documents using Reldi-Tokenizer[9], while for the rest of languages, we use the neural parser proposed by Kanerva et al. (2018). Further-

[3] labinform.ru/pub/named_entities
[4] github.com/dialogue-evaluation/factRuEval-2016
[5] nkjp.pl
[6] github.com/lang-uk/ner-uk
[7] clarin-pl.eu/dspace/handle/11321/270
[8] huggingface.co/dkleczek/bert-base-polish-cased-v1
[9] github.com/clarinsi/reldi-tokeniser

more, we over-tokenize all the files, i.e. we separate all the punctuation from tokens within a sentence, to solve some cases where abbreviation periods or dashes were not considered as part of a NE. For example, in Slovene, Roman numerals are followed by a period, such as in *Benedikt XVI.* nevertheless, some NE annotations did not consider the period. Some rules and manual corrections were applied to the tokenization where we determined the fix was critical. For instance, in Polish, *W. Brytania* (Great Britain) was being split into two sentences by the tokenizer. We automatically annotated the files by searching the longest match in the tokenized format and the annotation file. In case of ambiguity, the annotation tool requested a manual intervention. For the final submission, we converted the token-level output to a document-level one.

All the NEs types are encoded using BIOES (Beginning, Inside, Outside/Other, End, Single). As well, to reduce the number of entities types, we normalize those where the theoretical meaning is the same, i.e. PERS into PER or EVENT into EVT.

For the models where masked tokens have to be predicted, we only affect sentences in the training partitions that are longer than 3 actual tokens, i.e. not BPE tokens. At each epoch, we select randomly 25% of each sentence's tokens and substitute them with *[MASK]*. If a token after being processed by BERT's tokenizer produces more than one BPE token, we mask only one of them.[10] Regarding the models that are trained with marked uppercase tokens, at each training epoch, we randomly convert 5% of all the sentences into uppercase. This is done to provide some examples of uppercase sentences to datasets that do not present this phenomenon.

In Table 2, we present the final models created for recognizing NEs. As well, we detail which are the datasets used for training them and which are the additional features that they make use. The combinations of datasets and features used for the final models were selected according to their performance on internal models. To enrich the knowledge in Bulgarian, we added the Macedonian Wikiann dataset, as both languages are considered as mutually intelligible. All the models were trained up to 20 epochs using an early stop approach. In Table 1, we present a summary of the hyperparameters used for training the NER systems.

[10]For Polish BERT, we mask all the tokens as this model was trained using whole word masking.

Hyperparameter	**Value**
Maximum Epochs	20
Early Stop Patience	2
Learning Rate	2×10^{-5}
Scheduler	Linear with warm-up
Warm-up Ratio	0.1
Optimizer	AdamW with bias correction
AdamW ϵ	1×10^{-8}
Random Seed	12
Dropout rate	0.5
Weight decay	0.01
Clipping gradient norm	1.0
BERT's Sequence Size	128
Linear Layer 1 Size	512
Training Mini-Batch:	
Latin 1 & 2	5
Ru	8
Pl	28
Others	16

Table 1: Hyperparameters used for training the models.

6 Results

In Table 3, we present the performance of our systems in terms of strict micro F-score. We can observe, that the marking of uppercase words worked better, in general, for the *Covid-19* topic, specially on the Cyrillic-2 model. As well, single language models worked better on the *Covid-19* topic, while the model Latin-1 worked better on the *U.S. Elections* topic. In most languages, the hardest NEs to predict were related to products and events due to their weak regularity or because they never appeared on the training datasets.

From a manual inspection, we have observed that multiple events were considered as products, such as *Miss USA*, *Pizzagate* and *Covid-19*. Some products were marked as organizations such as *Zoom*, *COVAX*, *Apple TV+*, although fewer organizations were tagged as products, such as *Pfizer/Moderna* and *BBC*. Nonetheless, many of these NEs could be both types depending on the context in which happen. In certain documents, organizations were marked as locations and vice-versa, such as *Ostravské Fakultní Nemocnice* (Ostrava University Hospital) and *Szpitala Wojskowego w Szczecinie* (Military Hospital in Szczecin).

We have found interesting examples regarding products despite their irregularity. For example, the Cyrillic and Latin models managed to detect partially the 2020 book "*Nelojalen: resnična zgodba nekdanjega osebnega odvetnika predsednika Donalda Trumpa*" (Disloyal: A Memoir: The True Story of the Former Personal Attorney to President Donald J. Trump). Specifically, the entity was

Model		Features	BERT Model	*C*	Training datasets
Script-based	Cyrillic-1 Cyrillic-2	None Uppercase	LaBSE	8	SlavNER-17 (Ru, Uk); SlavNER-21 (Bg, Ru, Uk); Wikiann (Bg, Mk, Ru, Uk); FactRuEval; CNE5; NER-UK; Finnish NER
	Latin-1 Latin-2	None Uppercase	LaBSE	8	SlavNER-17 (Cs, Pl, Sl); SlavNER-21 (Cs, Pl, Sl); Wikiann (Cs, Pl, Sl); SSJ500k; KPWr; CNEC; Finnish NER
Single language	Bg	Uppercase	LaBSE	5	SlavNER-21 (Bg); Wikiann (Bg, Mk); Finnish NER
	Cs	Uppercase	LaBSE	5	SlavNER-21 (Cs); Wikiann (Cs); CNEC; Finnish NER
	Pl	Mask.+Upper.	Polish BERT	5	SlavNER-21 (Pl); Wikiann (Pl); KPWr; NKJP
	Ru	Mask.+Upper.	RuBERT	5	SlavNER-21 (Ru); Wikiann (Ru); FactRuEval; CNE5
	Sl	Mask.+Upper.	CroSloEngual	4	SlavNER-21 (Sl); Wikiann (Sl); SSJ500k
	Uk	Uppercase	LaBSE	4	SlavNER-21 (Uk); Wikiann (Uk); NER-UK

Table 2: Datasets used for training each of the model explored in this work. The number of classifiers (C) consider both the general and specialized ones used in the architecture.

	Covid-19							U.S. Elections							
Model	**Bg**	**Cs**	**Pl**	**Ru**	**Sl**	**Uk**	**All**	**Bg**	**Cs**	**Pl**	**Ru**	**Sl**	**Uk**	**All**	**Global**
Cyrillic-1	0.716	0.714	0.760	0.657	0.732	0.722	0.715	0.843	0.837	0.841	0.741	0.837	0.787	0.793	0.764
Cyrillic-2	0.720	0.730	0.783	0.642	0.744	**0.727**	0.721	**0.865**	0.857	0.849	0.746	0.858	**0.813**	0.807	0.775
Latin-1	0.730	0.765	0.791	0.662	0.752	0.706	0.733	0.850	**0.890**	**0.908**	**0.762**	**0.898**	0.789	**0.824**	**0.790**
Latin-2	**0.733**	0.763	0.792	**0.666**	0.758	0.688	**0.734**	0.854	**0.890**	0.891	0.759	0.884	0.782	0.819	0.787
Single lang.	0.725	**0.766**	**0.793**	0.611	**0.775**	0.701	0.729	0.813	0.889	0.887	0.742	0.891	0.781	0.807	0.778

Table 3: Strict micro F-scores obtained by each model for every language and topic. The *Global* column is the strict micro F-score regarding all the test data.

split into two “*Nelojalen: resnična zgodba nekdanjega osebnega odvetnika predsednika*” as a product and *Donalda Trumpa* (Donald Trump) as a person. But there were some exact matches, such as the book “*Cyberwar: How Russian Hackers and Trolls Helped Elect a President*” or the document “*Preveč in nikoli dovolj: kako je moja družina ustvarila najnevarnejšega moža na svetu*” (Treaty on Measures for the Further Reduction and Limitation of Strategic Offensive Arms). Furthermore, some scientific articles were tagged as products, such as “*A Study to Evaluate Efficacy, Safety, and Immunogenicity of mRNA-1273 Vaccine in Adults Aged 18 Years and Older to Prevent COVID-19*”, although they did not appear in the gold standard.

Some models considered *BioNTech* as an organization and *Instagram* as a product despite these NEs were never seen during the training. As well, some medication-related products were correctly found such as *AZD1222*, канакинумаб (Canakinumab), *Remdesivir* or *Zithromax*, even if they did not exist on the training corpora.

We observed, specially in Cyrillic-scripted languages, that some named entities were incorrect because they were predicted without punctuation marks. For example: *Moderna Inc* vs *Moderna Inc.*, гам-ковид-вак vs «гам-ковид-вак» and спутником vs "спутником". In Latin-scripted languages, we observed the opposite although less frequently. For instance, *Roberta F. Kennedyho Jr.* vs *Roberta F. Kennedyho Jr*. In some documents the punctuation mark is included in certain NEs but not in others, such as in *Korea Płn.* vs *Korea Płn* but *Korei Płn.*.

7 Conclusions and Future Work

This work presented the participation of Laboratory *L3i* in the 3rd Shared Task on SlavNER. Specifically, we proposed a collection of BERT-based NER systems that were trained using multiple datasets through FEDA.

The results showed us that our NER systems worked better on the *U.S. Elections* topic (strict micro F-score between 0.762 and 0.908) than on the *Covid-19* topic (0.666 - 0.775). Overall, a competitive strength of our NER systems is that they managed to predict named entities occurring with weak regularity or that were never seen before.

In the future, we will apply the proposed architecture on other languages and datasets.

Acknowledgments

This work has been supported by the European Union's Horizon 2020 research and innovation program under grants 770299 (NewsEye) and 825153 (Embeddia).

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.

Luis Adrián Cabrera-Diego, Jose G. Moreno, and Antoine Doucet. 2021. Simple ways to improve NER in every language using markup. In *Proceedings of the 2nd International Workshop on Cross-lingual Event-centric Open Analytics*, Online. CEUR-WS.

Hal Daumé III. 2007. Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. 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*. ArXiv cs.CL eprint: 2007.01852.

Jenna Kanerva, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 2018. Turku Neural Parser Pipeline: An End-to-End System for the CoNLL 2018 Shared Task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics.

Young-Bum Kim, Karl Stratos, and Ruhi Sarikaya. 2016. Frustratingly Easy Neural Domain Adaptation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 387–396, Osaka, Japan. The COLING 2016 Organizing Committee.

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.

Yuri Kuratov and Mikhail Arkhipov. 2019. *Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language*. ArXiv cs.CL eprint: 1905.07213.

Bill Yuchen Lin, Dong-Ho Lee, Ming Shen, Ryan Moreno, Xiao Huang, Prashant Shiralkar, and Xiang Ren. 2020a. TriggerNER: Learning with Entity Triggers as Explanations for Named Entity Recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8503–8511, Online. Association for Computational Linguistics.

Hongyu Lin, Yaojie Lu, Jialong Tang, Xianpei Han, Le Sun, Zhicheng Wei, and Nicholas Jing Yuan. 2020b. A Rigorous Study on Named Entity Recognition: Can Fine-tuning Pretrained Model Lead to the Promised Land? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7291–7300, Online. Association for Computational Linguistics.

Jouni Luoma, Miika Oinonen, Maria Pyykönen, Veronika Laippala, and Sampo Pyysalo. 2020. A Broad-coverage Corpus for Finnish Named Entity Recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4615–4624, Marseille, France. European Language Resources Association.

Michał Marcińczuk, Marcin Oleksy, Marek Maziarz, Jan Wieczorek, Dominika Fikus, Agnieszka Turek, Michał Wolski, Tomasz Bernaś, Jan Kocoń, and Paweł Kedzia. 2016. *Polish Corpus of Wrocław University of Technology 1.2*. CLARIN-PL digital repository.

Jose G. Moreno, Elvys Linhares Pontes, Mickaël Coustaty, and Antoine Doucet. 2019. TLR at BSNLP2019: A multilingual named entity recognition system. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 83–88, Florence, Italy. Association for Computational Linguistics.

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

Debora Nozza, Pikakshi Manchanda, Elisabetta Fersini, Matteo Palmonari, and Enza Messina. 2021. LearningToAdapt with word embeddings: Domain adaptation of Named Entity Recognition systems. *Information Processing & Management*, 58(3):102537.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual Name Tagging and Linking for 282 Languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. 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.

Rafal Powalski and Tomasz Stanislawek. 2020. *Uni-Case – Rethinking Casing in Language Models*. ArXiv cs.CL eprint: 2010.11936.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk. 2012. *Narodowy korpus jezyka polskiego*. Naukowe PWN, Warsaw, Poland.

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.

Lichao Sun, Kazuma Hashimoto, Wenpeng Yin, Akari Asai, Jia Li, Philip Yu, and Caiming Xiong. 2020. *Adv-BERT: BERT is not robust on misspellings! Generating nature adversarial samples on BERT*. ArXiv cs.CL eprint: 2003.04985.

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*, pages 142–147.

Tatiana Tsygankova, Stephen Mayhew, and Dan Roth. 2019. BSNLP2019 Shared Task Submission: Multisource Neural NER Transfer. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 75–82, Florence, Italy. Association for Computational Linguistics.

Matej Ulčar and Marko Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT. In *Proceeding of the 23rd International Conference on Text, Speech, and Dialogue (TSD 2020)*, pages 104–111, Brno, Czech Republic. Springer International Publishing.

Magda Ševčíková, Zdeněk Žabokrtský, and Oldřich Krůza. 2007. Named Entities in Czech: Annotating Data and Developing NE Tagger. In *Text, Speech and Dialogue*, pages 188–195, Pilsen, Czech Republic. Springer Berlin Heidelberg.