WikiGoldSK: Annotated Dataset, Baselines and Few-Shot Learning Experiments for Slovak Named Entity Recognition

Dávid Šuba **Marek Šuppa** **Jozef Kubík** **Endre Hamerlik** **Martin Takáč**

Comenius University in Bratislava, Slovakia

Contact: marek@suppa.sk

Abstract

Named Entity Recognition (NER) is a fundamental NLP tasks with a wide range of practical applications. The performance of state-of-the-art NER methods depends on high quality manually anotated datasets which still do not exist for some languages. In this work we aim to remedy this situation in Slovak by introducing `WikiGoldSK`, the first sizable human labelled Slovak NER dataset. We benchmark it by evaluating state-of-the-art multilingual Pretrained Language Models and comparing it to the existing silver-standard Slovak NER dataset. We also conduct few-shot experiments and show that training on a sliver-standard dataset yields better results. To enable future work that can be based on Slovak NER, we release the dataset, code, as well as the trained models publicly under permissible licensing terms at[1].

1 Introduction

Named Entity Recognition (NER) is a lower-level Natural Language Processing (NLP) task in which the aim is to both identify and classify named entity expressions in text into a pre-defined set of semantic types, such as Location, Organization or Person (Goyal et al., 2018). It is a key component of many downstream NLP tasks, ranging from information extraction, machine translation, question answering to entity linking and co-reference resolution, among others. Since its introduction at MUC-6 (Grishman and Sundheim, 1996), the task has been studied extensively, usually as a form of token classification. In recent years, the advent of pre-trained language models (PLMs) combined with the availability of sufficiently large high quality NER-annotated datasets has led to the introduction of NER systems with very high reported performance, sometimes nearing human annotation quality (He et al., 2021).

As the predominant method for adapting PLMs to a specific task of interest is model fine-tuning using training data, the availability of annotated NER datasets for both the training as well as the evaluation part of the process of creating a NER system is critical. Since their creation is expensive, many works have focused on extracting multilingual silver-standard NER datasets from publicly available corpora such as Wikipedia, exploiting the link structure to locate and classify named entities (Nothman et al., 2013; Al-Rfou et al., 2015; Tsai et al., 2016; Pan et al., 2017). While these methods have yielded NER-annotated datasets of significant size, with the recent follow-up work reporting quality comparable to that of datasets created via manual annotation (Tedeschi et al., 2021), their application has multiple limitations: only a limited amount of Wikipedia text is inter-linked, mapping Wikipedia links to the pre-defined NER classes is non-trivial and their application often depends on the existence of high quality knowledge bases which may not be available for some domains and languages.

In this paper we focus on Slovak, a language of the Indo-European family, spoken by 5 million native speakers, which is still missing a manually annotated NER dataset of substantial size. To fill this gap, we propose the following contributions:

- We introduce a novel, manually annotated NER dataset called `WikiGoldSK` built by annotating articles sampled from Slovak Wikipedia and labeled with four entity classes.

- We evaluate a selection of multilingual NER baseline models on the presented dataset to compare its quality with that of existing silver-standard Slovak NER datasets.

- We treat Slovak as a low-resource language and also assess the possibility of using few-shot learning to train a Slovak NER model using a small part of the introduced dataset.

---

[1]https://github.com/NaiveNeuron/WikiGoldSK

2 Related Work

NER datasets Much of the progress in NER over the past decades can be attributed to and evidenced by the results reported on standard benchmarks, which in turn originate from shared tasks. This is because they generally provide high-quality annotation datasets, which are key both for the evaluation as well as creation of NER systems. Shared tasks were first introduced for resource-rich languages, such as English, Spanish, German and Dutch (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) and later established for other language groups, such as Indic (Rajeev Sangal and Singh, 2008) or Balto-Slavic languages (Piskorski et al., 2017, 2019, 2021). The "First Cross-Lingual Challenge on Recognition, Normalization and Matching of Named Entities in Slavic Languages" (BSNLP 2017) (Piskorski et al., 2017) is of particular relevance for our work, as to the best of our knowledge, it introduced the only publicly available human annotated Named Entity Recognition dataset based specifically on Slovak newswire. The dataset, however, consists of less than 50 human-annotated articles and can at best only be used for evaluation of Slovak NER systems but not their training.

The over-reliance on newswire text in the shared tasks has been noticed by the authors of (Balasuriya et al., 2009) who introduced the manually annotated `WikiGold` dataset based on English Wikipedia articles. Despite its limited size, it is still used as an evaluation benchmark. As our aim is also to create a manually annotated (gold-standard) dataset based on Wikipedia articles, we use `WikiGoldSK` to refer to the dataset introduced in this work.

To alleviate the need for sizeable datasets at low cost across multiple languages, various methods of automatically generated NER-annotated datasets have been introduced. In (Nothman et al., 2013) the authors introduce the WikiNER datasets, which makes use of Wikipedia articles and spans 9 languages but does not include Slovak. Utilizing a similar approach, (Pan et al., 2017) first classified English Wikipedia pages to specific entity types and then used the cross-lingual links to transfer the annotations to other languages. As not all entries are linked, the authors also utilize self-training and translation methods to match as many entries as possible. This pipeline generates a dataset that covers 282 languages and includes Slovak as well.

With roughly 50 thousand entities annotated with categories Person, Location and Organization, it is the largest publicly available Slovak NER dataset to date.

Another approach of resolving the need for a sizable training dataset is to utilize few-shot learning, in which only a couple of expertly annotated training examples are provided. Recently, the methods based on the combination of cloze-style rephrasing with language models have been shown to perform comparably to GPT-3 (Brown et al., 2020) while having significantly fewer parameters (Schick and Schütze, 2020b). We consider a variant of Pattern-Exploiting Training (Schick and Schütze, 2020a) called PETER (La Gatta et al., 2021) and to the best of our knowledge for the first time evaluate its performance for a general-purpose NER system in a specific language.

Slovak NER The prior art in Slovak NER is limited. In (Kaššák et al., 2012) the authors identified potential named entities as words with capital letters and recognized new entities by finding the entity scope through Wikipedia parsing. For the purposes of this work they also created a dataset annotated by 60 human experts totalling 1620 entities. The authors of (Maruniak, 2021) and (Lupták, 2021) worked with datasets based on more than 5000 articles extracted from Slovak Wikipedia, containing more than 15 000 entities and used multiple well-established NLP toolkits and libraries (such as SpaCy) to train NER models on this dataset. Utilizing a different datasource, (Mičo, 2019) have focused on the Twitter account of one of the biggest Slovak journal and created a dataset with 10 000 of its NER-annotated tweets and almost 16 000 entities, and used it to train a NER model which utilized both FastText (Bojanowski et al., 2016) vectors and the BiLSTM neural network architecture. Unfortunately, none of the datasets and models mentioned in the aforementioned works are publicly available.

Despite having 5 million native speakers and being one of the official languages of the European Union, there are relatively few readily available NLP tools tailored specifically for Slovak, which might be to some extent caused by its linguistic and historical closeness to the much better resourced Czech. This creates a peculiar dichotomy: Slovak has too many native speakers to be considered "low-resource" but at the same time lacks readily available labelled datasets that are a prerequisite for

	WikiANN	BSNLP2017	WikiGoldSK
# doc	N/A	49	412
# sent	30 000	741	6 696
# tok	263 516	14 400	128 944
split	2:1:1	0:0:1	7:1:2
LOC	19 643	244	4 459
PER	18 238	255	2 739
ORG	15 286	273	1 929
MISC	N/A	55	1 668

Table 1: The comparison of WikiGoldSK to the other publicly available Slovak NER datasets. The terms # doc, # sent and # tok refer to the number of documents, sentences and tokens in the specific datasets, respectively. Note that WikiANN does not provide document-level split and is not labeled with the MISC entity.

	train	dev	test
# sent	4 687	669	1 340
# tok	90 616	12 794	25 534
split size	70%	10%	20%
LOC	3 040	461	958
PER	1 892	298	549
ORG	1 361	190	378
MISC	1 184	160	324

Table 2: The frequency distribution across the WikiGoldSK's train/dev/test splits.

many standard NLP tools. The "language richness" taxonomies such as (Joshi et al., 2020) consider Slovak among the "The Rising Stars" of languages, but it is, to the best of our knowledge, one of the few in this category that lacks a sizable, manually labelled NER dataset[2]. The introduction of SlovakBERT in (Pikuliak et al., 2021) does suggest, however, that there is interest in creating Slovak-specific NLP tools and resources. Our work aims to help push this trend further.

3 Dataset

When creating the WikiGoldSK dataset, our principal aim was to create a high quality, publicly available human annotated corpora that could be used to both evaluate and build Slovak NER systems and that would be comparable to well-established benchmark datasets in other languages. To ensure the resulting dataset can be used in the future for research as well as commercial use, we sampled 412 articles from the skwiki-20210701 dump of Slovak Wikipedia, licensed under the terms of the Creative Commons Attribution-Share-Alike License 3.0. In order for an article to be included, its last change date needed to be in 2021 and its length had to fit between 500 and 5 000 characters[3]. The raw text of the articles was tokenized by the generic English spaCy tokenizer, with a manual pass over the

dataset in which the Slovak-specific tokenization mistakes were remedied.

We use the same set of tags as the CoNLL-2003 NER Shared task (Tjong Kim Sang and De Meulder, 2003), that is Location (LOC), Person (PER), Organization (ORG) and Miscellaneous (MISC), and our annotation guidelines are inspired by the ones introduced by the BSNLP 2017 shared task (Piskorski et al., 2017). The annotation was done using Prodigy[4] in which the whole dataset was preloaded with the labels predicted by the SlovakBERT model finetuned on the training part of the Slovak portion of the WikiANN dataset. The dataset was annotated by three Slovak native speakers who are also authors of this paper. Two annotators provided annotations for the full dataset whereas one annotator corrected half of the dataset. The Cohen's kappa coefficient between the first two annotators is 0.90 when compared on the token level and 0.81 for the tokens where both annotators agreed that they were not a part of a named entity. As per the benchmark established in (Landis and Koch, 1977), the coefficient values in both cases suggest "almost perfect" strength of agreement and a high quality of the annotation. To arrive at the final dataset, the ambiguities were resolved in a discussion between the annotators.

The summary statistics of the resulting dataset, along with the existing Slovak NER datasets, can be found in Table 1. As we can see, WikiGoldSK is larger than the Slovak portion of the BSNLP2017 dataset but smaller than the Slovak portion of the WikiANN dataset. At the same time, one can see that the distribution of Named Entities in WikiANN and WikiGoldSK follows the same pattern, with the order of LOC, PER, ORG holding for both datasets in terms of entity frequency, which is not the case in

---

[2]The other languages lacking a sizable, manually NER-annotated datasets are Uzbek, Georgian, Belarusian, Egyptian Arabic and Cebauano.

[3]This is motivated by the observation that long articles may shift the dataset towards their domain, whereas short articles often do not contain any named entity.

[4]https://prodi.gy/

BSNLP2017. This is probably caused by the fact that both WikiANN and `WikiGoldSK` are based on Wikipedia articles whereas BSNLP2017 is based on newswire text.

To make the dataset compatible with existing benchmarks, we also introduce a standard train/dev/test split in the 7:1:2 ratio, described in detail in Table 2. We note that the size of the test portion of `WikiGoldSK` is on the same order as that of `WikiGold` which consists of 1 696 sentences and 39 007 tokens.

4 Experiments

We conduct two types of experiments with the newly introduced dataset. First, we establish a set of baselines based on existing state-of-the-art PLMs that were pre-trained on Slovak data. Next, we emulate a low-resource setup by only using a small sample of the training set and use it to evaluate a few-shot learning approach as well.

4.1 Baselines

To evaluate a broad set of baselines on `WikiGoldSK`, we choose three well-established NLP toolkits:

- **spaCy**[5], which provides a pipeline for converting words to embedding of user's choice and then models NER as a structured prediction task,

- **Trankit** (Nguyen et al., 2021), which is based on XLM-RoBERTa (Conneau et al., 2019), provides pre-trained models for 56 languages, including Slovak, along with the ability to finetune on custom NER datasets, and

- **Transformers** (Wolf et al., 2019), which has become the standard tool for training, storing and sharing Transformer-based models and also includes readily available scripts for finetuning PLMs on NER datasets.

When it comes to the models chosen as baselines, we again chose well-established models relevant to the task of Slovak NER:

- **XLM-RoBERTa** (`XLM-R-base`), a multilingual Transformer model pretrained on text spanning 100 languages, including Slovak,

- **SlovakBERT**, the only BERT-based model specifically optimized for Slovak, which was

---

[5] https://spacy.io

pre-trained on almost 20GB of Slovak text obtained from various sources, including crawling Slovak web, and

- **mDeBERTav3** (He et al., 2021), a multilingual Transformer model pretrained on the same dataset as XLM-RoBERTa using a different training objective which leads to more efficient training and better performance on various benchmarks.

Our experiments were generally conducted by finetuning a given model using a specific NLP toolkit on a selected dataset, while utilizing the test set of the `WikiGoldSK` for evaluation. We use three datasets for finetuning: WikiANN, WikiANN combined with `WikiGoldSK` and just `WikiGoldSK`. Only the training portions of the respective datasets were used for finetuning. Additionally, we also benchmark the models trained on the `WikiGoldSK` dataset on the Slovak portion of the BSNLP2017 dataset.

4.2 Few-shot learning

To evaluate the possibility of building a Slovak NER system, we chose the PETER (PET (Schick and Schütze, 2020a) for NER) method introduced in (La Gatta et al., 2021). At its core, it uses pattern-verbalizer pairs (PVP), in which the "pattern" part converts a sentence with a token that corresponds to a named entity and creates a cloze-style phrase containing exactly one [MASK] token and the "verbalizer" maps tokens predicted by a PLM in place of [MASK] to one of the considered Named Entity classes. Each labeled sentence s is converted into $|s|$ pairs of training inputs $x = (s, t)$ where t is a particular token from the sentence we are predicting a label to; the training set then consist of pairs (x, y) where y is the ground-truth label. A separate language model M is fine-tuned for each PVP, a soft-label dataset created from unlabeled data and finally, the resulting classifier is trained on this dataset.

In our experiments, we use two PVPs below. More details can be found in Appendix B.

- $P_1((s, t))$: "s. V predchádzajúcej vete slovo t označuje entitu [MASK]." (English translation: *"s. In the previous sentence, the word t refers to a/an* [MASK] *entity.*)

- $P_2((s, t))$: "s. t je [MASK]." (English translation: *"s. t is a* [MASK].)

	WikiANN			WikiANN + WikiGoldSK			WikiGoldSK			BSNLP2017		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
spaCy												
XLM-RoBERTa	0.5639	0.7413	0.6405	0.8809	0.8973	0.8890	0.9145	0.8955	0.9049	**0.8102**	0.7722	0.7907
SlovakBERT	0.5509	0.7285	0.6274	0.8754	0.8932	0.8842	0.8889	0.9122	0.9004	0.7186	0.7704	0.7436
mDeBERTaV3	0.5925	**0.7572**	**0.6648**	0.8621	0.8855	0.8737	0.9151	0.9167	0.9159	0.8024	0.8122	0.8073
Trankit												
XLM-RoBERTa	**0.6110**	0.7020	0.6533	0.8833	0.8805	0.8819	0.8869	0.9014	0.8941	0.7882	0.8252	0.8063
Transformers												
XLM-RoBERTa	0.5247	0.7423	0.6148	0.8815	0.9018	0.8915	0.9210	0.9339	0.9274	0.7760	0.8226	0.7986
SlovakBERT	0.5265	0.7428	0.6162	**0.9020**	**0.9208**	**0.9113**	0.9179	0.9262	0.9221	0.7900	0.8278	**0.8085**
mDeBERTaV3	0.5092	0.7471	0.6056	0.8835	0.9063	0.8948	**0.9302**	**0.9412**	**0.9357**	0.7793	**0.8322**	0.8049

Table 3: The results of finetuning various baselines using the three selected NLP toolkits on three dataset combinations and evaluating on the test set of `WikiGoldSK`. The P, R and F1 refer to Precision, Recall and the F1 score, respectively. Best result per metric and dataset is boldfaced.

5 Results

The results of the evaluation of baselines can be seen in Table 3. They suggest that XLM-RoBERTa can still be considered a strong baseline, as its performance is similar to that of SlovakBERT, despite the latter being specifically trained and optimized for Slovak. Across the three NLP toolkits, we observe that the performance of Trankit is generally lower than that of spaCy and Transformers, given the same dataset. Comparing the three models finetuned either using spaCy or Transformers, Table 3 suggests that mDeBERTaV3 obtains performance that is either very similar or better than that of XLM-RoBERTa across all considered datasets. A model based on mDeBERTaV3 also reported the best performance out of all models evaluated on `WikiGoldSK` and performance on par with SlovakBERT on the BSNLP2017 dataset. Finally, we also note that the choice of the training dataset has significant impact on the performance of the resulting NER model, as the difference between the F1 scores of the best performing model on the WikiANN dataset and the `WikiGoldSK` dataset is over 0.27. Despite the much larger size of the WikiANN dataset, the results in Table 3 suggest it is best not to combine it with the manually annotated dataset in order to obtain the best results.

When it comes to the few-shot learning experiments, the results can be seen in Table 4. We note that the combination of PVP 1 and PVP 2 (denoted "PVP 1 & 2" in Table 4) yields better results than when they are used separately. Comparing the results with those presented in Table 3, we can see that the supervised models outperform the few-shot learning approaches, even when trained on a silver-

	P	R	F1
10 shots			
PVP 1	0.4262	0.5290	0.4720
PVP 2	0.4320	0.6163	0.5079
PVP 1 & 2	0.4834	0.5937	0.5329
30 shots			
PVP 1	0.4853	0.5968	0.5353
PVP 2	0.4921	0.6502	0.5602
PVP 1 & 2	0.4857	0.6072	0.5397
50 shots			
PVP 1	0.5198	0.6176	0.5645
PVP 2	0.5041	**0.6688**	0.5749
PVP 1 & 2	**0.5321**	0.6484	**0.5845**

Table 4: The results of the PETER few-shot experiments for various shots and combinations of patter-verbalizer pairs (PVP) in terms of Precision (P), Recall (R) and F1 score. Best results are boldfaced.

standard dataset. This suggests that more work is necessary for few-shot NER approaches to be competitive with supervised approaches.

6 Conclusion

In this work, we introduce `WikiGoldSK`, the first sizable, manually annotated NER dataset in Slovak. We have established first baseline benchmarks on the dataset using state-of-the-art models, including multilingual and Slovak-specific models. The experiments with few-shot learning suggest that its performance does not reach that of supervised learning. The `WikiGoldSK` dataset is publicly released under permissible licensing terms, enabling training and evaluation of future models as well as tracking the progress in Slovak NER.

Limitations

While `WikiGoldSK` is currently the largest manually annotated Slovak NER dataset, it is still small in the great scheme of things, especially when its size (roughly 10 thousand labelled entities) gets compared to that of the CoNLL-2003 or Czech Named Entity Corpus 2.0 datasets (both with 35 thousand labelled entities). Our few-shot experiments have only been conducted in the case of Slovak and the newly introduced dataset, and may not generalize to other languages and datasets.

Ethics Statement

The dataset used for annotation was sampled from Slovak Wikipedia, which allows for reuse of its content under the terms of the Creative Commons Attribution-Share-Alike License 3.0. The annotated dataset is released under the same license.

Acknowledgements

References

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. Polyglot-ner: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 586–594. SIAM.

Dominic Balasuriya, Nicky Ringland, Joel Nothman, Tara Murphy, and James R. Curran. 2009. Named entity recognition in Wikipedia. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources (People's Web)*, pages 10–18, Suntec, Singapore. Association for Computational Linguistics.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

Archana Goyal, Vishal Gupta, and Manish Kumar. 2018. Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, 29:21–43.

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*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Ondrej Kaššák, Michal Kompan, and Mária Bieliková. 2012. Extrakcia pomenovaných entít pre slovenský jazyk. In *Krajči, S. Znalosti 2012: Proc. of the 11th Conference, Mikulov*, pages 52–66.

Valerio La Gatta, Vincenzo Moscato, Marco Postiglione, and Giancarlo Sperlì. 2021. Few-shot named entity recognition with cloze questions. *arXiv preprint arXiv:2111.12421*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Dávid Lupták. 2021. Rozpoznávanie pomenovaných entít metódami strojového učenia. Technická univerzita v Košiciach.

Jakub Maruniak. 2021. Anotácia a rozpoznávanie pomenovaných entít v slovenskom jazyku. Technická univerzita v Košiciach.

Jakub Mičo. 2019. Rozpoznávanie pomenovaných entít metódami strojového učenia. Slovenská technická univ. v Bratislave FIIT.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

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.

Matúš Pikuliak, Štefan Grivalský, Martin Konôpka, Miroslav Blšták, Martin Tamajka, Viktor Bachratý, Marián Šimko, Pavol Balážik, Michal Trnka, and Filip Uhlárik. 2021. Slovakbert: Slovak masked language model. *arXiv preprint arXiv:2109.15254*.

Jakub Piskorski, Bogdan Babych, Zara Kancheva, Olga Kanishcheva, Maria Lebedeva, Michał Marcinczuk, Preslav Nakov, Petya Osenova, Lidia Pivovarova, Senja Pollak, et al. 2021. Slav-ner: the 3rd cross-lingual challenge on recognition, normalization, classification, and linking of named entities across slavic languages.

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*. ACL.

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.

Dipti Misra Sharma Rajeev Sangal and Anil Kumar Singh, editors. 2008. *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*. Asian Federation of Natural Language Processing, Hyderabad, India.

Timo Schick and Hinrich Schütze. 2020a. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Timo Schick and Hinrich Schütze. 2020b. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

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.

Chen-Tse Tsai, Stephen Mayhew, and Dan Roth. 2016. Cross-lingual named entity recognition via wikification. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 219–228.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

A Annotation Manual

For the purpose of the `WikiGoldSK` dataset, we define the following classes of Named Entities:

- **PER** Names, surnames, nicknames of living beings, without titles. Groups of people that belong to a nation, city, family..., e.g. Slováci (*Slovaks* in English), Bratislavčania (meaning: people who live in the city of Bratislava), Kováčovci (meaning: family name). General adjectives are **not** entities e.g. rímsky vojak (*roman soldier* in English), slovenský jazyk (*slovak language* in English), but personal adjectives **are** PER entity, e.g. in "to je Petrov kufor" (*"it's Peter's suitcase"* in English), "Petrov" is a PER entity.

- **LOC** All territorial and and geo-political units, such as countries, cities, regions... Physical locations as rivers, parks, buildings, bridges, castles, roads... Streets were also classified as LOC entities, but without building numbers.

- **ORG** Political parties, companies, government institutions, political/sport/educational organizations, music bands. Museums, zoos and theaters were also annotated as ORG, although they are very close to LOC. However, in our opinion, their meaning exceed the location aspect. But if the context makes it clear that the described object is only a building and/or an area that belongs to an organisation, LOC should be used. Companies were also labeled with legal suffix, e.g. ESET, spol. s r.o. is all standing for one ORG entity.

- **MISC** Names of movies, awards, events, festivals, newspapers, TV or radio station names. Also sport series, cups and leagues were annotated as MISC.

In case of nested entities, the outer one is recognized as entity, e.g. whole "Národná Banka Slovenska" (*National Bank of Slovakia* in English) is ORG entity. Abbreviations following entity is separate entity, e.g. in "Úrad verejného zdravotníctva (UVZ)" (*Office of Public Health (OPH)* in English) we annotate 2 separate ORG entities.

The main differences between our guidelines and that of the BSNLP 2017 shared task (Piskorski et al., 2017) are as follows:

- For entities such as museums, theathers, zoos we've preferred ORG entity and only if it's clear from the context, LOC could be used. However, in BSNLP 2017 shared task these entities were always annotated as LOC.

- We've used MISC entity for newspapers, TV or radio stations. In BSNLP 2017 shared task guidelines it's not explicitly stated but in dataset these entities are mostly annotated as ORG.

B PETER training details

The unlabeled dataset is created by sampling 1000 sentences from the train split of `WikiGoldSK`. As a base model for training, we used SlovakBERT. To make the prediction comparable with that of the baselines, the token-level predictions were converted to the IOB2 form using a simple heuristic: whenever there is a sequence of entities of the same type, the tag of the very first entity is prefixed with B- while the rest is prefixed with I-. Note that this is a very imperfect heuristic, as it for instance cannot handle the cases where two entities from the same class are following straight after each other.