Named Entity Recognition for Low-Resource Languages - Profiting from Language Families

Sunna Torge[*†], **Andrei Politov**[*], **Christoph Lehmann**[*], **Bochra Saffar** and **Ziyan Tao**
TU Dresden, Center for Information Services and High Performance Computing (ZIH), Germany
Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) Dresden/Leipzig, Germany
{sunna.torge,andrei.politov,christoph.lehmann}@tu-dresden.de

Abstract

Machine learning drives forward the development in many areas of Natural Language Processing (NLP). Until now, many NLP systems and research are focusing on high-resource languages, i.e. languages for which many data resources exist. Recently, so-called low-resource languages increasingly come into focus. In this context, multi-lingual language models, which are trained on related languages to a target low-resource language, may enable NLP tasks on this low-resource language. In this work, we investigate the use of multi-lingual models for Named Entity Recognition (NER) for low-resource languages. We consider the West Slavic language family and the low-resource languages Upper Sorbian and Kashubian. Three RoBERTa models were trained from scratch, two mono-lingual models for Czech and Polish, and one bi-lingual model for Czech and Polish. These models were evaluated on the NER downstream task for Czech, Polish, Upper Sorbian, and Kashubian, and compared to existing state-of-the-art models such as RobeCzech, HerBERT, and XLM-R. The results indicate that the mono-lingual models perform better on the language they were trained on, and both the mono-lingual and language family models outperform the large multi-lingual model in downstream tasks. Overall, the study shows that low-resource West Slavic languages can benefit from closely related languages and their models.

1 Introduction

The success of recent large language models such as the GPTX-family (Brown et al., 2020) is due to a vast amount of training data and the availability of appropriate compute resources which allow to train these models. However, the availability of training data varies extremely between the languages of the world. High-resource languages such as English allow to train language models, performing impressively well on a variety of NLP tasks (Liu et al., 2019), whereas for the majority of the languages these large corpora are not available and thus, the same training concept does not necessarily yield well performing language models. This imbalance is addressed by multi-lingual models and transfer learning approaches.

Large multi-lingual language models such as XLM-R (Conneau et al., 2020) are trained on text data in 100 different languages and show good results on a variety of NLP downstream tasks in different languages, like e.g. Named Entity Recognition (NER) in German. However, there are many low-resource languages in the world, which are still not covered by these commonly available language models due to the lack of a reasonable amount of training data. This problem is addressed by different transfer learning approaches. One approach considers language families and the transfer based on the similarities of the languages of the same family (de Vries et al., 2021). In this case, the small amount of training data can partly be compensated by the similarities of the languages within the family. While training multi-lingual language models on languages from the same family, the training process profits from a larger amount of training data and from structural similarities of the languages at the same time.

In contrast to the training of multi-lingual language models, Ostendorff and Rehm (2023) consider the transfer from large language models for high-resource languages to large language models for lower-resource languages based on the overlapping vocabulary. In this approach, a large language model for a high-resource language (HRL) is used together with a small language model for a lower-resource language (LRL) in order to initiate the training of a large language model for LRL. This approach yields promising results and extensions to other language pairs which need to be investi-

[*] These authors contributed equally.
[†] Corresponding author.

gated. However, as this approach is based on an overlapping vocabulary, language families are of special interest.

In this paper, we present investigations on the West Slavic language family. The aim of this work was to assess the possibilities for low-resource languages like Upper Sorbian (Howson, 2017) and Kashubian (Nomachi, 2019) to profit from language models from the same language family. For this reason, we trained mono-lingual and multi-lingual language models from the same language familiy and evaluated them on the downstream task NER. Since there are several publicly available mono-lingual language models for slavic languages (Tikhonov et al., 2022), for comparison we evaluated some of them on the same downstream task.

Our contributions are as follows. We consider the languages Czech (cs), Polish (pl), Upper Sorbian (hsb), and Kashubian (csb), all being members of the West Slavic language family (Sussex and Cubberley, 2006). We trained three RoBERTa models (Liu et al., 2019) from scratch, two mono-lingual models for Czech and Polish respectively and one bi-lingual model for Czech and Polish, based on the Czech and Polish subset of the OSCAR data set (Abadji et al., 2022). For model evaluation, we used the downstream task, Named Entity Recognition (NER), as described in (Rahimi et al., 2019) and the corresponding wikiann dataset. We evaluated the three RoBERTa models on Czech and Polish NER and on Upper Sorbian and Kashubian NER. For comparison, we also considered existing SOTA mono- and multi-lingual models, namely the Czech RoBERTa model RobeCzech[1] (Straka et al., 2021), the Polish BERT model HerBERT[2] (Mroczkowski et al., 2021), and the multi-lingual RoBERTa model XLM-R[3] (Conneau et al., 2020), and evaluated them on Czech and Polish NER and on Upper Sorbian and Kashubian NER.

2 Related work

In de Vries et al. (2021), the impact of language families on low-resource languages was investigated. The authors used mono-lingual BERT models (source languages English, German, Dutch) and the multi-lingual mBERT to show, that linguistic structure can be transferred for the low-resource languages Gronings and West Frisian, which are closely related to the source languages.

A different approach is taken in Ogueji et al. (2021), where a transformer-based model is trained on 11 low-resource African languages belonging to a single language family. This expands the training data corpus by utilizing data within one language family. In contrast, we are interested in detecting those language combinations, which best support dedicated low-resource languages.

There is a variety of Czech and Polish language models available, as shown in Tikhonov et al. (2022). In Straka et al. (2021) RobeCzech, a Czech RoBERTa Model is presented and evaluated on several downstream tasks, including NER using two datasets (Ševčíková et al., 2007; Konkol and Konopík, 2013). A Polish RoBERTa model is described in (Dadas et al., 2020) and evaluated on NER, using the NKJP dataset (Przepiórkowski, 2011). In Mroczkowski et al. (2021) HerBert, a Polish BERT model is presented, trained on six different Polish datasets and evaluated on the NKJP dataset. For several reasons we decided to train models from scratch as baseline models. First, we wanted to compare mono-lingual and multi-lingual language models, which are trained on a subset of the languages of a language family, based on the same training corpora. Our particular focus was on the Sorbian language, which is spoken in a region of Germany adjacent to both Poland and the Czech Republic. As in practice geographic distances between countries, syntactic similarity and syntactic overlap play an important role for transfer learning (de Vries and Nissim, 2021), we wanted to train a czech-polish model. However, for comparison, we considered existing Czech and Polish language models in addition. Secondly, we were interested in performance analysis of distributed model training on our HPC infrastructure. These results are beyond the scope of this paper. Evaluating language models on NER is very common. Especially for balto-slavic languages there is a series of work, addressing the shared tasks of the Balto-Slavic NLP workshop series, e.g. (Suppa and Jariabka, 2021; Ljubešić and Lauc, 2021). In Piskorski et al. (2021) results of the last workshop are presented. As a starting point however, we restricted our investigations to NER for only three entites, namely Person, Organisation, and Location.

[1] https://huggingface.co/ufal/robeczech-base

[2] https://huggingface.co/allegro/herbert-base-cased

[3] https://huggingface.co/xlm-roberta-base

	N_D	low_LBP_D	RED_D	Meta_S
pl	443	209	607	10,121
cs	127	98	339	6,689

Table 1: Number of deleted documents and sentences (in thousands) after pre-processing

3 Training of Baseline Models

This investigation considers publicly available pre-trained language models such as RobeCzech, HerBERT and XLM-RoBERTa as well as models trained from scratch. In this section, the setup for training language models from scratch is described, which comprises training data, model architecture, tokenizer and the concrete training process.

3.1 Training Data

For the training of all models, the OSCAR (Open Super-large Crawled ALMAnaCH coRpus) dataset (version 22.01) (Ortiz Suárez et al., 2020; Abadji et al., 2022) was used. The Czech partition of the OSCAR dataset has a size of 58.6 GB, which comprises of 10,381,916 documents, and consists of 5,452,724,456 words. The Polish partition of the OSCAR dataset has a size of 139.0 GB, it comprises of 19,301,137 documents, and consists of 12,584,498,906 words. Before training a language model, we performed some preprocessing steps. Noisy documents, i.e. with high number of punctuation, were deleted. Documents were filtered, based on a low language-belonging probability (LBP) to the Czech and Polish languages respectively. The LBP is part of the meta data of the OSCAR dataset. We set the upper threshold for deletion to 0.6. A de-duplication step was performed in order to get rid of redundant documents. Sentences with less than 30 characters were deleted, as they have a high probability to be the meta data of web pages such as cookies, copy rights, urls etc. Table 1 depicts the deleted information, namely the number of noisy documents (N_D), documents with a low language-belonging probability (low_LBP_D), the redundant documents (RED_D), and the number of meta data sentences (Meta_S).

3.2 Model Architecture and Tokenizer

We used the RoBERTa architecture, a transformer-based architecture (Liu et al., 2019) with 125M parameters, 12 layers, 12 self-attention heads, and 768 hidden size for each of the models, we trained. As usual, models were trained on the masked language model objective. We trained three tokenizers, one each for Czech, Polish, and Czech and Polish, which are based on the Byte-Pair Encoding (BPE) tokenizer (Sennrich et al., 2016). The vocabulary size was set to 52K for each tokenizer, i.e. also for the multi-lingual tokenizer since the languages, both members of the West Slavic language family, are similar, especially in their lexical part. Given the same vocabulary size for each tokenizer, we also chose the same architecture for all models.

Overall, we trained two mono-lingual and one multi-lingual Roberta language models. The used models were trained using the official code released in the huggingface library[4], version 4.18.0. For training the multi-lingual model, the concatenation of the Czech and the Polish subset of the OSCAR dataset was used.

3.3 Training from Scratch of Mono- and Multi-Lingual Language Models

Within the concrete training process, all model weights were randomly initialized. The maximum sequence length was set to 512 tokens. All three models were trained with the same hyperparameters, which are presented in Table 2. We used the AdamW (Loshchilov and Hutter, 2017) optimizer for optimising the cross-entropy loss. The training

	Czech	Polish	Czech-Polish
Optimizer	AdamW		
Grad. acc.	10		
Warmup steps	55,700	75,000	117,000
Steps	1,160,200	1,563,900	2,434,100
Batch size	128		
Epochs	10		
Learning rate	4e-4		
Weight decay	0.01		
Adam β_1	0.9		
Adam β_2	0.98		

Table 2: Hyperparameter setting during training from scratch.

of models was done in a distributed manner on a node equipped with 2x AMD EPYC CPU 7352 (24 cores, multi-threading capable), 1 TB of RAM and 8x NVIDIA A100-SXM4 GPUs (40 GB HBM2 vRAM), in a fully connected intra-node topology (8x8 links, 3rd generation NVLink). We used PyTorch 1.11.0. While training, the data parallelism strategy of PyTorch DistributedDataParallel (DDP) was utilized. The training time for a mono-lingual model was approx. 48 hours. The training time for the multi-lingual model was approx. 100 hours.

[4] https://huggingface.co/roberta-base

During training of the three language models, the loss shows a strong decrease within the first 10% of the calculation steps and afterwards it decreases slowly. This structure remains the same for all three models. Figure 1 depicts the decrease of the cross-entropy loss during the training of the multi-lingual RoBERTa model. The loss is logged every 600 steps, and with gradient accumulation steps set to 10, this results in $\approx$ 400 data points. The warmup steps and the regularization term (weight decay) prevent the model from overfitting. The

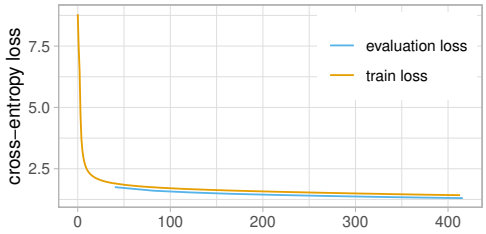

Figure 1: Cross-entropy loss on training (orange) and evaluation data (blue) during training of Czech-Polish LM

cross-entropy loss we obtained after training and evaluation, is shown in Table 3 for each of the trained language models. After this preparation

	Czech	Polish	Czech-Polish
Train loss	1.6	1.46	1.4
Evaluation loss	1.5	1.34	1.3

Table 3: Cross-Entropy Loss after training from scratch for each language model.

phase there are three models available that were trained from scratch.

4 Model Evaluation Results

Throughout the experiments, the main goal is to investigate whether the low-resource languages, Upper Sorbian and Kashubian, can benefit from language models that are trained for closely related higher-resource languages (here: Czech and Polish). Thereby, the experiment is twofold: First, the self-trained (from scratch) mono-lingual models and the multi-lingual language model are evaluated on the downstream task NER wrt the low-resource-languages of interest. Second, there is a further comparison with publicly available pre-trained mono- and multi-lingual language models such as e.g. RobeCzech or XLM-RoBERTa.

	cs	pl	hsb	csb
Train	20,000	20,000	150	150
Test	10,000	10,000	150	150
Size (MB)	9.860	9.764	0.073	0.088

Table 4: Number of sentences in training and test data for each language, Size of each data set (Bytes)

4.1 Evaluation Data

All of our evaluation experiments are based on the WikiANN dataset, which is a multi-lingual NER dataset consisting of Wikipedia articles annotated with LOC (location), PER (person), and ORG (organisation) tags in the IOB2 format. We used the subsets for Czech (cs), Polish (pl), Upper Sorbian (hsb), and Kashubian (csb) of the version (Rahimi et al., 2019). Table 4 depicts the number of sentences for each language and the size of each subset and clearly showing Upper Sorbian (hsb), and Kashubian (csb) as low-resource languages. Exemplary for the Czech language, in Table 5 the class distribution of the Czech wikiann subset is listed, which shows a sufficiently balanced dataset.

Class label	Number of sentences
Location	20,956
Organisation	17,938
Person	18,523

Table 5: Class distribution for the Czech wikiann subset.

4.2 Evaluation Setup on NER

In this section the evaluation setup of used models in connection with the wikiann data set is presented.

The following RoBERTa models which were trained from scratch are considered: 1. the mono-lingual models (Czech, Polish) and 2. the multi-lingual language model (Czech-Polish). Furthermore, three existing pretrained models are used, namely: 1. Czech RoBERTa (RobeCzech) (Straka et al., 2021), 2. Polish BERT model (HerBERT) (Mroczkowski et al., 2021), 3. the multi-lingual RoBERTa model (XLM-RoBERTa) (Conneau et al., 2020), for each using the official code released in the Huggingface library[5].

All models from above are evaluated on the downstream task NER based on the wikiann data set (see section 4.1) for the following languages:

[5] https://huggingface.co/ufal/robeczech-base, https://huggingface.co/allegro/herbert-base-cased, https://huggingface.co/xlm-roberta-base

Model	Evaluation NER (wikiann)			
Czech RoBERTa	cs	pl	hsb	csb
Polish RoBERTa	cs	pl	hsb	csb
Czech-Polish RoBERTa	cs	pl	hsb	csb
RobeCzech	cs	pl	hsb	csb
HerBERT	cs	pl	hsb	csb
XLM-R	cs	pl	hsb	csb

Table 6: Evaluation: Models and Languages

i) Czech (cs), ii) Polish (pl), iii) Upper Sorbian (hsb), and iv) Kashubian (csb). For the evaluation, which comprises of fine-tuning on training data and evaluation on the validation data, we used a stratified train - validation split; 80% for training and 20% for validation, keeping the same distribution of the entities in both splits. In the case of the low-resource languages hsb and csb, only 150 examples are available for fine-tuning. The hyperparameters for a full run of the fine-tuning process were chosen as follows: batch size 24, epochs 15. Based on different seeds a total of 20 runs was performed, whereby the integer seeds from $123, 124, \ldots, 142$ were used to control the data shuffling within the fine-tuning process. For each combination of language model and language data set, we chose the same 20 seeds in order to allow a reproducible comparison of the different models. An overview of all combinations within the evaluation is given in Table 6. Each of the trained language models is evaluated on NER for each of the languages under consideration.

4.3 Evaluation on Czech and Polish NER

We evaluated the language models on the downstream task NER on the languages cs, pl, hsb and csb as depicted in Table 6. In Figure 2 and Figure 3 we show the F1-score and accuracy, respectively, for all models we evaluated on the NER downstream task. The language, depicted in the header of each box plot is used for fine-tuning the corresponding model for the downstream task NER.

In this section we discuss our results concerning the languages cs and pl. First, we consider the models, trained from scratch, named Czech, Polish, and Czech-Polish. It can be seen that both of the mono-lingual models show a better accuracy on the language, they were trained on, in comparison with the Czech-Polish model. This is in line with the investigations on fine-tuning for NER on the majority of eight different languages (Rust et al., 2021). However, the decrease in performance is different in the two cases. This is possibly due to the training data size, since the Polish dataset (139.0 GB) is more than twice the size of the Czech (58.6 GB) (see section 3.1). Regarding the F1-score, in case of the Polish language, the Czech-Polish model performs slightly better than the Polish model.

We now compare our models with some existing models. In case of the Czech downstream task, it turns out that the Czech as well as the Czech-Polish models show a better F1-score than RobeCzech and XLM-R. Concerning the accuracy, RobeCzech performs comparable (slightly better) to our models, however the variance is more balanced. In case of the Polish downstream task, considering the F1-Score our Czech-Polish model performs slightly better than our Polish model and HerBERT. In contrast, concerning the accuracy our Polish model performs the best with a larger distance to our Czech-Polish model and HerBERT. The HerBERT model, we evaluated, was trained on a small, but high-quality data set. For the F1-score, the coverage is more important, which could explain this distance. For both downstream tasks, Czech and Polish, the mono-lingual model for the respective language and the language family model (Czech-Polish) perform better than the large multi-lingual model.

For a more detailed analysis, we consider the single entities. Exemplary, we compare the Czech model with the Czech-Polish model based on the Czech downstream task. The respective confusion matrices are shown in Table 7 and Table 8. The values in the confusion matrices are the mean values over 20 runs based on seeds of the corresponding combination of language model and language data set. Both matrices report quite similar results. In the referred tables, the discussed cells are highlighted. The Czech-Polish-LM identifies slightly more concrete entities for "I-ORG" and "B-LOC" (see main diagonal in confusion matrices Table 7 and 8, e.g. Czech-Polish-LM: approx. 85 entities classified as "I-ORG" vs. Czech-LM: approx. 80). The reverse holds for "O" entities. On the other hand, "B-ORG" and "B-LOC" as well as "I-ORG" and "I-LOC" are mixed up more often. Thereby, I-LOC is more often misclassified as I-ORG over the models than vice versa. The pairs of entities including "PER" are classified properly as shown in the confusion matrices.

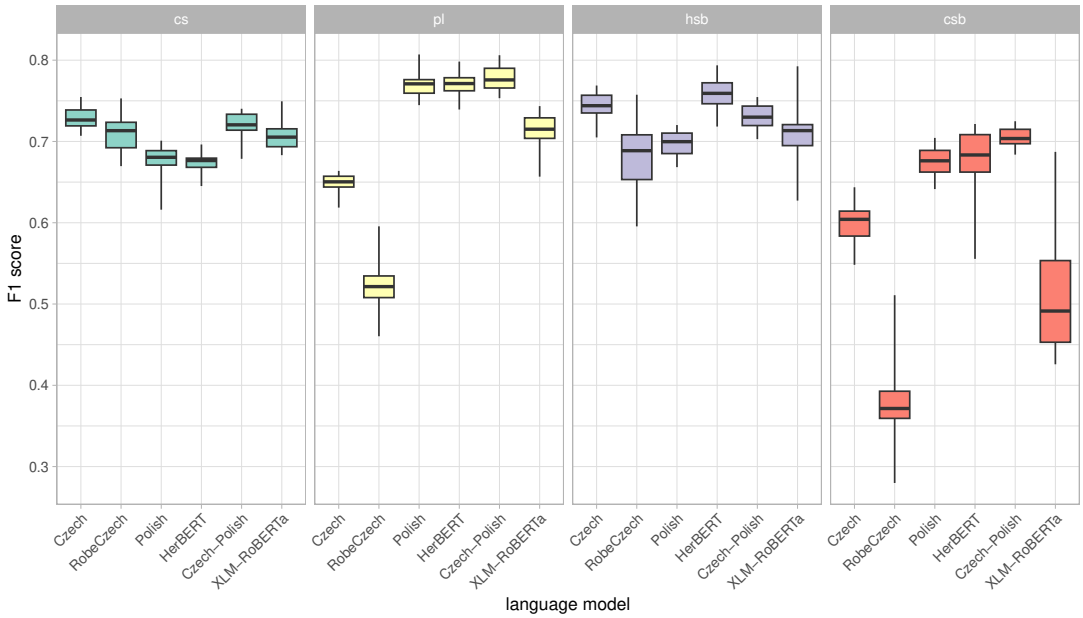

Figure 2: Boxplots of F1-score (20 runs) for all models and languages cs, pl, hsb, csb. The same set of seeds is used over all combinations.

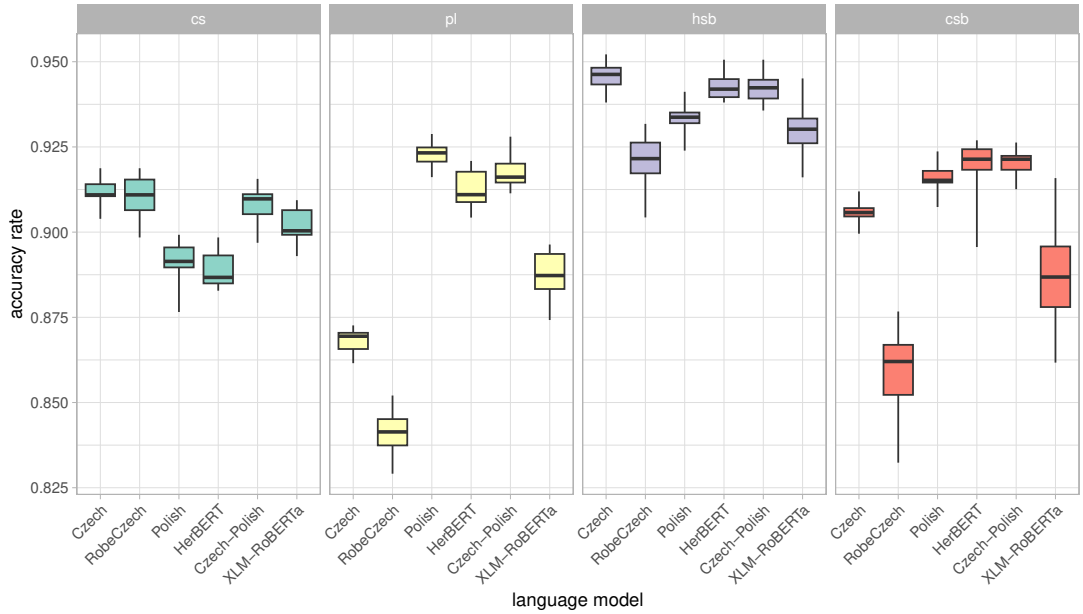

Figure 3: Boxplots of accuracy (20 runs) for all models and languages cs, pl, hsb, csb. The same set of seeds is used over all combinations.

	predicted label						
true label	O	B-PER	I-PER	B-ORG	I-ORG	B-LOC	I-LOC
O	788.40	1.20	2.25	2.40	3.60	4.50	5.65
B-PER	0.00	65.85	0.00	2.15	0.00	0.00	0.00
I-PER	6.95	0.00	82.60	0.10	4.30	0.95	1.10
B-ORG	2.10	3.35	1.05	45.90	0.00	9.60	0.00
I-ORG	7.65	2.05	6.85	3.20	80.10	2.80	6.35
B-LOC	2.80	1.15	0.00	7.55	2.00	58.50	1.00
I-LOC	1.00	0.00	1.75	0.00	15.25	0.90	45.10

Table 7: Mean values for confusion matrix (20 runs): Czech language model applied evaluated on cs data set. Highlighted cells refer to the discussion in the text.

	predicted label						
true label	O	B-PER	I-PER	B-ORG	I-ORG	B-LOC	I-LOC
O	775.85	1.30	2.30	4.90	7.65	9.30	6.70
B-PER	0.65	64.70	0.00	2.65	0.00	0.00	0.00
I-PER	6.50	0.80	82.95	0.00	2.85	2.00	0.90
B-ORG	2.35	2.10	1.00	46.95	0.10	9.50	0.00
I-ORG	3.90	1.45	5.95	3.45	84.70	1.45	8.10
B-LOC	1.50	0.80	0.00	7.95	1.65	61.05	0.05
I-LOC	1.95	0.00	0.70	0.00	14.55	0.30	46.50

Table 8: Mean values for confusion matrix (20 runs): Czech-Polish language model evaluated on cs data set. Highlighted cells refer to the discussion in the text.

4.4 Model Adaptation for Low-Resource Languages

The main goal of our work was to investigate, how language families may support low-resource languages within their family. For this purpose, we adapted the Czech, Polish, and Czech-Polish language models for the Upper Sorbian (hsb) and the Kashubian (csb) language. For each of the languages, the training data for fine-tuning for the NER downstream task comprises only 150 examples. The same holds for the evaluation data set.

In Figures 2 and 3, the F1-score and the accuracy is also presented for hsb and csb, comparing all considered models. For the downstream task NER in Upper Sorbian, the HerBERT model shows the best F1-score, which is surprising as the Upper Sorbian language is related more closely to Czech than to Polish (Howson, 2017). However, this might be caused by the high quality training data of the HerBERT model. Considering the accuracy, our Czech model performs the best, followed by our Czech-Polish model. The XLM-R model does perform worse than our Czech-Polish model, however the distance is not as large as in the case of the Polish language. The confusion matrices for our Czech model and the HerBERT model, evaluated on Upper Sorbian are shown in Table 9 and 10 resp. In general, the numbers are comparable, however, the HerBERT model does mix up less entities and identifies more "B-LOC" correctly, whereas our Czech model identifies more "I-ORG" entities.

For the downstream task NER in Kashubian, our Czech-Polish model shows the best F1-score, however the HerBERT model shows a similar accuracy, but a more balanced distribution. The interpretation of these results require a more thorough linguistic investigation, which is beyond the scope of this paper.

In Table 11, we summarize our results, presenting the mean F1-score and mean accuracy over 20 runs for all experiments.

We conclude, that language models, trained on languages within the same language family may improve downstream tasks for low-resource languages. This seems to be the case, if the language is not clearly related to a single language as in the case of Kashubian. However, mono-lingual models, trained on high-quality data may even outperform language family models, as it is the case with the Upper Sorbian language and the HerBERT model and our models, which were trained on a lower quality data set.

5 Conclusion and Future Work

In our paper, we investigated the West Slavic language family to evaluate the potential of language models for low-resource languages like Upper Sorbian and Kashubian. We trained three RoBERTa models from scratch, two mono-lingual models for both Czech and Polish respectively, and one multi-lingual model for Czech and Polish. These models were evaluated on the NER task for Czech, Polish, Upper Sorbian, and Kashubian. We also compared the performance of our models with existing SOTA mono- and multi-lingual models, namely RobeCzech, HerBERT, and XLM-R.

It can be seen that both mono-lingual models show better accuracy on the language they were trained on in comparison with the Czech-Polish model. The Czech and Czech-Polish models show a better F1-score than RobeCzech and XLM-R in the Czech downstream task. For both downstream tasks, the mono-lingual model for the respective language and the language family model (Czech-Polish) perform better than a large multi-lingual model. The adaptation of the language models for the Upper Sorbian and the Kashubian language was investigated. The HerBERT model shows the best F1-score for NER in Upper Sorbian. Our own

	predicted label						
true label	O	B-PER	I-PER	B-ORG	I-ORG	B-LOC	I-LOC
O	841.00	0.90	0.10	0.35	1.10	3.25	3.30
B-PER	0.30	46.15	0.05	0.60	0.00	1.45	0.45
I-PER	3.00	0.20	93.50	0.65	1.00	0.20	1.45
B-ORG	3.85	0.00	0.00	37.30	0.00	9.85	0.00
I-ORG	3.30	0.00	0.00	0.40	84.80	0.35	6.15
B-LOC	6.00	0.55	0.00	7.85	0.80	65.10	0.70
I-LOC	0.00	0.00	0.10	0.00	8.70	2.45	37.75

Table 9: Mean values confusion matrix (20 runs): Czech language model evaluated on hsb data set.

	predicted label						
true label	O	B-PER	I-PER	B-ORG	I-ORG	B-LOC	I-LOC
O	839.05	0.80	0.50	1.40	1.75	4.35	2.15
B-PER	0.60	45.55	0.25	1.20	0.00	1.35	0.05
I-PER	3.10	0.00	93.25	0.00	1.35	0.00	2.30
B-ORG	1.95	0.00	0.05	35.50	0.00	13.45	0.05
I-ORG	5.50	0.00	0.05	2.10	79.70	1.50	6.15
B-LOC	2.35	3.15	0.00	3.95	0.00	71.45	0.10
I-LOC	2.20	0.00	5.55	0.00	2.95	0.75	37.55

Table 10: Mean values confusion matrix (20 runs): HerBERT language model evaluated on hsb data set.

language_model	F1.cs	Acc.cs	F1.pl	Acc.pl	F1.hsb	Acc.hsb	F1.csb	Acc.csb
Czech	**0.729**	**0.911**	0.648	0.868	0.744	**0.946**	0.599	0.906
RobeCzech	0.710	**0.911**	0.521	0.841	0.679	0.921	0.377	0.860
Polish	0.676	0.892	0.769	**0.923**	0.697	0.933	0.677	0.916
HerBERT	0.674	0.888	0.771	0.912	**0.760**	0.943	0.676	0.920
Czech-Polish	0.720	0.908	**0.776**	0.918	0.730	0.942	**0.706**	**0.921**
XLM-RoBERTa	0.708	0.902	0.714	0.887	0.707	0.930	0.507	0.888

Table 11: Summary Results: Mean values of F1-score and accuracy over all 20 runs for all combinations of language model and language data set. Columnwise maximum values are bold.

Czech model performs the best for accuracy in Upper Sorbian. Our own Polish-Czech model shows the best F1-score for NER in Kashubian, while the HerBERT model shows similar accuracy.

Overall, the contribution has shown, that low-resource West Slavic languages such as Upper Sorbian or Kashubian can profit from closely related languages and their belonging models. But the crucial point seems to be the fundamental understanding of relatedness between low-resource languages and potentially promising high-resource languages. This requires a close collaboration with linguists, to successfully infer, where to profit from common training data and/or models. There is still a lot of potential to investigate more languages within a family and compare them with larger high-quality data sets (e.g. CNEC (Ševčíková et al., 2007), NKJP (Przepiórkowski, 2011)) and evaluate the models on modified NER tasks as described in Piskorski et al. (2021).

Furthermore, an interesting approach could be a cross-lingual and progressive transfer learning approach (Ostendorff and Rehm, 2023), where training of language models for low-resource languages starts with a large language model for a high-resource language and includes overlapping vocabulary. This method has yielded promising results for creating large models, but it refers to language families and not single languages.

Another development direction could be in building large corpora from existing parallel corpora. This would allow for the creation of high-quality training data for multi-lingual models and enable the training of models for low-resource languages that may not have sufficient training data available.

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF, SCADS22B) and the Saxon State Ministry for Science, Culture and Tourism (SMWK) by funding the competence center for Big Data and AI "ScaDS.AI Dresden/Leipzig“. The authors gratefully acknowledge the GWK support for funding this project by providing computing time through the Center for Information Services and HPC (ZIH) at TU

Dresden. The authors thank Alexander Fraser for helpful suggestions at the very beginning of this project, and the anonymous reviewers for supportive comments.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Sławomir Dadas, Michał Perełkiewicz, and Rafał Poundefinedwiata. 2020. Pre-training polish transformer-based language models at scale. In *Artificial Intelligence and Soft Computing: 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part II*, page 301–314, Berlin, Heidelberg. Springer-Verlag.

Wietse de Vries, Martijn Bartelds, Malvina Nissim, and Martijn Wieling. 2021. Adapting monolingual models: Data can be scarce when language similarity is high. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4901–4907, Online. Association for Computational Linguistics.

Wietse de Vries and Malvina Nissim. 2021. As good as new. how to successfully recycle English GPT-2 to make models for other languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 836–846, Online. Association for Computational Linguistics.

Phil Howson. 2017. Upper sorbian. *Journal of the International Phonetic Association*, 47(3):359–367.

Michal Konkol and Miloslav Konopík. 2013. Crf-based czech named entity recognizer and consolidation of czech ner research. In *Text, Speech, and Dialogue*, pages 153–160, Berlin, Heidelberg. Springer Berlin Heidelberg.

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 and Davor Lauc. 2021. BERTić - the transformer language model for Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42, Kiyv, Ukraine. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

Motoki Nomachi. 2019. *14. Placing Kashubian on the language map of Europe*, pages 453–490. De Gruyter Mouton, Berlin, Boston.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Malte Ostendorff and Georg Rehm. 2023. Efficient language model training through cross-lingual and progressive transfer learning.

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.

Adam Przepiórkowski. 2011. National corpus of polish. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

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.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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, Berlin, Heidelberg. Springer Berlin Heidelberg.

Milan Straka, Jakub Náplava, Jana Straková, and David Samuel. 2021. Robeczech: Czech roberta, a monolingual contextualized language representation model. In *Text, Speech, and Dialogue*, pages 197–209, Cham. Springer International Publishing.

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, Kiyv, Ukraine. Association for Computational Linguistics.

Roland Sussex and Paul Cubberley. 2006. *The Slavic Languages*. Cambridge Language Surveys. Cambridge University Press.

Alexey Tikhonov, Alex Malkhasov, Andrey Manoshin, George-Andrei Dima, Réka Cserháti, Md.Sadek Hossain Asif, and Matt Sárdi. 2022. EENLP: Cross-lingual Eastern European NLP index. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2050–2057, Marseille, France. European Language Resources Association.