Improving multilingual language models for the NER task

Anonymous ACL submission

Abstract

In this paper we test methods to improve the quality of cross-lingual transfer for under-resourced languages by means of more efficient mapping between embedding spaces, which helps to improve alignment between token embeddings. We test the method in the Named Entity Recognition task for a range of Slavic languages. The results of our experiments demonstrate improvement up to 8% of F1 measure in comparison to the XLM-RoBERTa few-shot baseline. Error analysis shows that our method especially helps for resolving ambiguous expressions, probably because the improved representation is able to take into account more information available in better resourced languages.

1 Introduction

Pre-trained Language Models (PLMs), such as BERT (Devlin et al., 2018), emerged as clear winners in a number of NLP tasks. Their multilingual versions also excel in few-shot and zero-shot learning, when a classifier initially trained in one language is applied to another one (Pires et al., 2019) with help of small training dataset or without any additional training data, since the models are capable of learning representations which are valid across languages (Conneau et al., 2020). However, the multilingual representations built for low-resourced languages often lead to a considerable performance drop (Blasi et al., 2022).

Our study is focused on improving alignment of PLM embeddings across Slavic languages with testing on the Named Entity Recognition (NER) task which is aimed at predicting types of Named Entities (NEs), specifically PERsons, ORGanisations or LOCations. Prediction of NEs across languages gives us a well-controlled semantic task, which matches forms with similar spelling to the same meaning: *Litwa, Litwy, Litwie, Litwę, Litwą* (all meaning 'Lithuania' in Polish) $\rightarrow$ (LOC). From a scientific point of view, the Slavic group is especially interesting because some languages have significantly more available data for training than others, while these languages exhibit rich morphology, which increases the number of forms per headword and leads to greater data sparsity in comparison to such languages as English or French. Below we refer to a donor language L_D as the language which has training data, while a fine-tuned classifier is applied in a few-shot way to a recipient language L_R. We show that the performance of zero-shot transfer in cross-lingual models deteriorates from L_D to lower-resourced L_R's, as there are no sufficient resources for building their representations and aligning them with better resourced languages. Moreover, we demonstrate that improving alignment between the embeddings by means of WECHSEL (Minixhofer et al., 2021) helps to resolve ambiguities and other problematic cases where the baseline cross-lingual model is wrong. More specifically our contributions concern:

1. better multilingual alignment between embeddings of multiple languages by transferring token embeddings of every recipient language to the embedding space of a donor language;
2. a new test corpus for seven Slavic languages and annotation guidelines explicitly covering difficult annotation cases;
3. error analysis of cross-lingual models in the NER.

2 Methodology

2.1 Datasets

For training and evaluation we used SlavicNER (Piskorski et al., 2019), which consist of Web pages on two news topics in six languages: Bulgarian, Czech, Polish, Russian, Slovene and

Entities	bg	cs	pl	ru	sl	uk	be
PER	321	377	414	373	360	298	479
LOC	974	530	716	552	690	948	821
ORG	87	195	195	292	104	138	125

Table 1: Number of entities in our gold test dataset.

Languages	bg	cs	pl	ru	sl	uk	be
SlavicNER	14957	11091	23902	16743	6472	8221	-
Our gold	3951	4457	4961	5123	4727	4875	4829
CC-100 $\times 10^6$	5487	2498	6490	23408	1669	7123	362

Table 2: Our train, gold test and pre-training datasets, in terms of the number of tokens.

Ukrainian. For evaluation of the classifiers on a wider range of topics we have also produced a new test dataset by manual cleaning of approximately 10,000 tokens taken from WikiNER (Pan et al., 2017) for the six languages of SlavicNER and also for Belarusian. The NE counts for this dataset are listed in Table 1. The other reason for creating a new gold dataset is because of numerous inconsistencies in the annotations of SlavicNER. While creating our gold dataset we have developed annotation guidelines to address three kinds of problems:

1. Entity boundaries;
2. Entity type disambiguation;
3. Entity presence.

One of the frequent problems with respect to the entity boundaries concerns nested entities, for example, for labelling Беларускі гуманітарны ліцэй імя Якуба Коласа 'Yakub Kolas Belarusian College for Humanities in Minsk' which contains a PER entity *Yakub Kolas* and a LOC entity *Minsk* in a bigger ORG entity. In our gold dataset we decided keep flat annotations, while there are inconsistencies in SlavicNER. Another important issue with the boundaries concerns ellipsis and conjunctions, e.g., Житомирская, Ровенская и Волынская области 'Zhytomyr, Rovenky and Volhynia regions' as the word 'regions' applies to each of the three entities, while the proper region should have been singular for each of them. We decided to include the common word in the last of the conjuncted entities. Other issues with respect to the entity boundaries concern abbreviated parts (Bulgarian г. София 'city Sofia'), titles (Polish *papież Franciszek* 'pope Francis'), and punctuation marks like quotes, which all need to be marked consistently as per the guidelines.

As for the disambiguation, the biggest issue concerns metonymy: in cases like *Italy defeated Spain in the World Cup* 'Italy' and 'Spain' obviously mean soccer teams (ORG), not countries (LOC). Finally, in some cases there is an open question of whether a recognizable entity is implied, as in *the Russian polar expedition of 1900–1902* (ORG?) or in predication *Mother called him* ***Peter***. Our annotation guidelines recommend no annotation in such cases, again with inconsistencies noted in existing resources, such as SlavicNER.

Both datasets for our study are compared in Table 2 with CC-100, the Common Crawl corpus subsets used for pre-training XLM-R, the size is given in terms of the millions of words. The Belarusian corpus is merely 1.5% of the Russian one.

2.2 NER setup

To test the contribution of our embeddings alignment mechanisms, we rely on a competitive NER approach, which is based on XLM-R (Conneau et al., 2019) and a linear token classification layer on top of the hidden states output from XLM-R. To evaluate models on our gold test dataset we used F1 metrics from seqeval framework, and for the SlavicNER dataset we used the evaluation script provided by the SlavicNER competition (Piskorski et al., 2019). A simple baseline is a few-shot transfer from SlavicNER with this model. Polish and Russian act as L_D because they have the largest amount of data, Table 2, while they differ in the character sets (Latin vs Cyrillic), which can impact XLM-R tokenization.

2.3 Embedding mapping methods

The main purpose of using embedding mapping in the proposed method is to help the language model accurately recognize named entities in several recipient languages using knowledge mostly from the donor language.

We implemented two models to make L_R embeddings closer to L_D:

M1 *Replacing embeddings for cognates* L_R embeddings are replaced with those of their L_D cognates. The motivation is to use known good cognates irrespectively of the cross-lingual alignment between L_R and L_D.

M2 *Iterative Wechsel* (Minixhofer et al., 2021) We initialize token embeddings from recipient language such that they are close to semantically similar tokens in donor language by utilizing multilingual static word embeddings covering both recipient and donor languages. The main idea is to replace recipient language embeddings with k nearest donor language tokens' embeddings. These k embeddings are then averaged with weights of softmaxed cosine similarity between recipient and current donor token.

In the M1 scenario, we used a Panslavonic dictionary of cognates (Sharoff, 2020) to replace the embeddings. We tested additional conditions for cognates such as frequency (occuring more than 2, 5, 8 times in training dataset), cosine similarity (score between token embedding and corresponding cognate embedding should be above 0.7, 0.8, 0.9), they all had a negative impact in comparison to replacing all full-word cognates. Usually, more than half of tokens in the train dataset can be replaced with cognates from our chosen dictionary, for instance, 54% of tokens in Ukrainian when L_D is Russian.

We refer to M2 as an *iterative* approach, because the mapping is gradually built in a loop for the seven Slavic languages in our study. Thus, we get an improved fully multilingual embedding space unlike the original WECHSEL. Unlike M1, the M2 approach does not use additional cognate dictionaries. The XLM-R model in all of our few-shot experiments was trained on the SlavicNER L_D dataset for three epochs and further fine-tuned on the respective SlavicNER L_R dataset for a single epoch.

3 Results and error analysis

Iterative WECHSEL is the best model overall, see Tables 3 & 4. Also, choosing the 5 closest words from L_D leads to the best result, probably, since at $k = 3$ there is shortage of relevant neighbors, and at $k = 7$ irrelevant tokens may get into the list of the nearest tokens. Furthermore, in this experiment we found that persons are recognized better than other entities, since the contexts of persons are probably less ambiguous.

The following examples in Slovenian illustrate cases when the baseline model is wrong and our multilingual WECHSEL model is right:

(1) Nekaj časa je potoval po Italiji in v Milanu obiskal **Cardana**.PER. (He traveled around Italy for a while and visited Cardano.PER in Milan.)

(2) Naslov prvaka je tridesetič osvojil EC KAC , ki je v finalni seriji s 4:0 v zmagah premagal **Vienna Capitals**. (The champion title was won for the thirtieth time by EC KAC, beating Vienna Capitals in victories in the final series 4-0.)

In the first example, the context for *Cardano* is similar to those of locations as it follows *visited* (the baseline model labeled it this way). The improved model is more likely to know about the person, *Gerolamo Cardano*. In the second example, *Vienna Capitals* is an organization, but the baseline labeled *Vienna* as a location. The iterative WECHSEL model made the correct predictions.

We also manually analyzed the errors of the baseline and M2 models over our gold WikiNER set for Russian to identify the frequent error types (Table 5 and Figure 1). The model copes quite poorly with entity types rare in the training dataset, e.g., treating football club names as LOC.

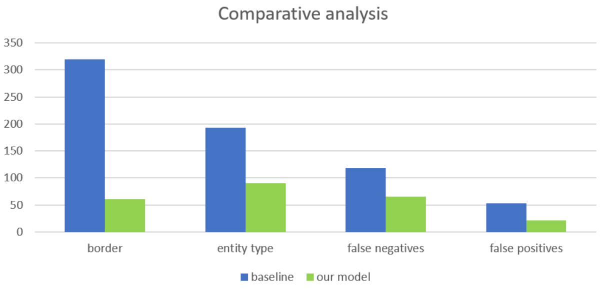

Figure 1: Error types for the baseline vs M2

The most frequent errors of the baseline model concern incorrect entity boundaries, which our modified WECHSEL model can improve, which is also true for the wrong entity type. We also discovered that many FN errors (i.e., an entity labelled as 'O') are caused by language-specific punctuation types, which impact the transfer, e.g., the guillemets are often used in Russian, but they are not used in other languages, hence the multilingual model does not consider them as features predicting NEs.

4 Related work

In addition to a number of studies which investigate the zero-shot abilities of PLMs on downstream tasks (Pires et al., 2019; Conneau et al., 2020), a recent study (Muller et al., 2021) analyzes which components of the architecture affect

Table 3: Polish as L_D. Few-shot testing on other 6 languages in our gold dataset (W) and SlavicNER (S).

	bg		**cs**		**ru**		**sl**		**uk**		**be**
	W	S	W	S	W	S	W	S	W	S	W
Baseline	0.77	0.82	0.84	0.87	0.76	0.82	0.80	0.83	0.78	0.81	0.77
M1	0.83	0.83	0.89	0.88	0.81	0.83	**0.85**	0.84	0.84	0.84	0.82
M2, k=3	0.84	0.83	0.89	0.89	0.83	0.84	**0.85**	0.85	0.84	0.85	0.83
M2, k=5	**0.85**	**0.84**	**0.90**	**0.91**	**0.84**	**0.85**	**0.85**	**0.86**	**0.86**	**0.86**	**0.84**
M2, k=7	0.84	0.82	0.89	0.88	0.81	0.84	0.84	0.84	0.84	0.83	0.82

Table 4: Russian as L_D. Few-shot testing on other 6 languages in our gold dataset (W) and SlavicNER (S).

	bg		**cs**		**pl**		**sl**		**uk**		**be**
	W	S	W	S	W	S	W	S	W	S	W
Baseline	0.79	0.80	0.78	0.84	0.79	0.85	0.77	0.82	0.83	0.79	0.81
M1	0.81	0.80	0.85	0.86	0.83	0.85	0.82	0.83	0.84	0.84	0.84
M2, k=3	**0.82**	0.80	0.85	0.87	**0.84**	**0.86**	**0.83**	0.83	0.85	0.85	0.84
M2, k=5	**0.82**	**0.81**	**0.86**	**0.88**	**0.84**	**0.86**	**0.83**	**0.84**	**0.86**	**0.87**	**0.85**
M2, k=7	0.81	0.80	0.85	0.85	**0.84**	0.84	**0.83**	0.82	0.84	0.83	0.84

Table 5: WECHSEL error types by category

Category	**Errors**
Sports organizations	36
Ambiguous cases	22
Embedded entity	17
Unusual name form	15
Context-caused	12
Acronym	3

the ability to transfer knowledge between the languages. Their study shows that in the lower layers there is a multilingual encoder necessary for transferring knowledge between languages and aligning internal representations, while in the upper layers there is a language-agnostic predictor that is focused on the target task. In our future studies, we want to analyze the effect of embedding replacements in the lower layers.

There has been also some research on which training elements are essential for multilinguality via re-training under different conditions (Dufter and Schütze, 2020), which show that the lexical overlap is less important than the structural similarity. This is also supported by our M1 experiment. (Zhao et al., 2020) showed that quality of the cross-lingual model, which was trained in a few-shot scenario, strongly depends on sampling of few-shot L_R data, such their diversity. In addition, if the target language is very different from the source language, then the adaptation step, i.e. fine-tuning, significantly improves the target metric on the downstream task. In this study We work with closely related languages, but we will experiment with curriculum learning methods to select more suitable few-shot data.

In article (Liu et al., 2021), the authors proposed a method for generating a massive dataset for the task of recognizing named entities in low-resource languages. The creation of a new dataset for the NER task is motivated by several problems. Firstly, the known datasets are significantly smaller than the data corpora on which the models from BERT family were initially trained. And secondly, in known datasets the set of possible entities is commonly limited to three or four, which limits the capabilities of the resulting NER model.

5 Conclusions

We demonstrated how to improve prediction quality in the NER task for lesser resourced languages by building a better multilingual representation space for several Slavic languages. We also created a multilingual test corpus with consistent annotations. Further work includes better understanding of the geometric properties of the original cross-lingual space and its improved version across better- and lesser-resourced languages and across representation layers, as well curriculum learning for better selection of few-shot data.

References

Blasi, D., Anastasopoulos, A., and Neubig, G. (2022). Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale.

Conneau, A., Wu, S., Li, H., Zettlemoyer, L., and Stoyanov, V. (2020). Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dufter, P. and Schütze, H. (2020). Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Liu, Z., Jiang, F., Hu, Y., Shi, C., and Fung, P. (2021). Ner-bert: A pre-trained model for low-resource entity tagging.

Minixhofer, B., Paischer, F., and Rekabsaz, N. (2021). Wechsel: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. *arXiv preprint arXiv:2112.06598*.

Muller, B., Elazar, Y., Sagot, B., and Seddah, D. (2021). First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.

Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., and Ji, H. (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.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual BERT? *arXiv preprint arXiv:1906.01502*.

Piskorski, J., Laskova, L., Marcińczuk, M., Pivovarova, L., Priban, P., Steinberger, J., and Yangarber, R. (2019). The second cross-lingual challenge on recognition, normalization, classification, and linking of named entities across slavic languages. pages 63–74.

Sharoff, S. (2020). Finding next of kin: Cross-lingual embedding spaces for related languages. *Natural Language Engineering*, 26(2):163–182.

Zhao, M., Zhu, Y., Shareghi, E., Vulić, I., Reichart, R., Korhonen, A., and Schütze, H. (2020). A closer look at few-shot crosslingual transfer: The choice of shots matters.