Exploring Contextual Embedding Spaces in Multilingual Models

Anonymous ACL submission

Abstract

Pre-trained multilingual language models such as BERT and XLM-RoBERTa are reasonably successful in zero-shot cross-lingual transfer because of the similarities in geometry of contextual embedding spaces for the donor and recipient languages. However, there has been little research on the relationship between the embeddings of individual tokens and the final predictions in downstream tasks. In this paper, we investigate the impact of (1) lexical similarity between the tokens, (2) differences in tokenization, and (3) similarity of embedding spaces. We test this on zero-shot cross-lingual transfer with Named Entity Recognition (NER) as the downstream task.

1 Introduction

Pre-trained Language Models (PLMs) such as BERT (Devlin et al., 2018) are widely used in all kinds of NLP tasks nowadays. By representing every subword in a language BERT creates the so-called contextual embedding space which can be visualized and further studied from the point of view of its geometric properties (Cai et al., 2021).

The multilingualism of modern PLMs, such as multilingual BERT or XLM-RoBERTa (Conneau et al., 2019), allows to perform zero-shot cross-lingual transfer (CLT), and recent research shows that when English is used as a donor language, the performance of the model on the recipient language data would not drop lower than 25%, and often it is merely 2-3% (Hu et al., 2020). This leads to a question on how the quality of multilingual embedding space affects the quality of CLT. A natural hypothesis would be that a) closely related languages, such as Catalan and Spanish, would have more similar embedding spaces and therefore a higher quality of CLT (bidirectionally) b) high-resourced languages, such as English or Russian, would have a fine-grained embedding space which again would allow higher quality of CLT.

In our experiments we found out that multilingual language models like XLM-RoBERTa have a bias in contextual word representations (CWRs) of ambiguous named entities (NEs) between low-resourced and high-resourced languages even after fine-tuning for the NER task. It causes CWRs of the languages that have more pre-training data to be placed nearer to each other than to other languages, even when the recipient languages are more closely related to to the donor. Also, CWRs of these NEs differ more by the language they came from than by the NE type they have. It is counter-intuitive with the distributional hypothesis and lowers the representativeness of the NE embeddings after fine-tuning. Also, we showed that isotropy of multilingual embedding space is affected differently by fine-tuning on different language groups. It means that the CWRs of Russian NEs are transformed in a similar way to Belarusian ones. In addition, we noticed a strong correlation between similarity of NE spelling between languages and the quality of zero-shot CLT between them. The more similar NEs are in terms of spelling, the better the CLT quality.

2 Related Work

Neural language models represent words and tokens as embedding vectors with a large number of dimensions (768 dimensions in BERT), which leads to many unexpected properties, such as a large number of nearest neighbors (Radovanović et al., 2010). PLMs further increase these problems by combining embeddings with parameters of the layers of attention transformers, thus leading to research in BERTology (Rogers et al., 2020), a study of how PLMs make their predictions. A case closely related to ours is a study by Cai et al. (2021), which explores the geometry of embedding spaces. While the parameters of the model are difficult to scrotinise, the contextual embeddings research can help in better understanding of the embedding topology across languages, so this may lead to improving the quality of zero-shot CLT.

Rajaee and Pilehvar (2021) studied the impact of fine-tuning on the isotropy of the contextual embedding space by considering the semantic text similarity (STS) as a downstream task. Authors showed that despite fine-tuning the embedding space stays highly anisotropic. Also, the local structure of CWRs undergoes a massive change during fine-tuning. In our work we are interested in the way fine-tuning on different languages impacts isotropy of monolingual embeddings in multilingual embedding space.

Their subsequent work (Rajaee and Pilehvar, 2022) analysed geometry of multilingual embedding space in terms of isotropy. Multilingual BERT (mBERT) has other distribution of dimensions than the English BERT but still is highly anisotropic. Also, in both models there is a frequency bias, which causes CWRs to form clusters according to the number of times they meet in a corpus. We investigated this bias between high-resourced and low-resourced languages for NEs before and after fine-tuning for the NER task.

However, not only the amount of pre-training data has a positive impact to the downstream task performance as shown by Rust et al. (2021). The languages adequately represented in the dictionary of a multilingual model have less performance gap with their monolingual counterparts. Below we report our experiments which show more specifically how differences in tokenization affect closely related languages in terms of their embedding space geometry even after fine-tuning.

Maronikolakis et al. (2021) investigated the importance of tokenization for multilingual models. Authors proposed a compatibility measure that correlates with downstream performance. In our work we extended this work and showed the impact of different tokenizations across languages on the topology of CWRs in parallel contexts.

3 Methodology

In this study we observe different geometrical properties and the impact of languages on multilingual embedding space after fine-tuning for NER as our downstream task.

3.1 Data and models

For our research we have expanded a synthetic NER dataset for 11 languages based on Slavic-NER (Lobov et al., 2022). The main idea behind its creation was to use machine-translated contexts taken from the English annotated WikiNER (Pan et al., 2017) and entities parsed from Wikipedia itself. The algorithm is to combine the corresponding entities and contexts; the contexts are chosen so that each sentence contains only one NE and the case of the NE would be the one desired (e.g., Nominative; the sentences which were translated with a different case in a language would be discarded as well as their counterparts in other languages). The original NE would be replaced with a placeholder, which can be filled with any other NE from the Wikipedia list. Thus, we can obtain a very large corpus of the size of the number of the contexts multiplied by the number of the entities.

In comparison with the original version we added languages, cleaned the contexts and added Accusatuve/Dative case contexts for LOCations. The languages present in the dataset are: Belarusian, Bulgarian, Catalan, Czech, English, Polish, Russian, Slovenian, Spanish, Turkish, Ukrainian.

Each context and each NE is strictly parallel (as machine translation and Wikipedia language links for parallel articles allow). The PER contexts take gender of the name into account: we distinguish male and female personal names. The PER and the ORG entities are only in Nominative case, while there is a certain amount of LOC entities (and corresponding contexts) in Accusative (Russian, Belarusian contexts of a type 'I am going to London'), Dative (the same type for Turkish) and Locative cases. The quality of machine translation for every language was manually assessed and the overall consistency of the synthetic data was selectively checked as well.

The size of the dataset is described in the Table 1.

Table 1: The sizes of SyntheticNER

Type	Quantity of Possible Sentences
PER	20,646,346
LOC	3,047,088
ORG	362,876

For all our experiments we used the XLM-RoBERTa model pretrained on 2.5TB of filtered CommonCrawl data.

The languages which interest us the most are Belarusian, English, Russian and Turkish. The reasons for that are as follows. The English and Russian languages are the best represented in the LM we use; Belarusian is closely related to Russian: it has the same word order (SVO) and it also uses Cyrillic alphabet, which is important for tokenisation, while Turkish, on the other side, is the most different from Belarusian: Turkish has the SOV word order and a high index of agglutination. In some of our experiments we also use the other languages in our dataset, e.g. Polish, as it is another Slavonic language, but it uses the Latin alphabet, while its NE spellings often differ from English.

In order to get the final dataset for NER task, we consider a subset of Cartesian product between the set of contexts and the set of entities. Formally, let C be a set of all context sentences with NE slots and E a set of all NEs available. Then, resulting dataset is

$$D \subset \{c(e),\ c \in C,\ e \in E\},$$

where $c(e)$ is a sentence which is produced by placing a NE e in a slot of a context sentence c. As NEs and contexts exist independently, we split both sets into train and test parts with 80% and 20% proportion respectively. Let's denote the train part of dataset as D_{train} and the test part as D_{test}. Then,

$$D_{train} = \{c(e),\ c \in C_{train},\ e \in E_{train}\} \subset D,$$
$$D_{test} = \{c(e),\ c \in C_{test},\ e \in E_{test}\} \subset D,$$

where $C_{train}, C_{test} \subset C$, $E_{train}, E_{test} \subset E$ and $C_{train} \sqcup C_{test} = C$, $E_{train} \sqcup E_{test} = E$, $|C_{train}| = 0.8 \cdot |C|$, $|E_{train}| = 0.8 \cdot |E|$.

3.2 Tokenization

PLMs use sub-word tokenizers which split a character sequence of the entire text into pieces called tokens and maps those tokens to natural numbers that represent the ordinal of tokens in a dictionary. One of the ways of splitting character sequences into tokens is byte-pair-encoding (BPE) (Sennrich et al., 2016; Gage, 1994). As BPE can split any word in a sequence into several tokens, in our experiments we consider embeddings of whole **words** defined as $e(w) = \frac{1}{k}\sum_{l=1}^{k} e(t_l)$, where w is a word, $t_1, t_2, \ldots, t_k$ its tokens and $e(t_1), e(t_2), \ldots, e(t_k)$ their contextual embeddings.

One of the problems of multilingual PLMs is underrepresentation of some languages in the pre-training dataset, which causes inadequate tokenization of some words (Maronikolakis et al., 2021). Also, there is an ambiguity problem as some NEs can be used either in PER contexts or in LOC contexts. This complicates the solution of NER task during CLT and may lead to inadequate distances between CWRs of such words in low-resourced and high-resourced languages.

An example of an ambiguous NE with considerable differences in tokenization across the four languages is *Washington*, which can be either PER or LOC, and it is rendered into Belarusian as Вашынгтон, Russian as Вашингтон, and Turkish as *Vaşington*. The tokenizer of pre-trained XLM-RoBERTa model uses a single token for English and Russian. However, for lesser-resourced languages it is split into tokens as:

be Ваш ын г тон
tr Va ş ington

We fine-tuned the XLM-RoBERTa model on the train part of the English NER corpus, generated 100 PER and 100 LOC samples for "Washington" in all of the languages using contexts from the test part, and collected CWRs of this NE from the output layer. In order to represent complexity and non-linearity of the multilingual embedding space we used t-SNE with perplexity=70 to display token embeddings in two dimensions (Figure 1).

We found that despite the similarity of Russian-Belarusian and English-Turkish CWRs in terms of cosine similarity of fine-tuned model for the NER task (Table 2), Russian and English as high-resourced languages are closer to each other than to low-resourced Belarusian and Turkish languages for this particular NE.

Also, we compared the quality of fine-tuning on different languages for the NER task. We fine-tuned XLM-RoBERTa model on the train parts of languages and tested it on the test parts of all other languages. While testing we measured the amount of wrong answers as the number of sentences where the model was wrong. Also, we measured the similarity between NEs of train languages and test languages by the transliterated normalized Levenshtein distance (TNLD). It's defined as a normalized Levenshtein distance between entities which are transliterated to the English language. Formally, let e_1 and e_2 be the

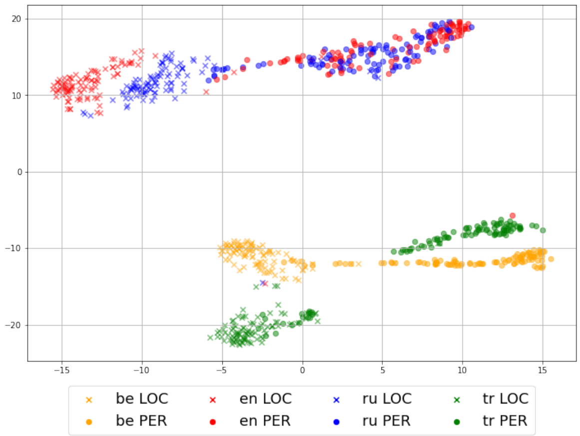

Figure 1: Output layer normalized embeddings of the "Washington" word transformed by t-SNE after fine-tuning on the English NER task

	be	en	ru	tr
be	1.0000	0.8457	**0.9159**	0.8329
en	0.8457	1.0000	0.8840	**0.9372**
ru	**0.9159**	0.8840	1.0000	0.8306
tr	0.8329	**0.9372**	0.8306	1.0000

Table 2: Average cosine similarities between parallel named entities from the output layer of fine-tuned model on the English NER task

entities from languages l_1 and l_2 respectfully and $t(e_i), i = 1, 2$ be their transliterations. Then TNLD is defined as

$$TNLD(e_1, e_2) = \frac{LD(t(e_1), t(e_2))}{\max(|t(e_1)|, |t(e_2)|)},$$

where LD is the Levenshtein distance. This metric allows to measure similarity between tokens even with different alphabets.

3.3 Embeddings

This set of experiments is dedicated to better understanding of the topology of NE embeddings in the multilingual embedding space of the XLM-RoBERTa model before and after fine-tuning on the NER task. Here we considered Belarusian, English, Russian, and Turkish languages with their training and testing parts of the SyntheticNER dataset. Before fine-tuning we projected contextual embeddings of entities from the test parts to the plane using t-SNE. After that we fine-tuned the model on the English train part for one epoch and did the same procedure with the resulting contextual embedding space (Figure 2). In this experiment we took 8,534 train sentences (6,082 PER, 1,580 LOC and 872 ORG) and 1,613 test sentences (1,000 PER, 395 LOC and 218 ORG).

In the initialization and output layers of the pre-trained model there are clear clusters divided by languages (Russian with Belarussian and English with Turkish), while after fine-tuning these clusters are less noticeable in the last layer. This explains the partial success of CLT. Also, in addition to language separation the embeddings from the output layer of the pre-trained model form some entity type clusters, especially persons and organizations. Obviously, in the fine-tuned model clusters based on the relation to a certain entity group prevail against the relation to the language this entity comes from, and this entity-language link is not entirely lost.

One of the features of the SyntheticNER dataset is a large number of sports organizations, which are named after their cities or districts. In this experiment we concluded that the embeddings from the output layer of a fine-tuned model for clubs named by their cities are placed in the LOC cluster by t-SNE ("Empoli", "Perugia", "Troyes"). Moreover, clubs with such names are near to the border between LOC and ORG clusters ("Swansea City", "Chicago Bulls"). It means that even after fine-tuning the multilingual models often fail to properly distinguish contexts during zero-shot transfer and rely mostly on the morphological properties of NEs.

4 Experiments

In the process of our research we conducted a set of experiments which can show the significance of NE similarity in zero-shot transfer for the NER task and different behaviour of the multilingual embedding space while training on the different language groups.

4.1 Fine-tuning impact of language groups

In this section we observe the impact of different languages to the isotropy change of the multilingual embedding space during fine-tuning. As the cosine similarity is a common measure of the isotropy, we observe a difference of average cosine similarities inside language samples between training steps. Formally, while training our model on a language l_{train} we define average language cosine similarity on the step t for language l_{test}, which can be equal to l_{train}, as $sim(l_{test}, t) = \mathbb{E}_{\phi, \psi} \cos(\phi, \psi)$, where ϕ, ψ are random word em-

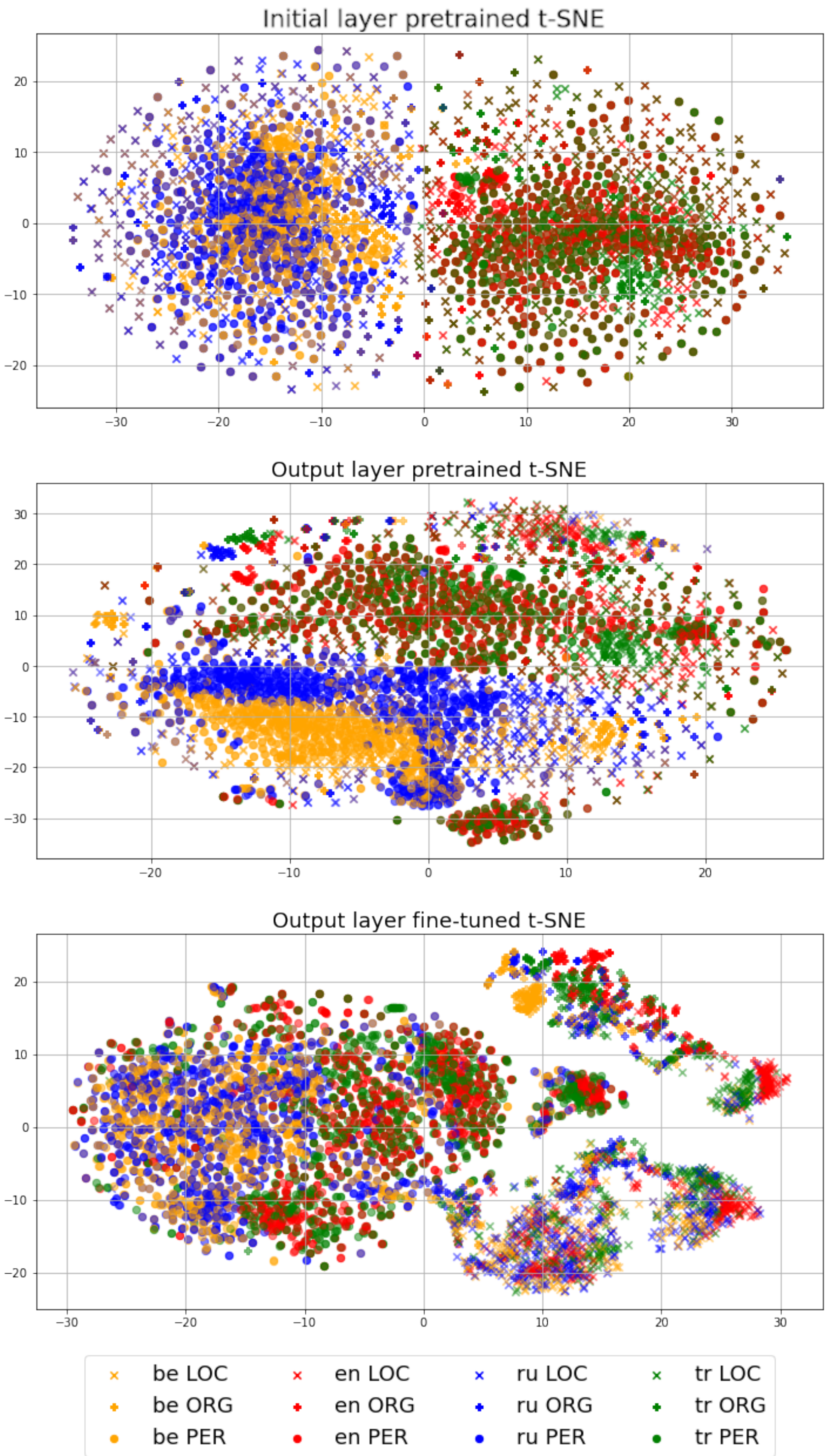

Figure 2: Embeddings of NE types in the initial and output layers before and after fine-tuning with t-SNE transformation

beddings for language l_{test}. After that we measure the difference $sim(l_{test}, t + h) - sim(l_{test}, t)$ for fixed value $h = 50$ during training (Figure 3).

Also, we consider correlations between these differences (Figure 4). According to plots, training on Turkish and Polish improves isotropy mostly for their embedding spaces but it is not so for other language embedding spaces. Training on the Russian part of dataset leads to the almost similar transformations for all languages in a sample as well as training on the Ukrainian part. It is also seen that six languages from this experiment are split into two groups according to the similarity of embedding space transformations during training. The Russian and Ukrainian languages have the greatest correlation coefficient while both of them have near zero or negative correlations with other languages. Another group is Polish, Turkish, Spanish and English languages. They also have high positive correlations which shows that their embeddings behave in a similar manner while fine-tuning.

4.2 NER task: pairwise comparison

The experiment with the "Washington" NE shows that there is a big impact of word tokenization to the NE embeddings topology. Even the same NEs from parallel sentences of closely related languages can be placed in different locations following their spelling and tokenization. In this section we would like to explore if there is a dependency between the spelling of NEs in different languages and the quality of zero-shot transfer between them.

Here we consider all available languages from the SyntheticNER corpus. For each language l_{train} we fine-tuned the XLM-RoBERTa model on the train part and measured the number of errors on the test parts of each language $l_{test} \neq l_{train}$. We also measured the average TNLD between parallel NEs in the test parts of l_{train} and l_{test} (Figure 5). This process allows to check the quality of zero-shot transfer from a single train language l_{train} to the languages l_{train} without revealing test contexts and NEs during fine-tuning.

We observe a high impact of parallel NE spelling to the quality of solving the NER task. If the two languages have NEs with a similar spelling, then the zero-shot transfer from one language to another will have a better quality than the transfer between languages with big differences in NE spelling.

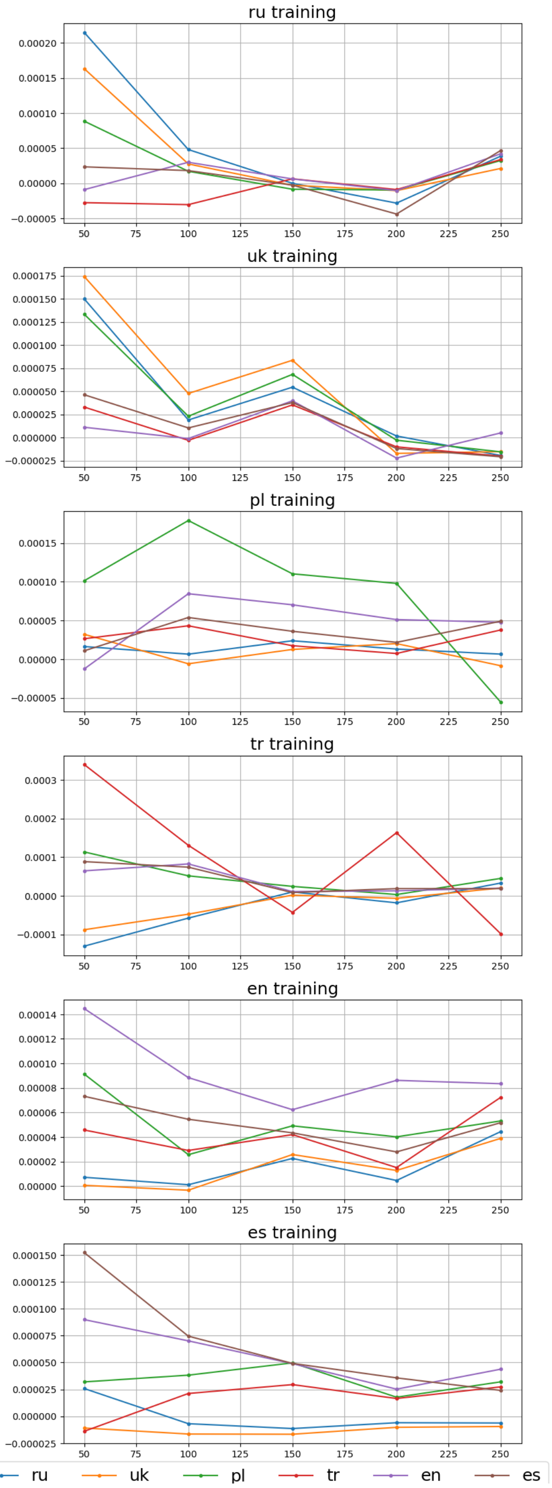

Figure 3: Differences of average cosine similarities inside languages between $h = 50$ training steps.

5 Conclusions

In our work we have demonstrated

1. the extent multilingual PLMs such as XLM-RoBERTa rely on the morphological information about words rather than on the con-

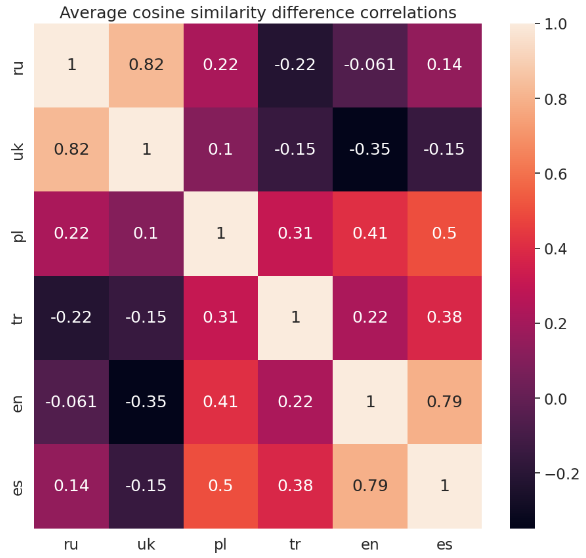

Figure 4: Correlations between average differences of cosine similarities during training. Languages appeared to form two clusters according to the similarity of transformations embeddings.

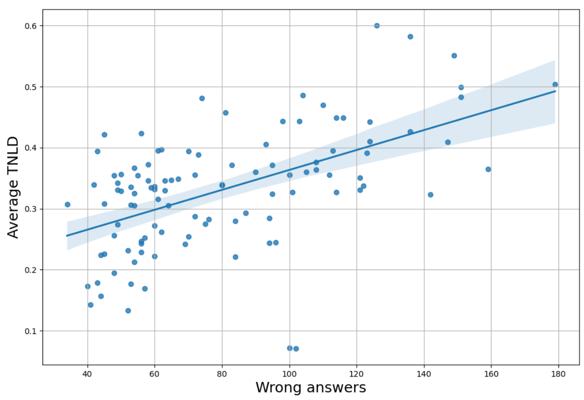

Figure 5: Dependence between number of wrong samples on the test dataset from the average TLND of parallel NEs

text information during zero-shot transfer for the NER task.

2. Multilingual model tokenization plays crucial role in the multilingual embedding space topology. Differences in tokenization and ambiguity of NEs cause the embeddings for closely related languages like Belarusian and Russian to be placed inside different manifolds.

3. The multilingual embedding space is affected in different ways while fine-tuning for the NER task according to the language group. Training affects closely-related languages in a similar way.

4. There is a correlation between model performance for the NER task and the named entities similarity expressed as TNLD. It also emphasizes the importance of tokenization in model's performance because similarity of tokens causes similarity of tokenization which positively affects quality in a downstream task like NER.

References

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*.

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. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv preprint arXiv:2003.11080*.

Valeriy Lobov, Alexandra Ivoylova, and Serge Sharoff. 2022. Applying natural annotation and curriculum learning to named entity recognition for underresourced languages. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4468–4480.

Antonis Maronikolakis, Philipp Dufter, and Hinrich Schütze. 2021. Wine is not v i n. on the compatibility of tokenizations across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2382–2399, Punta Cana, Dominican Republic.

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.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(Sep):2487–2531.

Sara Rajaee and Mohammad Taher Pilehvar. 2021. How does fine-tuning affect the geometry of embedding space: A case study on isotropy. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3042–3049, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sara Rajaee and Mohammad Taher Pilehvar. 2022. An isotropy analysis in the multilingual BERT embedding space. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1309–1316, Dublin, Ireland. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

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.