Comparing Multilanguage Models for Similar Language translation

Natural Language Processing with Deep Learning Project

Edyta Rozczypała
School of Applied Mathematics and Computer Science
University Josip Juraj Strossmayer of Osijek

February 14, 2026

Abstract

Multilingual language models allow for translations between languages that have a small amount of direct training data. However, for languages from the same language family, this type of model may prove to be less accurate than a model focusing only on translating between those languages. In this project I would like to focus on analyzing different multilingual models and their weak points, in particular looking into Croatian and Polish translations. The BLEU score will be used for evaluation, as well as back and-forward and cycle translations will be examined to compare the result with the original. After the analysis, I will also look into finding possible improvements for those translations.

1 Key Information

- Mentor: Dr. Domagoj Ševerdija

2 Introduction

Multilingual language models allow for translations between languages that have a small amount of direct training data. However, as mentioned in Wu and Dredze (2020) one inherent problem with multilingual LLMs is the unbalanced dataset used for training. The languages with large amounts of training data (for example English or Spanish) are usually well represented in the model, whereas languages with limited resources are constrained by the size of training data. Even high resource languages in those models don't compete with state of the art specialized models. As shown in Papadimitriou et al. (2023), models trained in only a few languages (bilingual/trilingual LLMs) or in a single language (monolingual LLMs) perform better in a number of NLP tasks than models trained in many languages (multilingual LLMs). This becomes especially clear for languages from the same language family, where interferences from other languages may negatively influence the accuracy of translations (for example translating through English may cause the translation to loose grammatical genders or noun cases).

The goal of this project is to compare the different language models on the quality of the Polish-Croatian translations. There is no specialized model, so a general Slavic translation model from the

Helsinki-NLP group will be used. This will be compared with a general multilingual model mBART (Tang et al., 2020). The models will be used to translate some sentences and then evaluated on their correctness. Additionally, the sentences will be translated back and then examined in order to find patterns in the types of errors that occur. The multilingual model will also be used to translate the text through a non-Slavic language in order to evaluate the quality of the translations. Finally, I will propose some improvements to the models.

3 Related Work

There is very little work dedicated to creating Polish-Croatian language models. However, there have been efforts into creating general Slavic language models, mainly by the Helsinki NLP group, which created the aforementioned general model, as well as models specialized in Southern, Eastern and Western Slavic language subgroups. Aside from the translation models, there are also specialized BERT models, mainly the Slavic BERT model (Arkhipov et al., 2019), which specifically includes Bulgarian, Czech, Polish and Russian. There are also some smaller specialized models - BERTić (Ljubešić and Lauc, 2021), which is a model that focuses on the Croatian, Bosnian, Serbian and Montenegrin languages, and CroSloEngual BERT (Ulčar and Robnik-Šikonja, 2020), which is trained on the Croatian, Slovenian, and English corpora. Additionally, Polish has a specialized BERT - HerBERT (Mroczkowski et al., 2021) and the Croatian model is being created (Gaurish Thakkar, 2024). In time, it should be possible to leverage those two models in order to create a translation focused model.

4 Approach

4.1 Model

There were two models used in this project. The first one was Helsinki's Slavic model. It is a transformer model, trained on the OPUS and Tatoeba data, which are one of the largest parallel corpora. The model has been pretrained using normalization and the SentencePiece tokenizer (Kudo and Richardson, 2018). It can take 16 Slavic languages as both source and target languages and accepts the Cyrillic alphabet.

The second model was Facebook's mBART-large-50 (Tang et al., 2020), which also is a transformer model. It has been pretrained using multilingual finetuning, which is a technique where instead of training a model from language i to language j, a model is trained to translate N languages to N other languages. The model can translate directly between any pair of 50 languages, of which 7 are from the Slavic language family.

4.2 Data

The dataset that I used was the MultiParaCrawl v7.1 (Bañón et al., 2020) dataset from OPUS, which can be found here. It contains parallel data from all European languages collected through web crawling. The languages used in the project are Croatian and Polish. The dataset contains 1,472,014 sentences, 21,185,000 Croatian tokens and 20,882,516 Polish tokens. Due to its size as well as computing constraints it was not used in its entirety.

4.3 Evaluation method

As mentioned in Kishore Papineni and Zhu (2002) the best way to automatically evaluate machine translation that correlates with human evaluation is the BLEU score. BLEU (bilingual evaluation understudy) is an algorithm for evaluating the quality of text which calculates the score for individual translated segments (in this case sentences) by comparing them with a set of good quality reference translations (here - the corresponding sentence from the dataset). Those scores are then averaged over the whole corpus to reach an estimate of the translation's overall quality. The implementation of BLEU used here is the implementation from the NLTK Python package. Additionally, some translated sentences were manually evaluated by the author of this project, in order to better understand the types of mistakes in translations.

5 Experiments

5.1 Experimental details

I ran 3 experiments with these two models. My first experiment was using the multilingual model and the Slavic model to translate back-and-forth between the languages and compare the translations with the original by using the BLEU score as well as manually looking at the translated texts. The second experiment was translating through English using the multilingual model and comparing the results to just using this model to translate between Polish and Croatian directly. My third experiment was to compare the translations made by the specialized Slavic model to the general multilingual model.

5.2 Results

In the below table 1 are BLEU scores for cycle translations for the multilingual model. Note that the reverse translations have a lower average BLEU score than the direct translations.

Hr-Pl	reverse Hr-Pl	Pl-Hr	reverse Pl-Hr
0.35	0.30	0.27	0.18

Table 1: BLEU scores for cycle translations with the multilingual model.

In the below table 2 are BLEU scores for cycle translations for the Slavic model. Note that the reverse translations have a higher average BLEU score than the direct translations, which is surprising considering the results of the multilingual model as well as what one may assume would happen.

Hr-Pl	reverse Hr-Pl	Pl-Hr	reverse Pl-Hr
0.56	0.68	0.58	0.67

Table 2: BLEU scores for cycle translations with the Slavic model.

In the below table 3 are BLEU scores for direct and indirect translations for the multilingual model. Note that the translations that go through English have a higher average BLEU score than the direct translations, which may be surprising, but mostly results from the low quality of Polish

and Croatian translations. If we compare these results to the results from the Slavic model in 2, than as expected the direct model does a better job of translating the texts.

Hr-Pl	Hr-En-Pl
0.35	0.54

Table 3: BLEU scores for direct and indirect translations with the multilingual model.

Finally, in the below table 4 are BLEU scores for translations for both models. As expected, the specialized model has significantly better results. What is interesting is that the results differ depending on in which direction the text is being translated. For the Slavic model the difference is marginal and may be accidental, but for the multilingual model this difference is bigger and it may be due to the amount of data from each of the language that the model was trained on.

Multilingual Hr-Pl	Slavic Hr-Pl	Multilingual Pl-Hr	Slavic Pl-Hr
0.35	0.56	0.27	0.58

Table 4: BLEU scores for translations with the multilingual and Slavic model.

6 Analysis

By analyzing the BLEU scores it can be concluded that the specialized model does a better overall job of translation. However, the difference between this model and using the multilingual model with the indirect translation method is not that big. This is probably due to the sizes of training data for those models, but especially for the specialized model not being big enough.

Looking at the specialized model translations there can be found two major types of errors. The first error has to do with the model choosing the wrong meaning of the word, for example *lekcje gotowania*, which means *cooking lessons* were translated by the model into *gotowanie wykładu*, which mean *cooking the lecture*. The other error was present in both models and had to do with incorrect word creation. The models were creating words according to word building rules and based on some other related words, but ended up creating non-existent words. For example the model created the word *wisienia*, which was supposed to mean *for hanging* as in something to hang clothes on. The correct translation of that is *wieszania*, but the model probably based its translation on the word *wisieć*, which means *to hang*.

The biggest problem with the multilingual model was that it often translated sentences into English when it wasn't supposed to. This happened when the sentence contained a word that existed in English, for example *cure*, even if it also existed in Croatian or Polish. This is the main reason why the BLEU score for this model was this low. The model also had occasional issues with picking the right pronouns, it translated *ask him* to *zapytaj mu* instead of the correct *zapytaj go*, as well as picking the right preposition, it translated *at the Bari Airport* to *w Bari Lotnisko* instead the correct *na Lotnisku Bari* (the word order and accusative case is also incorrect here).

Interestingly, in one of the examples the multilingual model translated the word *teacher* while correctly applying the gender marker based on the name even in the case where the original had a mistake and didn't do that.

7 Conclusion

In this work, I evaluated two models based on the quality of their Croatian-Polish translations. Predictably, the more specialized model produced better results. After manual evaluation it can be said that a lot of the translations were correct, even if they got lower BLEU scores, which can be attributed to the way that the BLEU score was constructed. The translations were mostly correct and the ones that had errors were mostly due to choosing the wrong translation of the word, that had a different meaning.

The multilingual model produced worse results, in some cases even translating into English instead of the target language when there was an English word in the sentence. Besides that the model had some problems with verb tenses and grammatical modes (like subjunctive or conditional). This model definitely has space for improvement, which can be done by finetuning it with a large corpora of parallel data in Polish and Croatian.

Interestingly, there were a few instances of the models creating a word according to word building rules, but producing a non-existent word, which resulted in an incorrect translation.

As mentioned before, only a small amount of available data was used. This was mainly due to the speed of the multilingual model, which was significantly slower than the Slavic model and was not able to process to many sentences at a time. It would be beneficial if in future work the evaluation could happen on a bigger corpora.

References

Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy. Association for Computational Linguistics.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Daša Farkaš Marko Tadić Gaurish Thakkar, Vanja Štefanec. 2024. Building a large language model for moderately resourced language: A case of croatian.

Todd Ward Kishore Papineni, Salim Roukos and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL).*

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing.

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.

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.

Isabel Papadimitriou, Kezia Lopez, and Dan Jurafsky. 2023. Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1194–1200, Dubrovnik, Croatia. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Matej Ulčar and Marko Robnik-Šikonja. 2020. Finest bert and crosloengual bert: less is more in multilingual models.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.