How Does a Multilingual LM Handle Multiple Languages?

Shashank Agarwal 2024AIY7543

November 26, 2024

1 Introduction

In recent years multilingual models have revolutionized natural language processing by enabling robust understanding and generation across multiple languages by a single model. BLOOM (BigScience Large Open-science Open-access Multilingual Language Model) is a state-of-the-art multilingual model trained on over 46 different sources of languages and multiple programming languages. These models provide opportunities to explore different cross linguistic relationships, transfer learning, and universal linguistic representations.

This mini-project investigates how multilingual language models process multiple languages at once and whether or not they exhibit cross-lingual transfer capabilities in doing so. Specifically, the project examines whether the embeddings of semantically similar words are shared across different languages so as to check if the 'word meaning' as a dimension spans the embedding space learnt by multilingual models. We will also check if finetuning the model using data of a particular language improves efficiency in case of other low-resource language. Such insights can provide a deeper understanding of the linguistic generalization capabilities of multilingual models and their application in cross-lingual tasks like translation and sentiment analysis.

2 Similarity between word embeddings in different languages

2.1 Objective:

The goal of this task is to determine whether the embeddings of semantically identical or similar words in different languages are close in the model's representation space. This involves calculating metrics such as cosine similarity to quantify the alignment of embeddings across languages.

2.2 Methodology:

2.2.1 Using the BLOOM-1.7B Model:

In this project we use the BLOOM-1.7B multilingual language model, a large-scale transformer trained on 46 natural and programming languages. The model was selected for its diverse linguistic coverage and due to its fewer number of parameters. For this task, we aim to analyze the embeddings generated by BLOOM-1.7B for words across different languages.

2.2.2 Language Selection:

BLOOM was trained on a diverse set of languages, making it ideal for this task.The selected languages for this analysis are Spanish, French, Portuguese, Hindi, and Chinese. These languages were chosen because they represent diverse linguistic families (e.g., Roman, Indo-Aryan, Sino-Tibetan) and scripts (e.g., Latin, Devanagari, Han). This diversity allows for an insightful evaluation of cross-lingual embedding similarity. capabilities

2.2.3 Word Sampling and Dataset Creation:

Medium-length words were randomly sampled from an English dictionary to ensure a focus on words that are less likely to include extremely common or obscure terms. After sampling, 1,500 words were

initially selected. These were carefully filtered to exclude proper nouns such as names, places, brand names, and other entities that might skew semantic evaluations. Each word was translated into the target languages (Spanish, French, Portuguese, Hindi, and Chinese) using the Google Translate Python library. The final dataset contains 1,200 entries each comprising the English word and its corresponding translations in the five selected languages.

2.2.4 Embedding Extraction:

All words in the dataset were tokenized using BLOOM's tokenizer. Multi-token outputs for single words were aggregated to create unified embeddings. Each tokenized word was passed through the BLOOM-1.7B model to extract its embeddings.

2.2.5 Similarity

For each word, cosine similarity was calculated between the embedding of the English word and the embeddings of its translations in the other languages.

2.3 RESULTS

2.4 similarity analysis

The results of the study revealed notable trends in the cross-lingual similarity of word embeddings generated by the BLOOM-1.7B model:

Figure 1: Cosine similarity heatmap of word embeddings across different languages. Higher similarity is observed between English and Romance languages.

The embeddings showed significantly high cosine similarity between English and Romance languages, particularly French, Spanish, and Portuguese. This observation aligns with the linguistic and scriptural closeness of these languages to English, as they share Latin roots and similar grammatical structures.

Conversely, embeddings for Hindi and Chinese words demonstrated low cosine similarity with their English counterparts. This can be attributed to the substantial differences in linguistic families, scripts (Devanagari and Han, respectively), and phonetic structures when compared to English.

2.4.1 Top 100 Words Similarity

The top 100 words with the highest similarity scores were selected for each language.

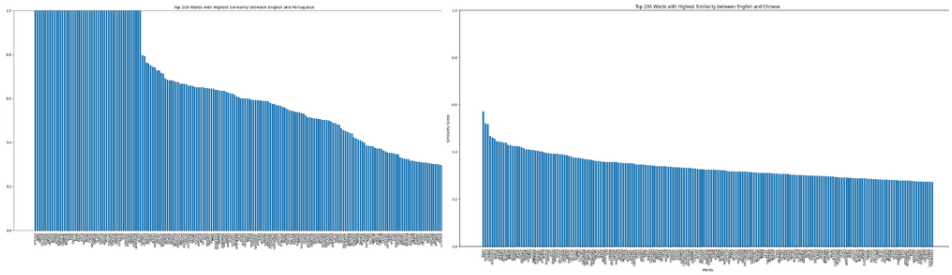

Figure 2: Top 100 words from both french and hindi that are similar to english.

3 Cross Lingual Transferability

3.1 Objective:

The objective of this task is to evaluate the cross-lingual transferability of a pre-trained multilingual model (BLOOM-1.7B) by fine-tuning it on a high-resource language (English) and assessing its performance on a low-resource language (Swahili). This aims to understand how well the model can transfer knowledge across languages without direct fine-tuning on the target language. The goal is to analyze the effectiveness of leveraging resources from a high-resource language to improve the model's understanding and performance in a low-resource language, thereby exploring the robustness of the multilingual model's generalization capabilities.

3.2 Methodology:

For this cross-lingual transferability task, we used the XNLI dataset, focusing on English as the high-resource language for training and Swahili as the low-resource language for evaluation. The BLOOM-1.7B multilingual model was fine-tuned for 12,000 iterations, with only the top 23rd layer updated while keeping the other layers frozen. This approach aimed to investigate how well the model could transfer knowledge from the high-resource English language to the low-resource Swahili language without additional fine-tuning on Swahili data. The evaluation was based on the accuracy of Swahili NLI tasks from the XNLI dataset.

3.3 Results:

The initial evaluation with the pretrained BLOOM-1.7B model on the XNLI dataset showed relatively low performance, achieving 30% accuracy for both English and Swahili. This outcome indicates that the model, while multilingual, struggled to effectively handle the cross-lingual transfer between the high-resource (English) and low-resource (Swahili) languages.

However, after fine-tuning the model on the top 23rd layer for 12,000 iterations, the performance significantly improved, with the model achieving 70% accuracy on English and 40% accuracy on Swahili. This result suggests that fine-tuning on the high-resource language (English) led to better transferability of knowledge to the low-resource language (Swahili), though the performance gap between the two languages remains significant. The improvement demonstrates the potential of selective fine-tuning for enhancing cross-lingual transferability.

Screenshot 2024-11-26 122552

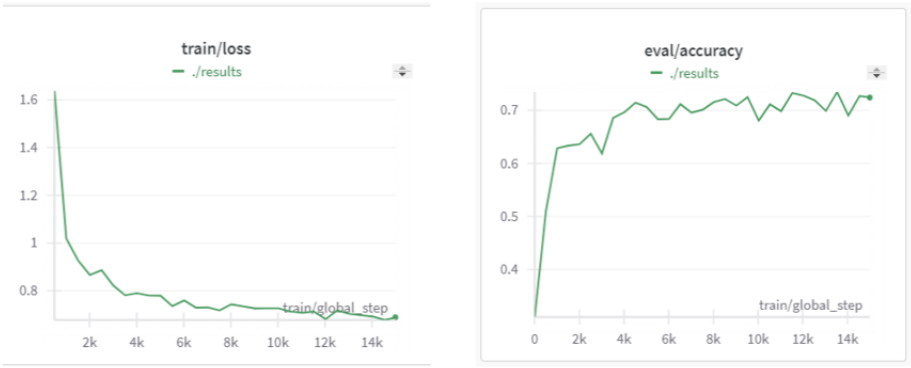

Figure 3: Metrics of finetuning.

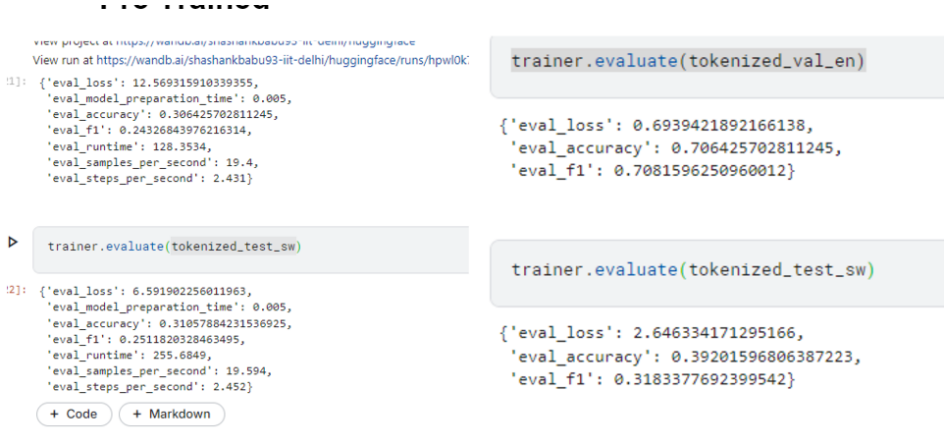

Figure 4: Results on both english and swahili dataset before fine-tuning and after fine-tuning

4 Conclusion

In this project, we explored the capabilities of large multilingual language models, particularly the BLOOM-1.7B model, to process and transfer knowledge across languages. Through tasks like analyzing word embedding similarities and evaluating cross-lingual transferability, we gained insights into how multilingual models handle different languages and their potential for transferring knowledge between high-resource and low-resource languages.

The results of the word embedding similarity analysis demonstrated that languages with closer linguistic and cultural roots (such as Spanish, French, and Portuguese) showed higher similarity in word representations, while more distant languages like Hindi and Chinese displayed weaker similarities. This highlights the model's strength in languages with shared linguistic characteristics, while also emphasizing the challenges it faces in handling languages that diverge significantly in structure and usage.

Furthermore, the cross-lingual transferability analysis showed that fine-tuning a model on a high-resource language (English) and testing it on a low-resource language (Swahili) led to substantial improvements, though performance discrepancies still existed. This reinforces the idea that multilingual models, while powerful, still struggle with language pairs that lack sufficient data or linguistic resources.

Overall, this work sheds light on the strengths and limitations of current multilingual models, suggesting that while they exhibit impressive cross-lingual capabilities, further advances are necessary to ensure better performance across a broader range of languages, especially those with limited resources.