How Does a Multilingual LM Handle Multiple Languages?

Anshul Patel
Yardi School of Artificial Intelligence
IIT Delhi
aib232072@iitd.ac.in

Bogam Sai Prabhath
Yardi School of Artificial Intelligence
IIT Delhi
aib232079@iitd.ac.in

Abstract

In this study, we explore the multilingual capabilities of the pre-trained BLOOM-1.7B model through a series of tasks: measuring the similarity between word embeddings across languages, probing the model for language understanding tasks, and assessing cross-lingual transferability. The findings indicate that BLOOM-1.7B effectively encodes semantic relationships across languages and demonstrates cross-lingual transferability, with performance varying based on language similarity and resource availability. Our analysis provides insights into the model's ability to handle multilingual tasks and its potential for zero-shot learning in low-resource languages. Code is available here.

1 Introduction

1.1 Background

As pretrained deep learning models, particularly those based on transformer architectures, continue to push the boundaries of performance in natural language processing (NLP), it becomes increasingly important to understand the types of linguistic knowledge they capture. One key area of interest is whether these models effectively encode information such as part-of-speech, morphology, and syntactic structures. Probing methods, which involve supervised models designed to detect specific linguistic information, are crucial for investigating the depth of these representations.

Recent research has focused on how these models, especially those using contextualized word embeddings, can represent complex syntactic information. Unlike traditional static embeddings that assign a fixed vector to each word, contextualized embeddings, generated through networks like LSTMs or Transformers, adapt to the specific context of each word in a sentence. This shift has improved model performance across a wide range of tasks, from parsing to coreference resolution, by capturing richer, more dynamic word representations.

Building on this background, our work investigates how the BLOOM-1.7B(Scao et al., 2022) multilingual model processes multiple languages. We explore its ability to transfer knowledge across languages and its capacity for semantic alignment in a shared embedding space. By evaluating its performance on tasks such as word similarity, probing for syntactic knowledge, and cross-lingual transferability, we aim to understand the model's effectiveness in multilingual contexts and its potential for zero-shot learning in low-resource languages.

1.2 Related work

Multilingual models have made significant strides, particularly with large-scale pre-trained architectures such as BERT(Devlin et al., 2019), XLM-R (Conneau et al., 2019), mBART(Liu, 2020), mBERT(Pires et al., 2019), and GPT-based (Achiam et al., 2023) models. These models are designed to process multiple languages by learning shared representations across diverse linguistic structures. mBERT demonstrated that a single model could perform well across various languages, especially when languages are closely related. XLM-R, an improved version of mBERT, has been shown to perform better on downstream tasks due to its larger multilingual corpus and more robust architecture.

In addition to multilingual representation learning, probing techniques have been crucial in understanding how well models capture linguistic features. For instance, probing tasks for BERT have shown that higher layers capture semantic information, while lower layers encode syntactic features. Studies on multilingual models like XLM-R and mBART have highlighted how these models handle cross-lingual similarities and are effective for tasks like translation and sentiment analysis.

Cross-lingual transferability, or the ability to transfer knowledge between languages, is vital for low-resource languages. Models like XLM-R have demonstrated strong cross-lingual transfer, particularly on tasks like XNLI. Research has shown that factors such as language similarity, shared vocabulary, and available data influence the success of transfer learning.

Alignment of multilingual embedding spaces is another key challenge. Models like MUSE have explored techniques for aligning embeddings across languages, improving tasks such as word similarity comparison and machine translation. BLOOM-1.7B benefits from shared subword tokenization and cross-lingual pretraining, enabling better alignment of its embedding spaces for multilingual tasks.

This project builds on these foundations, aiming to explore how BLOOM-1.7B processes multiple languages and transfers knowledge across linguistic boundaries.

2 Methodology

2.1 Task 1: Similarity between word embeddings in different languages

This task investigates whether BLOOM-1.7B (Scao et al., 2022) captures and aligns semantic representations of words across multiple languages within its embedding space. To achieve this, we created a parallel dataset of 992 words translated into 4 languages (French, Spanish, Chinese, Hindi). The English dataset was generated automatically by sampling words from the NLTM Brown corpus. Further, the translation was done into these 5 languages using the Google Translate library. For each word the embeddings were extracted for each language from the model, normalized and their semantic similarity was evaluated using cosine similarity. When a word is not present in BLOOM's vocabulary, the model performs subword tokenization, splitting the word into multiple subword tokens. In such cases, the word embedding is obtained by averaging the embeddings of all its subword tokens. The similarity matrix is visualized in Figure 3. The results provide insights into whether the "word meaning" is consistently represented across languages in BLOOM's embedding space.

2.2 Task 2: Probing to understand model behavior

In this task, we probed the multilingual capabilities of BLOOM-1.7B (Scao et al., 2022) by performing a textual entailment task using the XNLI (Conneau et al., 2018)(Cross-lingual Natural Language Inference) dataset. The XNLI dataset contains multilingual sentence pairs annotated with entailment labels, making it suitable for evaluating the model's understanding of sentence level semantic relationships across languages. The classification task was conducted on all languages in XNLI, allowing us to assess the model's ability to encode and generalize semantic information across diverse linguistic contexts.

Dataset we used the XNLI dataset which contains premise-hypothesis sentence pairs in 15 languages and their entailment labels. Input prompt for the BLOOM model was generates as *"Premise: sentence1 <sep> Hypothesis: sentence2"*, where languages for sentence1 and sentence2 were uniformly sampled from the set of 15 languages to capture the multilingual capability of the model. The training set consists of 10k datapoints and validation set contains 500 datapoints.

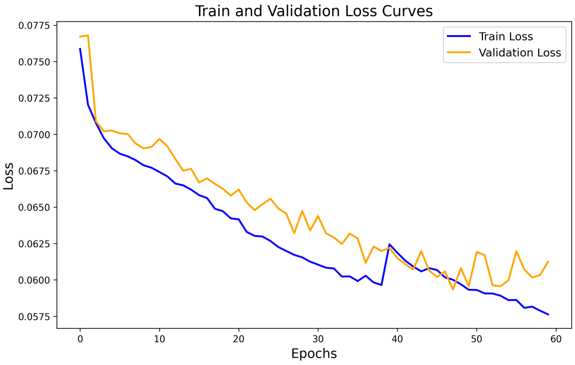

Figure 1: Task 2 training and validation loss

Architecture the probing classifier consists of a Transformer block with 8 attention heads, followed by a two-layer classification Multi-Layer Perceptron (MLP) that outputs the final probabilities. For each layer of the Transformer, the hidden representations were extracted. A special [CLS] token was appended to these hidden representations to serve as an aggregate feature representation for classification. These modified representations, including the [CLS] token, were then passed through the Transformer block. This process allowed the [CLS] token to gather and encode the necessary information from the input for the classification task.

Finally, the updated [CLS] token was passed through the MLP, which processed it to produce the output probabilities for the classification. This design enables the classifier to effectively utilize layer-specific information encoded within the Transformer model.

Training the complete Bloom1.7B model is frozen and we only train the classification head which is added on top of each of the transformer layers. The model is trained for 60 epochs end to end using batch size 16. This approach enabled us to identify which layers of BLOOM contribute most to multilingual understanding and semantic alignment. We also analyzed how these multilingual capabilities evolve during training. The results highlight the model's ability to generalize across languages and provide insights into the effectiveness of BLOOM's representations for multilingual NLU tasks. The train and validation loss are shown in Figure 1 and The accuracy of each layer over the validation set is shown in Table 1 .

2.3 Task 3: Cross-Lingual Transferability

In this task, we evaluated the cross-lingual transferability of BLOOM-1.7B using the XNLI dataset. The experiment involved fine-tuning the model on textual entailment task in a high-resource language (English) and testing its performance on the same task in a low-resource language (e.g., Swahili) without additional fine-tuning. This approach allowed us to assess the model's ability to generalize knowledge from high-resource to low-resource languages in a zero-shot setting. By analyzing the results, we investigated factors influencing cross-lingual transfer, such as language similarity, script differences, and resource availability. The findings shed light on the model's robustness and its capacity to perform well in under-represented languages, highlighting strengths and limitations in its multilingual representation learning.

Dataset the dataset for this task was generated using a process similar to that of Task 2, with one key distinction: the training dataset was created in English, while 13 separate test datasets were generated in the languages listed in Table 2.

Architecture the classification head used in this task is identical to that in Task 2. However, in this case, it was attached only to the final transformer layer, rather than being added after each transformer layer.

Training training is identical to Task 2.

3 Results

3.1 Task 1: Similarity between word embeddings in different languages

Across all cosine similarity matrices Figure 3, there is a noticeable high similarity along the diagonal. This indicates that words in their respective languages align closely to their semantic equivalents, reflecting the effectiveness of the cross-lingual representations in the BLOOM model.

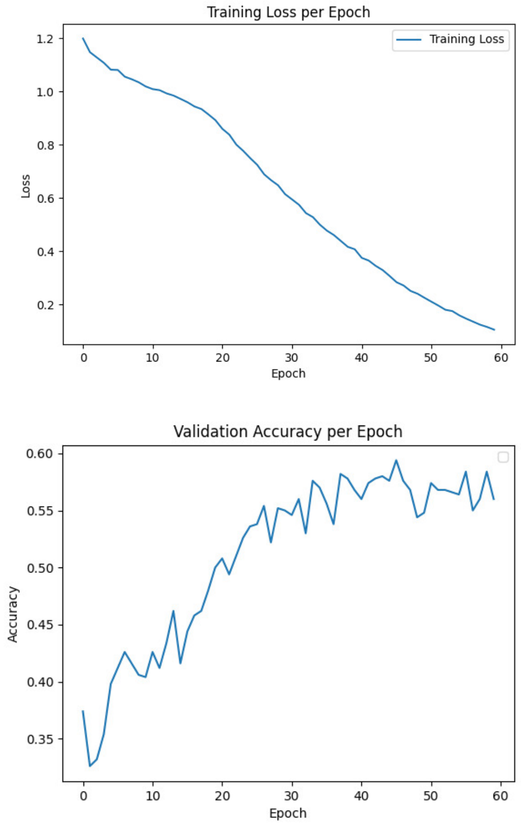

Figure 2: Task 3 training loss and validation accuracy

Dominance of English English constitutes the largest proportion of the training data, which may explain its robust alignment with other languages in the similarity matrices. The BLOOM model's embeddings for English are likely the most well-trained and accurate, leading to stronger cross-lingual alignments with other languages.

High Presence of Chinese Chinese is the second most represented language, which justifies its relatively better alignment compared to languages with lower representation. Despite its typological differences, the high volume of data likely enhances its embedding quality.

English and French The similarity matrix for English and French shows strong alignment for semantically similar words, given the high lexical borrowing and similarity in sentence structures between these languages.

Representation of French and Spanish These languages are reasonably well-represented, explaining their good alignment in the similarity matrices.

Their shared Latin roots further enhance the alignment between them.

Clustering by Language Families Languages belonging to similar language families (e.g., French and Spanish from the Romance group) show stronger embedding alignments. This trend is less pronounced between languages from unrelated families (e.g., Hindi and English).

Cross-Lingual Adaptation Challenges For typologically distant languages like Chinese and Hindi, the embedding similarities are lower, potentially highlighting challenges in aligning language representations with vastly different grammatical structures and vocabularies. The BLOOM model's ability to capture these relationships still outperforms random alignment, reflecting its capacity for cross-lingual understanding.

3.2 Task 2: Probing to understand model behavior

Layer-Specific Performance The results, as shown in Table 1, indicate that the middle layers (e.g., layer16, layer13, and layer14) exhibit the highest validation accuracy, with layer16 achieving the top performance (average accuracy: 57.78%) (Figure 4). This suggests that the middle layers of the BLOOM-1.7B model capture the most semantically rich and generalizable representations for the multilingual textual entailment task. Interestingly, the final layer (layer24) does not demonstrate the highest accuracy (47.29%), indicating that the deeper layers may be more specialized or task-specific and less effective for general multilingual understanding. By freezing the parameters of the BLOOM-1.7B model and training a lightweight classification head on top of each layer, the results provide an unbiased analysis of the semantic encoding capabilities across layers. This approach confirms that the hidden states from intermediate layers of transformer models often contain the most useful features for multilingual tasks.

Training Loss The training loss curve (Figure 1) exhibits a steady and consistent decrease over the 60 epochs, indicating that the model is effectively optimizing the classification head.

Validation Loss The validation loss (Figure 1) also shows a similar downward trend, consistently decreasing as training progresses. This indicates that the lightweight classification head generalizes well to unseen data without overfitting.

Layer Number	Average Accuracy (%)
layer_16	57.78
layer_13	57.54
layer_15	57.32
layer_14	57.22
layer_17	56.98
layer_12	56.70
layer_11	56.62
layer_18	56.42
layer_19	55.58
layer_10	55.52
layer_20	54.14
layer_21	53.74
layer_9	53.52
layer_22	52.32
layer_8	52.16
layer_7	51.82
layer_23	51.82
layer_6	49.48
layer_5	47.74
layer_4	46.88
layer_3	45.62
layer_2	44.44
layer_24	43.18
layer_1	42.50

Table 1: Layers sorted by validation accuracy (average over last 10 epochs) in decreasing order.

3.3 Task 3: Cross-Lingual Transferability

High-Accuracy Performance for High-Resource Languages As shown in Figure 5 and Table 2, English achieves the highest accuracy (59.4%), reflecting its dominant representation in the training data and the robust embeddings produced by the model for high-resource languages. Languages closely related to English, such as French (47.4%) and Spanish (42.0%), also exhibit high performance. Their shared linguistic features, such as similar syntax, vocabulary, and orthography, likely contribute to this alignment. Russian (41.0%) and Hindi (40.0%), while not as closely related to English as French or Spanish, still perform relatively well, possibly due to their moderate representation in the training data.

Moderate-Accuracy Performance for Mid-Resource Languages German (39.6%) and Bulgarian (38.4%) fall into a moderate-performance category, where the embeddings still capture semantic relationships but are less effective than those for

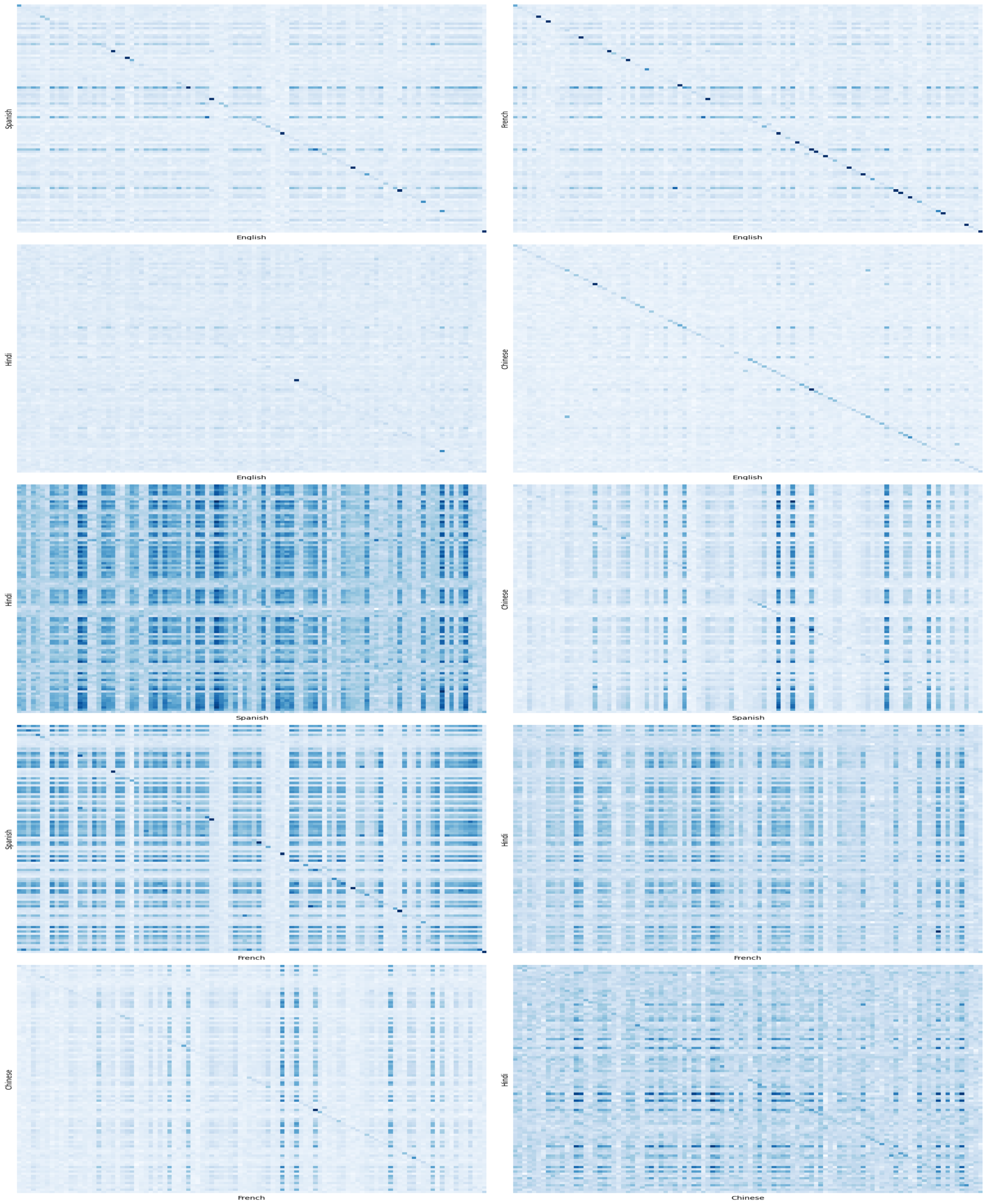

Figure 3: Cosine similarity matrix between two different languages.

high-resource languages. This performance could result from moderate representation in the training data and linguistic divergence from English.

Low-Accuracy Performance for Low-Resource Languages Low-resource languages, such as Swahili (0.312), Arabic (0.294), and Urdu (0.294), exhibit the lowest performance in the task, as seen in Figure 5 and Table 2. These languages are under-represented in BLOOM's training data, leading to weaker embeddings and limited cross-lingual transferability. Additionally, typological and script differences further exacerbate the performance gap. Languages like Swahili and Arabic, which use non-Latin scripts and have unique morphological structures, face additional challenges in embedding alignment with English-trained models.

Influence of Script and Language Similarity Languages with Latin-based scripts (e.g., French, Spanish, German) generally perform better, highlighting the importance of script similarity in cross-lingual transfer. Typological and orthographic dif-

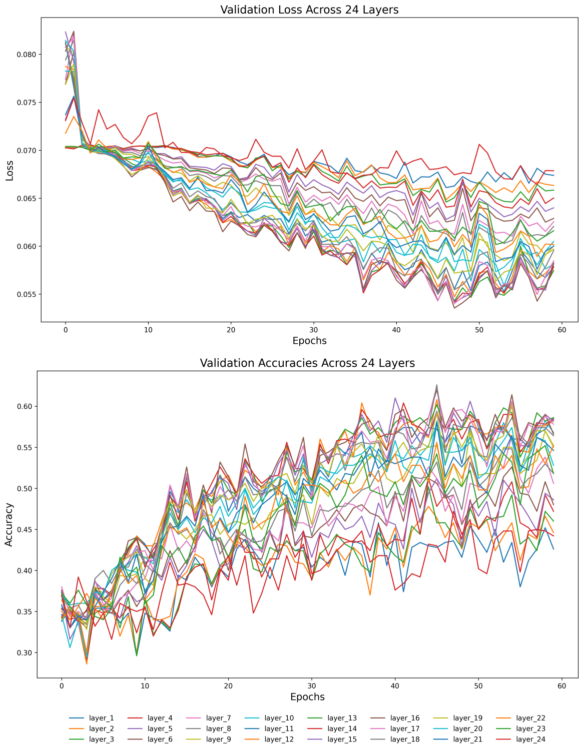

Figure 4: Task 2 loss and accuracies of individual layers.

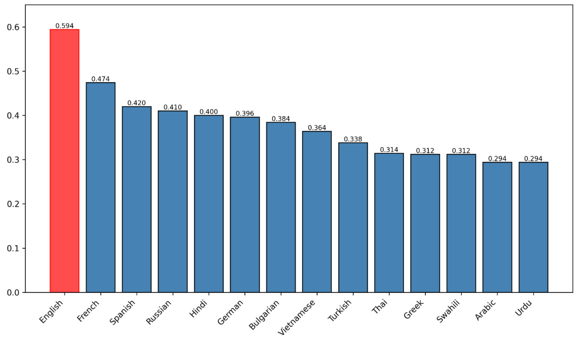

Figure 5: Zero shot accuracies

ferences negatively impact low-resource languages like Arabic and Swahili, underscoring the need for targeted fine-tuning or additional data for such languages.

4 Conclusion

Our study on the BLOOM-1.7B model reveals its robust multilingual capabilities. Through evaluating word embedding similarity, language understanding tasks, and cross-lingual transferability, we find that BLOOM-1.7B effectively captures semantic relationships across languages. Its performance, however, varies depending on language similarity and the availability of language resources. This analysis highlights the model's proficiency in multilingual tasks and underscores its potential for zero-shot learning, particularly in low-resource language contexts.

Language	Value
English	0.594
French	0.474
Spanish	0.420
Russian	0.410
Hindi	0.400
German	0.396
Bulgarian	0.384
Vietnamese	0.364
Turkish	0.338
Thai	0.314
Swahili	0.312
Greek	0.312
Urdu	0.294
Arabic	0.294

Table 2: Zero shot accuracies for different languages sorted in descending order by value.

5 Contribution

2023AIB2072: Task1, Task3, Report.
2023AIB2079: Task2, Report.

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, and many more.. 2023. Gpt-4 technical report.

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. In *Annual Meeting of the Association for Computational Linguistics*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Conference on Empirical Methods in Natural Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Y Liu. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *ArXiv*, abs/1906.01502.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ili'c, and many more.. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100.