Is it easy to be multilingual

Rishikesh Ksheersagar and **Karan Anand**
University of Michigan / Ann Arbor
rishiksh@umich.edu and karanand@umich.edu

1 Introduction

Multi-lingual language models like mBERT, XLM-R, mT5, and mBART have been instrumental in enabling natural language tasks in low-resource languages through cross-lingual transfer. This process involves pre-training and fine-tuning the models, resulting in top-tier performance on various tasks. In a typical scenario, a model is pre-trained on multiple languages and fine-tuned for a specific task in a source language, allowing it to perform this task effectively in another language with minimal or no human-labeled data (zero-shot or few-shot transfer).

Recent research has examined factors like model parameters, corpus size, and training steps, revealing that data transfer follows a power law. Probing studies suggest that large multi-lingual models possess zero-shot transfer capabilities for low-resource languages, emphasizing the importance of language structural similarity. This study extends these findings by providing an interpretable statistical framework for cross-lingual transfer, assessing the influence of various factors, and identifying the critical role of syntactic, morphological, and phonological similarities in predicting cross-lingual transfer performance.

2 Approach

We begin by setting up a framework to study cross-lingual transfer. We analyze a pre-trained M-BERT and M-BART model for pairs of languages (that are distributed into Source Languages and Target Languages) by observing its performance on standard Natural Language Processing tasks. The model is fine-tuned for the tasks using the Source Language and is evaluated on the same tasks in the Target Language. The tasks we will be utilizing are Named Entity Recognition (NER), Cross-Lingual Natural Language Inference (XNLI), and Question Answering (QA).

Along with the results derived from these tasks, we will also be modeling the similarity through lexical, morphological, phonological, and syntactic properties so that we can assess their impact on cross-lingual transfer.

The lexical similarity of languages is determined by computing a normalized Jenson-Shanon Divergence of the distribution of character n-grams of the source language against a similar distribution of the target language. The morphological similarity is determined by deriving a Type-Token-Ratio similarity of two languages. For the purpose of Phonological and Syntactic similarities we extract the corresponding features of the languages from the World Atlas of Language Structures database and compute the intersection of the properties over the union of the list of properties.

To determine the impact of all the derived predictors on cross-lingual transfer, we will be performing uni and bi-variate analyses on these predictors. We will fit a regression model to get a statistical understanding of the impact of the above mentioned predictors on cross-lingual transfer thus getting a framework to understand whether models trained on similar languages can be used to get zero-shot inferences on other languages.

3 Related Work

Recent research in the field of language models and cross-lingual transfer capabilities has yielded valuable insights. Probing studies, such as those conducted by [1] and [2], highlighted the remarkable zero-shot transfer capabilities of large multi-lingual language models, particularly their efficacy in low-resource languages. [3] emphasized the importance of "structural similarity" between source and target languages, transcending mere lexical overlap or word frequency. Furthermore, research by [4], [5], and [6] employed different methods to predict cross-lingual task performance. To extend these

findings, our work introduces an interpretable statistical framework to elucidate zero-shot and few-shot cross-lingual transfer across three tasks and 90 language pairs. Our research underscores the significance of language similarity in cross-lingual transfer while also highlighting the substantial role played by corpus size and language model performance in pre-trained models.

In our study, we aim to gain a deeper understanding of how multi-lingual pre-trained language models, such mBERT [10] and mBART [11], implicitly transfer linguistic and semantic knowledge across languages. Notably, these models are not explicitly provided with cross-lingual signals during pre-training. Instead, they learn cross-lingual connections through exposure to unannotated texts from various languages. This observation suggests that these models align their "semantic spaces" across different languages [7, 8].

The Cross-lingual TRansfer Evaluation of Multilingual Encoders (XTREME) [9] benchmark assesses the cross-lingual generalization capabilities of multilingual representations across 40 languages and 9 tasks. It reveals that while models tested on English excel in many tasks, there is room for improvement in cross-lingual transfer performance, especially in syntactic and sentence retrieval tasks.

4 Datasets

For the purposes of NER we will be using WiNER which is a Wikipedia annotated corpus, for XNLI we will be using the XNLI dataset from Huggingface, and for QA we will use the TyDiQA dataset.

The models we will be using for M-BERT and M-BART will be the bert-base-multilingual-cased and the MBart models from HuggingFace.

For our lexical, morphological, phonological, and syntactic observations we will be using the World Atlas of Language Structures (WALS) database.

5 Implementation Roadmap

Date	Task
12 Nov 2023	Fine-tune mBERT and mBART for selected NLP Tasks
19 Nov 2023	Derive the Predictors
26 Nov 2023	Analyze the Predictors and build a Regression Model to predict zero-shot cross-lingual transfer
3 Dec 2023	Collate Results and Conclusions

6 Division of work

The proposed division of work is as follows - Karan will work on fine-tuning the models for the purpose of NER and XNLI, find the lexical and morphological similarities of the chosen languages. Rishikesh will work on fine-tuning the models for the purpose of QA, find the phonological and syntactic similarities of languages, and create the regression model from the derived results as predictors.

References

[1] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics.

[2] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Sid- dhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934, 2020.

[3] Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual bert: An empirical study. In International Conference on Learning Representations, 2020.

[4] Anne Lauscher, Vinit Ravishankar, Ivan Vulic, and Goran Glavaš. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4483–4499, Online, November 2020. Association for Computational Linguistics.

[5] Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. Choosing transfer languages for cross-lingual learning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages

3125–3135, Florence, Italy, July 2019. Association for Computational Linguistics.

[6] Wietse de Vries, Martijn Wieling, and Malvina Nissim. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7676–7685, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[7] Jindˇrich Libovicky‘, Rudolf Rosa, and Alexander Fraser. How language-neutral is multilingual bert? arXiv preprint arXiv:1911.03310, 2019.

[8] Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 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, April 2021. Association for Computational Linguistics.

[9] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, Melvin Johnson. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization. arXiv preprint arXiv:2003.11080v5 [cs.CL] 4 Sep 2020.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding arXiv preprint arXiv:1810.04805 [cs.CL]

[11] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, Luke Zettlemoyer. Multilingual Denoising Pre-training for Neural Machine Translation. arXiv preprint arXiv:2001.08210 [cs.CL]