# Is it easy to be multilingual

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

## Abstract

Multi-lingual language models have redefined natural language processing in low-resource languages through cross-lingual transfer. By pre-training on multiple languages and fine-tuning for specific tasks, these models showcase potent zero-shot transfer capabilities, requiring minimal human-labeled data for proficient task performance. This study introduces an interpretable statistical framework, systematically evaluating the impact of model-related factors while highlighting the crucial role played by syntactic, morphological, lexical, and phonological similarities in predicting cross-lingual transfer performance. Offering nuanced insights into the determinants of successful cross-lingual transfer, this research provides valuable guidance for optimizing multi-lingual language models across diverse linguistic contexts, facilitating robust natural language processing in low-resource settings.

## 1 Introduction

In the dynamic landscape of natural language processing, the advent of multi-lingual language models, exemplified by mBERT, XLM-R, mT5, and mBART, has proven instrumental in transcending linguistic barriers and empowering applications in low-resource languages. These models, underpinned by the cross-lingual transfer paradigm, undergo a transformative process involving pre-training on diverse languages and subsequent fine-tuning for specific tasks, showcasing unparalleled performance across a spectrum of linguistic challenges. Recent investigations into the intricacies of cross-lingual transfer dynamics have unveiled compelling patterns, with probing studies emphasizing a power-law distribution in data transfer and the remarkable zero-shot transfer capabilities of large multi-lingual models for low-resource languages.

Building on these foundational insights, our research introduces a novel dimension by presenting an interpretable statistical framework for cross-lingual transfer assessment. Beyond exploring conventional factors like model parameters and training steps, our study takes a deeper dive into the nuanced influence of syntactic, morphological, and phonological similarities on cross-lingual transfer performance. By leveraging the rich insights provided by the Cross-lingual TRansfer Evaluation of Multilingual Encoders (XTREME) benchmark, we extend our exploration to three fundamental natural language processing tasks—Named Entity Recognition, Cross-Lingual Natural Language Inference, and Question Answering—across an extensive array of language pairs.

As we navigate the intricacies of multi-lingual pre-trained language models, our research endeavors to unravel the implicit transfer of linguistic and semantic knowledge across languages. The unique proposition of our study lies in its comprehensive approach, incorporating lexical, morphological, phonological, and syntactic properties to provide a holistic understanding of cross-lingual transfer dynamics. By examining both the impact of these linguistic properties and the performance of models on specific tasks, our goal is to contribute a robust framework for optimizing cross-lingual transfer capabilities, fostering a deeper comprehension of language diversity within the realm of natural language processing.

## 2 Related Work

Recent explorations in language models have unveiled remarkable insights into the domain of cross-lingual transfer capabilities. Groundbreaking studies by [1] and [2] have illuminated the extraordinary zero-shot transfer capabilities within extensive multilingual language models. These investigations have notably highlighted the efficacy of these models in handling low-resource languages, symbolizing a breakthrough in linguistic inclusivity. [3]

further emphasizes the pivotal role of "structural similarity" between source and target languages, surpassing mere lexical overlap or word frequency considerations.

Building upon these pivotal insights, the collective research led by [4], [5], and [6] has adopted diverse methodologies to predict cross-lingual task performance. This sets the stage for our innovative statistical framework, delving into zero-shot and few-shot cross-lingual transfer across three tasks and 90 language pairs. Our study accentuates the significance of language affinity in cross-lingual transfer while highlighting the substantial role played by corpus size and pre-trained model performance benchmarks.

Our investigation delves into the mechanisms through which multilingual pre-trained language models like mBERT [10] and mBART [11] implicitly propagate linguistic and semantic knowledge across languages. These models learn cross-lingual connections from unannotated texts across various languages, suggesting their ability to align "semantic spaces" across linguistic boundaries [7, 8].

The XTREME benchmark [9], assessing cross-lingual generalization in multilingual representations across 40 languages and nine tasks, reveals substantial insights. While excelling in English-centric tasks, models exhibit room for enhancement in syntactic intricacies and sentence retrieval tasks. This underscores the continuous quest for refining cross-lingual proficiency across diverse linguistic landscapes.

[10] explores cross-lingual transfer with an mT5 model, uncovering the influence of linguistic features such as syntax, morphology, and phonology on transfer, surpassing lexical similarity. This study also underscores language model performance as a reliable indicator of cross-lingual success, offering a practical assessment metric.

[14] analyzes cross-lingual transfer using an mT5 model, identifying predictive linguistic features such as syntax, morphology, and phonology. Notably, these aspects outperform lexical similarity in their influence on transfer. The study also highlights language model performance as an accessible metric for improving cross-lingual transfer processes.

Our study advances by introducing a novel framework designed to ascertain the most optimal source language for zero-shot cross-lingual transfer. This framework integrates mBERT's performance metrics across diverse languages and incorporates multiple language similarity metrics. By leveraging mBERT's multilingual capabilities and assessing linguistic similarity through various metrics, our framework seeks to identify the source language most conducive for zero-shot cross-lingual transfer to a target language without explicit training in that specific language. This strategic approach aims to optimize the precision and efficiency of zero-shot cross-lingual transfer by strategically selecting the source language based on mBERT's performance and linguistic similarity metrics.

## 3 Approach

In order to unravel the intricacies of cross-lingual transfer within the mBERT model, we establish an empirical framework inspired by transfer learning literature . Our objective is to discern how the effectiveness of cross-lingual transfer, denoted as ($\mathrm{S_T}$), for a given language pair (source language S and target language T) relates to specific characteristics of that language pair. Given the absence of a standard methodology for such studies and the lack of an evident theoretical model for transfer learning across languages, we draw inspiration from established methodologies to formulate our analysis.

Our empirical framework involves analyzing a pre-trained mBERT model for language pairs through observations of its performance ($\mathrm{S_T}$) in the target language on natural language processing (NLP) tasks (Named Entity Recognition, Question Answering, and Cross Lingual Natural Language Inference) after fine-tuning it using source language training data and evaluating the task on target language test data. This cross-lingual transfer can be represented as a function f as follows:

$$S_T = f(S_S, LS_{S,T}, LM) \tag{1}$$

where $\mathrm{S_S}$ is the performance of the model in the source language on the NLP tasks, LM is the performance of the model without any fine-tuning, and $\mathrm{LS_{S,T}}$ is a measurable language similarity metric that we introduce.

This formulation allows us to seek an optimal combination of linguistic and/or data-driven features that accurately estimate target language performance, providing valuable insights into factors influencing cross-lingual transfer within the mBERT model. To operationalize our analysis, we explore various possibilities for defining language similarity and modeling language performance.

### 3.1 Language Similarity

Assessing the similarity of languages($LS_{T,S}$) offers a nuanced perspective on cross-lingual transfer. In our framework, we consider multiple approaches to defining language similarity, each capturing different aspects of linguistic commonality.

**Lexical Similarity:** Lexical similarity refers to the degree of similarity between the vocabularies of two languages. It involves comparing the words and expressions used in one language with those in another to identify commonalities and differences.

We define lexical language similarity by computing the distribution of character n-grams for each source and target language. To capture dataset-specific similarities, we compute these distributions using the training dataset of each task and then measure the normalized Jensen-Shanon divergence (JSD), which is a symmetric and smoothed version of Kullback-Leibler divergence, of the source distribution against the target distribution.

**Morphological Similarity:** Morphological similarity refers to the degree of similarity between the morphological structures of two languages, which is a study of the internal structure of words and the rules governing how words are formed. Morphological features include prefixes, suffixes, root words, and grammatical markers that indicate aspects such as tense, number, gender, and case.

**Phonological Similarity:** Phonological similarity refers to the degree of similarity in the sounds and phonetic characteristics between two languages. Phonology is the branch of linguistics that deals with the systematic organization of sounds in languages, including the study of phonemes, syllables, intonation patterns, and other aspects of speech sounds.

**Syntactic Similarity:** Syntactic similarity refers to the degree of similarity in the structural organization and rules governing sentence construction between two languages. Syntax is the branch of linguistics that deals with the arrangement of words into phrases, clauses, and sentences, as well as the relationships between them.

Our methodology involved evaluating the degree of similarity in syntax and phonology across languages. To achieve this, we obtained syntactical and phonological vectors for all languages from the comprehensive World Atlas of Language Structures (WALS) database, utilizing the lang2vec tool.

For consistent vectors related to the Morphology metric, we sifted through the WALS dataset, extracting a total of 41 features. We included elements from both the Morphology and Nominal Categories sections, considering them indicative of morphological traits. These encompassed aspects like gender counts, usage patterns of definite and indefinite articles, and instances of reduplication.

Subsequently, we employed a calculation method based on determining the intersection over the union between the acquired vector representations of the source language and the target language. This approach allowed us to quantitatively measure the extent of resemblance in syntax, phonology, and morphology among languages, providing valuable insights into language similarity metrics rooted in linguistic elements. Through this comprehensive assessment, we aimed to discern and delineate the nuanced linguistic similarities among diverse languages, shedding light on their syntactic, phonological, and morphological proximities.

### 3.2 NLP tasks

The NLP tasks we attempted to create a framework for are Named Entity Recognition (NER), Question Answering (QA), and Cross Lingual Natural Language Inference) (XNLI).

**Named Entity Recognition:**NER is an NLP technique used to identify and classify named entities in text into predefined categories. Named entities are real-world objects that have a proper name, such as persons, organizations, locations, dates, numerical values, and more. The primary goal of Named Entity Recognition is to extract and classify these entities from unstructured text, providing a structured representation of information. This process is crucial for various NLP applications, including information retrieval and knowledge graph construction.

**Question Answering:** QA is an NLP task that involves developing systems capable of understanding and responding to questions posed in human language. The goal of QA systems is to provide relevant and accurate answers to user queries by extracting information from a given knowledge base or dataset. These systems find applications in various domains including Virtual Assistants, Search Engines, and Education.

**Cross Lingual Natural Language Inference:** XNLI is an NLP task that involves evaluating the ability of models to understand and infer relationships between sentences in different languages. The task extends the traditional natural language inference (NLI) to a cross-lingual setting, where the goal is to determine the logical relationship between a premise and a hypothesis in multiple languages. In the NLI task, models are trained and evaluated on pairs of sentences within the same language, typically involving tasks such as determining whether a hypothesis contradicts, entails, or is neutral with respect to a given premise. NLI models have applications in Information Retrieval systems, Text summarization, and Sentiment Analysis.

### 3.3 Datasets and Languages

Our research utilized different datasets for the three distinct tasks: Natural Language Inference (NLI) utilizing the XNLI dataset, Name-Entity Recognition (NER) employing the PANX dataset, and Question Answering (QA) utilizing the TyDiQA dataset. Each task's assessment criterion consists of F1 score for NER and QA, adhering to established and widely-accepted evaluation standards.

Throughout this project, our focus spans across 9 distinct languages chosen deliberately due to their prevalence within the datasets utilized for the Question Answering (QA) and Named Entity Recognition (NER) tasks. These languages were selected based on their common occurrence and significance within the TyDiQA and PANX datasets. The languages are : *Arabic, Bengali, English, Finnish, Indonesian, Korean, Russian, Swahili and Telugu.* This deliberate selection of nine languages serves as a crucial component of our investigation, facilitating a comprehensive analysis of cross-lingual transfer capabilities across diverse linguistic landscapes within the context of these specific tasks.

Acquiring datasets encompassing similar languages for distinct NLP tasks posed a considerable challenge. Specifically, sourcing appropriate datasets for the XNLI task proved notably problematic, precluding us from producing results for this particular task. To address this constraint in future endeavors, a proposed strategy involves generating comprehensive datasets for all languages by translating an existing dataset from one language into all necessary languages. This approach aims to ensure data consistency across languages. However, it introduces the challenge of securing an accurate translator capable of effectively translating datasets comprising numerous data points, potentially reaching hundreds of thousands in volume.

We utilized scripts from XTREME, a benchmark for the evaluation of the cross-lingual generalization ability of pre-trained multilingual models, to load the necessary data for each task from the available datasets. This data was then utilized by the scripts to fine-tune and test the pre-trained mBERT model.

### 3.4 Method

To comprehensively assess the influence of the obtained predictors on cross-lingual transfer, we gathered language similarity metrics for all 81 language pairs (9 * 9). Additionally, to acquire model performance metrics for each task, we trained nine individual models, one for each language, subsequently evaluating the performance of each model across all languages. This comprehensive process involved training nine distinct models and obtaining evaluations a total of 81 times, ensuring a thorough exploration of cross-lingual transfer capabilities.

Following this exhaustive data gathering and model training phase, we conducted bivariate analyses on these predictors, examining their impact on cross-lingual transfer. Subsequently, we employed regression modeling techniques to gain statistical insights into how the aforementioned predictors influence cross-lingual transfer. This analytical framework aims to decipher whether models trained on languages sharing similarities can effectively generate zero-shot inferences for other languages, providing a robust framework to comprehend and potentially leverage cross-lingual transfer phenomena.

## 4 Evaluation

Our experimental methodology initiates by applying the structured framework articulated in equation 1, incorporating the previously expounded features from this section. Our investigation commences with a rigorous bi-variate analysis, systematically examining the discrete impact of each feature on the efficacy of cross-lingual transfer. Subsequently, we present an exhaustive meta-regression, consolidating and synthesizing the collective influence of these features on the overarching proficiency of cross-lingual transfer mechanisms.

Our primary research focus centers on mBERT, an abbreviation denoting Multilingual BERT. This sophisticated language model is an extension of the BERT (Bidirectional Encoder Representations from Transformers) architecture, specifically trained on a diverse spectrum of languages. mBERT's training across varied linguistic contexts equips it with a comprehensive understanding of diverse language structures. Consequently, it can proficiently transfer this accrued knowledge to execute tasks across multiple languages without necessitating specific training for each individual language.

To incorporate all features outlined in equation 1, we conduct fine-tuning of the mBERT (multilingual-bert-base-uncased) model for the Question Answering (QA) task using the GoldP Task within the TyDiQA dataset. Additionally, for the Named Entity Recognition (NER) task, we finetune the model using the PANX dataset, ensuring a comprehensive incorporation of the outlined features in our analysis.

| Feature | QA | NER |
|---|---|---|
| Optimizer | AdamW | AdamW |
| Epochs | 3 | 5 |
| Learning Rate | 2e-5 | 1e-5 |
| Weight Decay | 0.01 | 0.01 |
| Batch Size | 32 | 32 |

### 4.1 Bi-variate Analysis

We start our study by conducting a bi-variate correlation analysis using the predictors introduced in section 3 and report the correlation of each predictor with cross-lingual transfer performance for each task below. By employing all features mentioned in equation 1, we calculate the Correlation coefficient between the Model Performance on the Target Language and obtained features, expecting those with high correlation to act as strong predictors for zero-shot multilingual cross transfer.

| Task | $S_S$ | Syn | Phon | Morph | Lex | LM |
|---|---|---|---|---|---|---|
| QA | 0.32 | 0.60 | 0.46 | 0.56 | -0.7 | 0.23 |
| NER | 0.41 | 0.62 | -0.17 | 0.4 | -0.83 | 0.45 |

The correlations between the model performance on target languages and various linguistic features across the two tasks, Question Answering (QA) and Named Entity Recognition (NER), reveal noteworthy patterns. In the QA task, we observe moderate to strong positive correlations between model performance and features such as syntax (0.60), phonology (0.46), and morphology (0.56), suggesting a significant relationship between these linguistic aspects and the model's effectiveness in cross-lingual transfer. Conversely, the lexical similarity (Lex) exhibits a notably strong negative correlation (-0.7), implying an inverse relationship, where lower lexical similarity potentially aids in better cross-lingual transfer performance. Interestingly, the language model performance (LM) also showcases a positive albeit weaker correlation (0.23), indicating a potential but less influential connection between the overall model performance and cross-lingual transfer in the QA task.

In contrast, the correlations in the NER task portray a similar trend in some linguistic aspects while showcasing distinct associations in others. Notably, syntax (0.62) and morphology (0.4) maintain positive correlations with model performance, emphasizing their potential importance in cross-lingual transfer for NER. However, phonological similarity shows a weaker positive correlation (0.17), suggesting a comparatively lesser impact on transfer effectiveness in this task. Moreover, the lexical similarity demonstrates a strong negative correlation (-0.83), aligning with the inverse relationship seen in the QA task, highlighting its substantial influence on cross-lingual transfer for NER. Additionally, the language model performance (LM) exhibits a relatively stronger positive correlation (0.45) in the NER task compared to the QA task, indicating a more pronounced association between overall model performance and cross-lingual transfer effectiveness specifically in Named Entity Recognition.

Significant disparities emerge among the observed features, underscoring the necessity for tailoring a task-specific multivariate linear model. While these correlations offer glimpses into associations, they each present a limited perspective on the intricate interplay between language proximity and cross-lingual transfer. The correlation coefficients, consistently distant from the value of 1, signify the inability of any single metric to comprehensively elucidate cross-lingual transfer dynamics. One striking limitation surfaces in character-level 3-grams, confined to languages sharing identical scripts. Consequently, languages with disparate scripts exhibit substantial lexical divergence. However, intriguingly, even between languages scripted differently, the divergence remains finite, although notably elevated. This phenomenon stems from numerical representations and residual Latin tokens inherent in datasets like Arabic, Russian, and Korean, attributing to this persistent divergence.

Remarkably, an intriguing scenario unfolds concerning Arabic's transfer dynamics, where its similarity with Swahili closely mirrors that with Indonesian. Despite this parity, Indonesian showcases notably superior transfer performance. Furthermore, instances arise where language similarity metrics inadequately explicate cross-lingual transfer phenomena. These nuanced observations highlight the intricate nature of cross-lingual transfer, showcasing that diverse linguistic and contextual factors contribute to its efficacy beyond simplistic similarity metrics.

In summary, we've introduced multiple predictors that collectively exhibit a strong linear correlation with cross-lingual performance. Subsequently, in the upcoming sections, we demonstrate that these predictors, when combined linearly, effectively anticipate cross-lingual transfer with considerable accuracy.

### 4.2 Regression framework

Based on our bi-variate analysis in the previous section, we propose that the fundamental factor governing zero-shot learning during cross-lingual transfer is the similarity between the source and target languages.

We can simplify the equation 1 as follows:

$$S_T = f(S_S, SYN_{S,T}, PHON_{S,T}, MORPH_{S,T}, LM) \tag{2}$$

In simplifying our analysis to focus on the most relevant features, we utilize Lasso Regression [12] coupled with recursive feature elimination based on absolute coefficient values.

Our assessment methodology, drawing inspiration from [13], employs a meta-regression model trained via k-folds cross-validation. This method ensures that each fold contains observations exclusively from an individual target language. By fitting the model on a concatenation of k-1 language-folds and evaluating it on the kth fold, we derive an average k-cross-validation score computed across all target languages. For assessing the regression's performance, we rely on the Root Mean Squared Error (RMSE), a widely used metric in evaluating regression models.

Moreover, our analysis extends to predicting the optimal source language for a specific target language with a given number of annotated samples. This prediction involves solving an argmax function across all possible language pairs using Equation 2. Evaluating the accuracy of this prediction, referred to as $A_{SRC}$, entails comparing the predicted best source language with the verified best source language.

## 5 Results

Our analysis underscores the effective representation of zero-shot cross-lingual transfer across the three tasks through a linear combination of the mentioned features. The outcomes of this regression are concisely detailed below. It's noteworthy that a refined model, incorporating the most pertinent attributes, demonstrates an impressive explanatory capability for the variability observed in Question Answering (QA) performance during zero-shot transfer. This streamlined model achieves an exceptionally low Root Mean Squared Error (RMSE) of 0.066, highlighting its proficiency in capturing the variance within QA performance.

Moreover, our model showcases a reasonable accuracy in forecasting the optimal source language, achieving a commendable accuracy rate of 62.5% in QA instances. The coefficients derived from the regression, encompassing all language-folds, offer valuable insights into the interpretability of these relationships. Specifically, certain features like syntactic, morphological, and lexical similarity emerge as robust predictors significantly impacting transfer learning performance across tasks.

Below is table with RMSE and $A_{SRC}$ for the Regression models for both QA and NER tasks:

| Task | RMSE | $A_{SRC}$ |
|---|---|---|
| QA | 0.066 | 62.5% |
| NER | 0.236 | 50% |

Another noteworthy observation is the pronounced and positive coefficient exhibited by the language model performance in the target language across the tasks. This finding indicates that a more proficient language model specific to the target language substantially enhances the predictive capacity for task-specific performance improvements.

The below equations elucidate intricate associations between predictors and performance within the target language, yielding deeper insights into these nuanced interconnections:

### 5.1 Question Answering

$$\begin{aligned} S_T = {} & 0.04 * SYN_{S,T} - 0.03 * PHON_{S,T} \\ & - 0.131 * MORPH_{S,T} - 2.023 * LEX_{S,T} \\ & + 0.574 * LM + 0.547 * S_S \end{aligned} \tag{3}$$

The regression equation for QA models the target language's model performance ($S_T$) based on several predictors. Each predictor's coefficient indicates its impact on the target language's performance. Notably, positive coefficients for syntactic similarity (SYN), language model performance (LM), and source language's model performance ($S_S$) suggest that increases in these factors correspond to higher performance in the target language's model. Conversely, negative coefficients for phonological (PHON), morphological

(MORPH), and lexical (LEX) similarities imply that higher similarity in these linguistic aspects could potentially decrease the target language's model performance. The statistical significance of these coefficients indicates the relative strength and direction of influence each predictor holds in determining the performance of the target language's model in this regression framework.

### 5.2 Named Entity Recognition

$$\begin{aligned} S_T &= 1.33 * SYN_{S,T} - 1.17 * PHON_{S,T} \\ &+ 1.43 * MORPH_{S,T} - 0.06 * LEX_{S,T} \\ &+ 0.46 * LM + 1.99 * S_S \end{aligned} \tag{4}$$

The regression equation for NER task unveils influential factors dictating the target language's model performance. Syntactic and morphological similarities exhibit substantial positive coefficients, indicating their considerable positive impact on the target language's model performance. Conversely, phonological similarity demonstrates a pronounced negative coefficient, implying a significant adverse effect on the target language's model performance. In contrast, lexical similarity appears to have a minimal impact, indicated by its relatively negligible coefficient. Moreover, both the language model performance and the source language's model performance continue to exhibit positive coefficients, highlighting their significant contributions to enhancing the target language's model performance. This equation underscores the varied and distinct impacts of different linguistic similarities and model performances on the target language's model proficiency compared to the QA's regression inference.

Additionally, we conduct a comparison between the Observed Model Performance on the Target Language and the Predicted Model Performance on the Target Language subsequent to fine-tuning the model on a designated Source Language. Figures 1 and 2 depict the graphical illustrations that delineate this comparative analysis for the QA task. These graphs provide a visual depiction of the contrast between the actual model performance observed in the target language and the anticipated model performance derived from fine-tuning the model on a specific source language. As we can observe, the framework's predictions do not stray from the observed performance by a large margin. There are no major outliers in the prediction.

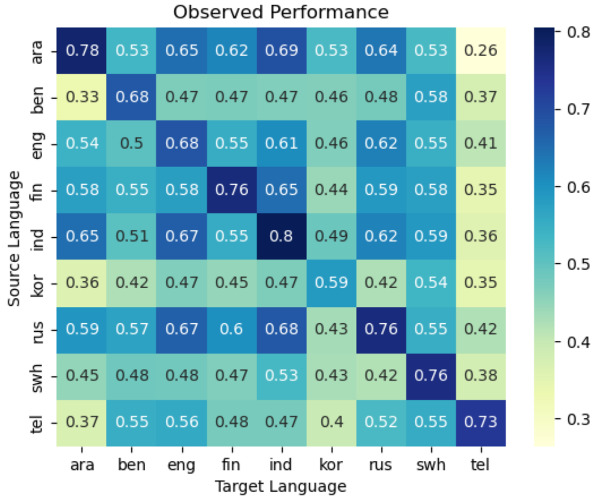

Figure 1: Observed Performance of a Target Language on a model fine-tuned on a Source Language

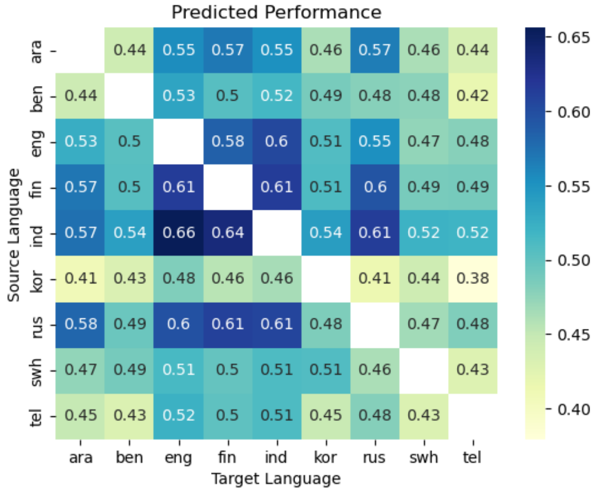

Figure 2: Predicted Performance of a Target Language on a model fine-tuned on a Source Language

The model demonstrates remarkable explanatory capability for Question Answering (QA) performance, achieving a low Root Mean Squared Error (RMSE) of 0.066 and showcasing an accuracy rate of 62.5% in predicting the optimal source language for QA instances. Notably, syntactic, morphological, and lexical similarities emerge as robust predictors significantly influencing transfer learning performance. The regression equations for QA and Named Entity Recognition (NER) tasks further elucidate the intricate associations between predictors and target language performance. The graphical illustrations comparing observed and predicted model performance highlight the framework's accuracy, with predictions closely aligning with actual performance. This suggests that the model, especially in the context of QA, performs well in capturing the nuances of cross-lingual transfer and underscores the importance of specific linguistic similarities in predicting target language proficiency.

## 6 Conclusion

This study delves into the examination of a pretrained mBERT model to delve deeper into the mechanics of cross-lingual transfer. By conducting comprehensive model interpretation experiments across various language pairs and tasks, we've unearthed significant insights. Our findings highlight the possibility of statistically modeling transfer through a select set of linguistic and data-derived features. Notably, we've established that the syntax, morphology, and phonology of languages serve as robust predictors of cross-lingual transfer, surpassing the predictive capacity of lexical similarity between languages. Moreover, our analysis underscores the relevance of language model performance as a crucial indicator of cross-lingual prowess, presenting a readily available metric to better understand and facilitate cross-lingual transfer processes.

**Future Work:** Presently, our evaluation focuses on assessing the zero-shot inference capabilities of the mBERT model. To enhance our understanding and provide a more comprehensive analysis, we propose expanding our methodology to include training the model on a reduced subset of the target languages dataset. This modification enables us to measure the few-shot capabilities of the model, offering insights into its accuracy when exposed to a limited amount of training data. Incorporating this additional dimension into our evaluation framework has the potential to serve as a valuable predictor, providing a more nuanced and robust assessment of the model's cross-lingual transfer performance.

By integrating Latent Dirichlet Allocation (LDA) as an additional predictor, we could tackle diverse corpora disparities. LDA, a probabilistic topic modeling technique, allows us to unveil latent themes within multilingual text datasets. Leveraging LDA empowers us to decipher the underlying thematic structures of individual language corpora, shedding light on their unique content nuances. This enriched understanding aids in quantifying and accounting for dissimilarities across languages. Moreover, utilizing LDA-derived topics as features expands our predictive framework's scope, enabling us to assess how topic distributions impact cross-lingual transfer effectiveness.

## 7 Division of work

The work distribution was structured as follows: Karan dedicated efforts to refining models for Named Entity Recognition (NER) and investigating lexical similarities among chosen languages. Meanwhile, Rishikesh focused on optimizing Question Answering (QA) models and exploring phonological, morphological, and syntactic language commonalities. Additionally, Rishikesh formulated a regression model utilizing these linguistic analyses as predictive factors. Both of us contributed equally to every facet of the project, including the development of presentations and the compilation of this report.

## 8 Code Base

We have created a repository on GitHub that contains all the code we have used in our project. To visit the repository click here (url: https://github.com/EECS595-Multilingual/Is-it-easy-to-be-multilingual).

Please feel free to reach out to us if you have any questions about the code or need help reproducing it.

## 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]

[12] Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288, 1996.

[13] Mengzhou Xia, Antonios Anastasopoulos, Ruochen Xu, Yiming Yang, and Graham Neubig. Predicting performance for natural language processing tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8625–8646, Online, July 2020. Association for Computational Linguistics.

[14] Benjamin Muller, Deepanshu Gupta, Siddharth Patwardhan, Jean-Philippe Fauconnier, David Vandyke, Sachin Agarwal. Languages You Know Influence Those You Learn: Impact of Language Characteristics on Multi-Lingual Text-to-Text Transfer. arXiv:2212.01757 [cs.CL]