
# Is It Easy To Be Multilingual?

Rishikesh Ksheersagar , Karan Anand
(rishiksh@umich.edu, karanand@umich.edu)

# Introduction

- Study investigates multi-lingual models' efficacy (mBERT) for cross-lingual transfer.
- Focuses on fine-tuning models, highlighting their zero-shot capabilities.
- Statistical framework explores lexical, morphological, phonological, and syntactic similarities' impact on cross-lingual transfer across tasks (NER, QA, and XNLI) in 81 language pairs.
- Aims to unveil mechanisms behind cross-lingual transfer, emphasizing language similarity and model characteristics' roles in facilitating this process.
- **Goal** is to design a statistical framework to identify best source language for zero-shot cross-lingual transfer for a target language.

| **Languages** |
|---|
| Arabic |
| Bengali |
| English |
| Finnish |
| Indonesian |
| Korean |
| Russian |
| Swahili |
| Telugu |

# Previous Work

- [1] and [2] highlighted the remarkable zero-shot transfer capabilities of large multi-lingual language models, emphasizing their efficacy in low-resource languages.
- [3] underscored the importance of "structural similarity" between source and target languages, surpassing mere lexical overlap or word frequency considerations.
- [4] and [5] demonstrated several methods to predict cross-lingual task performance.

---

## 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 Siddhant, 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.

# Methods

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

We try to model the cross-lingual transfer between 81 pairs of languages based on the equation above.

- $S_S$ and $S_T$ are the performances on the source and target languages respectively.
- $LM$ term refers to performance of mBERT as a language model before any task-specific fine tuning is done.
- $S_S$ , $S_T$ and $LM$ metrics are based on commonly utilized NLP Tasks which comprises of:
  - Question Answering
  - Named Entity Recognition
  - Cross-Lingual Natural language Inference

- $LS_{S,T}$ is the Language SImilarity metric which comprises of:
  - Syntactic similarity
  - Phonological similarity
  - Morphological similarity
  - Lexical similarity

# Methods | NLP tasks

For the following tasks, the pre-trained mBERT model is fine-tuned using a source language and then evaluated for its performance against a target language.

- QA (Question Answering):
  - QA tasks focus on leveraging natural language processing to enable machines to comprehend and respond to user questions by extracting relevant information from textual data.
  - Used **Ty**pologically **Di**verse **Q**uestion **A**nswering dataset (TyDi QA - GoldP Task)

- NER (Named Entity Recognition):
  - NER tasks involve identifying and classifying entities, such as names of people, organizations, locations, and more, within a given text.
  - Used WikiANN dataset

- XNLI (Cross-Lingual Natural Language Inference):
  - NLI tasks involve evaluating the ability of natural language processing models to understand and perform textual entailment.

# Methods | Language Similarity

- Syntactic Similarity:
  - Resemblance in the structure and arrangement of words or phrases within sentences
- Phonological Similarity:
  - Likeness in the sounds, pronunciation, and phonetic characteristics between words or language units
- Morphological Similarity:
  - Similarity in the structure and formation of words, including prefixes, suffixes, and root words

We obtained the above 3 metrics from **World Atlas of Language Structures (WALS)** dataset using *lang2vec*

- Lexical Similarity:
  - Likeness or resemblance in vocabulary, words, or lexical items

Computed character 3-gram distributions for both source and target languages using their respective **training datasets** in each experiment. These distributions are used to calculate a normalized Jensen-Shannon divergence (JSD) between the source and target distributions.

# NLP Tasks | QA & NER

- Used the TyDiQA dataset's GoldP Task for Question Answering, WikiANN dataset for Named Entity Recognition.
- TyDiQA consists of 9 languages, used data for same 9 languages from WikiANN dataset.

- Trained separate models for each of the 9 languages, and obtained inferences for each model on all the 9 languages. (9 models trained, 9x9 inferences)

- Fine tuned mBERT (multilingual-bert-base-uncased) with the following hyperparameters:
  - Optimizer: Adam
  - Epochs: 3
  - Batch SIze: 32
  - Learning Rate: 2e-5
  - Weight Decay: 0.01

- **Generated the metrics $\mathbf{LM}$ , $\mathbf{S_S}$ and $\mathbf{S_T}$**
  - $LM$ is the F1 score on Target Language before fine-tuning.
  - $S_S$ is the Testing F1 score on the Source Language after training on the same language.
  - $S_T$ is the Testing F1 score on the Target Language after training on a different source language.

# Results | QA

| | s | t | syn | phon | morph | lex | LM | S_s | S_t |
|---|---|---|---|---|---|---|---|---|---|
| 0 | ben | ara | 0.296296 | 0.615385 | 0.111111 | 0.343902 | 0.124259 | 0.679622 | 0.333755 |
| 1 | eng | ara | 0.421053 | 0.600000 | 0.025641 | 0.304094 | 0.124259 | 0.683939 | 0.542399 |
| 2 | fin | ara | 0.438596 | 0.600000 | 0.076923 | 0.298701 | 0.124259 | 0.764442 | 0.575941 |
| 3 | ind | ara | 0.470588 | 0.600000 | 0.025641 | 0.309252 | 0.124259 | 0.804405 | 0.648013 |

- A regression model was fitted using k-fold cross validation.

$$S_T = 0.04*SYN - 0.03*PHON - 0.131*MORPH - 2.023*LEX + 0.574*LM + 0.547*S_S$$

| Task | Root Mean Squared Error | Top-2 Source Prediction Accuracy |
|---|---|---|
| Question Answering | 0.066 | 62.5% |

- Lexical Divergence and Morphological Similarity are key predictors for the cross-lingual transfer performance along with $S_S$ and LM.
- **The framework reasonably predicts the best Source Language for any given Target Language for cross-lingual transfer for QA Task based on the language similarity and performance metrics.**

# Next steps

- Immediate (within scope of the project):
  - Fitting Regression model for the NER task. (Next slide)

- Future Improvements:
  - Few shot training will help improve the model's accuracy and could be added as a key predictor in this framework.
  - Testing the effects of the dataset on the effectiveness of transfer - Some tasks did not have an appropriate dataset to be worked on, so a future plan could be to generate our own datasets by translating one from a standard language to the required languages.
  - Using a larger variety of NLP tasks and similarity metrics to gain a clearer understanding of language similarity.
  - To account for dissimilarity in corpora of different languages, we can try topic modeling using Latent Dirichlet Allocation as another predictor.

# Results | NER

| | s | t | syn | phon | morph | lex | LM | S_s | S_t |
|---|---|---|---|---|---|---|---|---|---|
| 0 | ben | ara | 0.296296 | 0.615385 | 0.111111 | 0.013625 | 0.110902 | 0.241917 | 0.111306 |
| 1 | eng | ara | 0.421053 | 0.600000 | 0.025641 | 0.000000 | 0.110902 | 0.620021 | 0.121354 |
| 2 | fin | ara | 0.438596 | 0.600000 | 0.076923 | 0.000000 | 0.110902 | 0.667381 | 0.399154 |
| 3 | ind | ara | 0.470588 | 0.600000 | 0.025641 | 0.010320 | 0.110902 | 0.113821 | 0.507614 |

- A regression model was fitted using k-fold cross validation.

$$S_T = 1.33*SYN - 1.17*PHON + 1.43*MORPH - 0.06*LEX + 0.46*LM + 1.99*S_S$$

| Task | Root Mean Squared Error | Top-2 Source Prediction Accuracy |
|---|---|---|
| Question Answering | 0.236 | 50% |

- All the Similarity Features apart from Lexical Divergence are key predictors for the cross-lingual transfer performance along with $S_S$ and LM for the NER task.

# Presentation Questions | Q1

## How is sound similarity calculated?

- We calculated the “sound” similarity of languages through their phonological similarity. Phonological similarity refers to the similarity in sound or pronunciation of words or linguistic elements. We extracted the phonological vectors for all Languages from the **World Atlas of Language Structures (WALS)** database by using `lang2vec.get_features({lang}, “phonology_wals”)` and calculated the similarity by computing the intersection over the union of the obtained vector of the source language with the vector of the target language.

WALS dataset: https://wals.info/

Lang2vec: https://github.com/antonisa/lang2vec

# Presentation Questions | Q2

**Were the best source languages dependent on the target language or was there a universal best source language? Did being multilingual help in any of the monolingual tasks?**

- **Language specificity**: Our framework emphasized the specific relationship between source and target languages, highlighting the importance of linguistic alignment over a universal best target language. Clearly, the model performance on Target Language $S_T$ is highly dependent on that of the specific Source Language $S_S$ for both the tasks.

- **Multilingual advantage and complexities**: Multilingualism offered performance advantages by leveraging shared knowledge across languages, impacting both multilingual and monolingual tasks. Understanding these complexities is crucial for diverse multilingual applications.

# Presentation Questions | Q3

**Were there any particular bottlenecks to your development of the project ?**

- It was difficult to find datasets for the different NLP tasks that shared similar languages. This was particularly the case for the **XNLI** task, so we could not generate results for this task. A plan to tackle this issue in the future was to generate our own datasets for all the languages by translating a dataset from one language to all other required languages. This would make the data consistent across languages, but we would run into the issue of finding a translator that is accurate and can be used to translate datasets that content many hundreds of thousands of data points.


Thank you
Ευχαριστώ
Kiitos
Dziękuję
Obrigado
谢谢
Hvala
Merci
תודה
Danke
Gracias
Thank you
Grazie
شكرا
謝謝
감사합니다
Спасибі
Mulţumesc
Asante