Beyond Static Models and Test Sets: Benchmarking the Potential of Pre-trained Models Across Tasks and Languages

Kabir Ahuja[1] Sandipan Dandapat[2] Sunayana Sitaram[1] Monojit Choudhury[2]
[1] Microsoft Research, India
[2] Microsoft R&D, India
{t-kabirahuja,sadandap,sunayana.sitaram,monojitc}@microsoft.com

Abstract

Although recent Massively Multilingual Language Models (MMLMs) like mBERT and XLMR support around 100 languages, most existing multilingual NLP benchmarks provide evaluation data in only a handful of these languages with little linguistic diversity. We argue that this makes the existing practices in multilingual evaluation unreliable and does not provide a full picture of the performance of MMLMs across the linguistic landscape. We propose that the recent work done in Performance Prediction for NLP tasks can serve as a potential solution in fixing benchmarking in Multilingual NLP by utilizing features related to data and language typology to estimate the performance of an MMLM on different languages. We compare performance prediction with translating test data with a case study on four different multilingual datasets, and observe that these methods can provide reliable estimates of the performance that are often on-par with the translation based approaches, without the need for any additional translation as well as evaluation costs.

1 Introduction

Recent years have seen a surge of transformer (Vaswani et al., 2017) based Massively Multilingual Language Models (MMLMs) like mBERT (Devlin et al., 2019) , XLM-RoBERTa (XLMR) (Conneau et al., 2020), mT5 (Xue et al., 2021), RemBERT (Chung et al., 2021). These models are pretrained on varying amounts of data of around 100 linguistically diverse languages, and can in principle support fine-tuning on different NLP tasks for these languages.

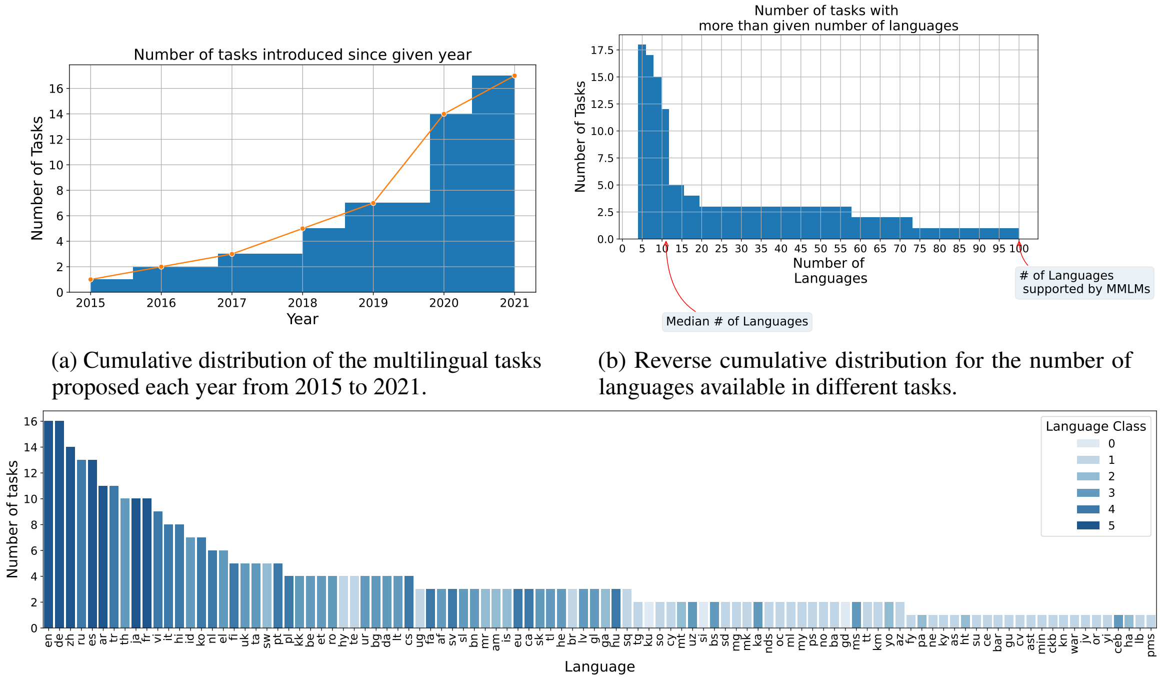
These MMLMs are primarily evaluated for their performance on Sequence Labelling (Nivre et al., 2020; Pan et al., 2017), Classification (Conneau et al., 2018; Yang et al., 2019; Ponti et al., 2020), Question Answering (Artetxe et al., 2020; Lewis et al., 2020; Clark et al., 2020a) and Retrieval (Artetxe and Schwenk, 2019; Roy et al., 2020; Botha et al., 2020) tasks. However, most these tasks often cover only a handful of the languages supported by the MMLMs, with most tasks having test sets in fewer than 20 languages (cf. Figure 1b).

Evaluating on such benchmarks henceforth fails to provide a comprehensive picture of the model's performance across the linguistic landscape, as the performance of MMLMs has been shown to vary significantly with the amount of pre-training data available for a language (Wu and Dredze, 2020), as well according to the typological relatedness between the *pivot* and *target* languages (Lauscher et al., 2020). While designing benchmarks to contain test data for all 100 languages supported by the MMLMs is be the ideal standard for multilingual evaluation, doing so requires prohibitively large amount of human effort, time and money.

Machine Translation can be one way to extend test sets in different benchmarks to a much larger set of languages. Hu et al. (2020) provides pseudo test sets for tasks like XQUAD and XNLI, obtained by translating English test data into different languages, and shows reasonable estimates of the actual performance by evaluating on translated data but cautions about their reliability when the model is trained on translated data. The accuracy of translation based evaluation can be affected by the quality of translation and the technique incurs non-zero costs to obtain reliable translations. Moreover, transferring labels with translation might also be non-trivial for certain tasks like Part of Speech Tagging and Named Entity Recognition.

Recently, there has been some interest in predicting performance of NLP models without actually evaluating them on a test set. Xia et al. (2020) showed that it is possible to build regression models that can accurately predict evaluation scores of NLP models under different experimental settings using various linguistic and dataset specific features. Srinivasan et al. (2021) showed promising

Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP, pages 64 - 74
May 26, 2022 ©2022 Association for Computational Linguistics



(a) Cumulative distribution of the multilingual tasks proposed each year from 2015 to 2021.

(b) Reverse cumulative distribution for the number of languages available in different tasks.

(c) Number of multilingual tasks containing test data for each of the 106 languages supported by the MMLMs (mBERT, XLMR). The bars are shaded according to the class taxonomy proposed by Joshi et al. (2020).

Figure 1

results specifically for MMLMs towards predicting their performance on downstream tasks for different languages in zero-shot and few-shot settings, and Ye et al. (2021) propose methods for more reliable performance prediction by estimating confidence intervals as well as predicting fine-grained performance measures.

In this paper we argue that the performance prediction can be a possible avenue to address the current issues with Multilingual benchmarking by aiding in the estimation of performance of the MMLMs for the languages which lack any evaluation data for a given task. Not only this can help us give a better idea about the performance of a multilingual model on a task across a much larger set of languages and hence aiding in better model selection, but also enables applications in devising data collection strategies to maximize performance (Srinivasan et al., 2022) as well as in selecting the representative set of languages for a benchmark (Xia et al., 2020).

We present a case study demonstrating the effectiveness of performance prediction on four multilingual tasks, PAWS-X (Yang et al., 2019) XNLI (Conneau et al., 2018), XQUAD (Artetxe et al., 2020) and TyDiQA-GoldP (Clark et al., 2020a) and show that it can often provide reliable estimates of the performance on different languages on par with evaluating them on translated test sets without any additional translation costs. We also demonstrate an additional use case of this method in selecting the best pivot language for fine-tuning the MMLM in order to maximize performance on some target language. To encourage research in this area and provide easy access for the community to utilize this framework, we will release our code and the datasets that we use for the case study.

2 The Problem with Multilingual Benchmarking

The rise in popularity of MMLMs like mBERT and XLMR have also lead to an increasing interest in creating different multilingual benchmarks to evaluate these models. We analyzed 18 different multilingual datasets proposed between the years 2015 to 2021, by searching and filtering for datasets containing the term *Cross Lingual* in the Papers with Code Datasets repository.[1] The types and language specific statistics of these studied benchmarks can be found in Table 3 in appendix.

As can be seen in Figure 1a, there does appear to be an increasing trend in the number of multilingual datasets proposed each year, especially with a sharp increase observed during the year 2020. However,

[1] https://paperswithcode.com/datasets

if we look at the number of languages covered by these different benchmarks (Figure 1b), we see that most of the tasks have fewer than 20 languages supported with a median of 11 languages per task which is substantially lower than the 100 supported by the commonly used MMLMs.

The only tasks which have been able to support a large fraction of these 100 languages are the Sequence Labelling tasks WikiANN (Pan et al., 2017) and Universal Dependencies(Nivre et al., 2020) which were a result of huge engineering, crowd sourcing and domain expertise efforts, and the Tatoeba dataset created from the parallel translation database maintained since more than 10 years, consisting of contributions from tens of thousands of members. However, we observed a dearth of supported languages in the remaining tasks that we surveyed, especially in NLU tasks.

We also observe a clear lack of diversity in the selected languages across different multilingual datasets. Figure 1c shows the number of tasks each language supported by the mBERT is present in and we observe a clear bias towards high resource languages, mostly covering class 4 and class 5 languages identified according to the taxonomy provided by Joshi et al. (2020). The low resource languages given by class 2 or lower are severely under-represented in the benchmarks where the most popular (in terms of number of tasks it appears in) class 2 language i.e. Swahili appears only in 5 out of 18 benchmarks.

We also categorized the the languages into the 6 major language families at the top level genetic groups [2] each of which cover at least 5% of the world's languages and plot language family wise representation of each task in Figure 2. Except a couple of benchmarks, the majority of the languages present in these tasks are Indo-European, with very little representation from all the other language families which have either comparable or a higher language coverage as Indo-European.

There have been some recent benchmarks that address this issue of language diversity. The TyDiQA (Clark et al., 2020a) benchmark contains training and test datasets in 11 typologically diverse languages, covering 9 different language families. The XCOPA (Ponti et al., 2020) benchmark for causal commonsense reasoning also selects a set of 10 languages with high genealogical and areal diversities.

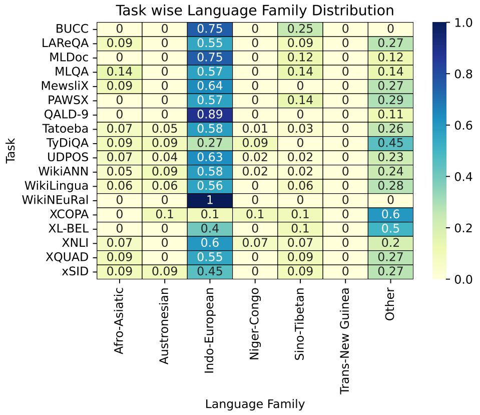


Figure 2: Task wise distribution of language families i.e. fraction of languages belonging to a particular language for a task.

While this is a step in the right direction and does give a much better idea about the performance of MMLMs over a diverse linguistic landscape, it is still difficult to cover through 10 or 11 languages all the factors that influence the performance of an MMLM like pre-training size (Wu and Dredze, 2020; Lauscher et al., 2020), typological relatedness (syntactic, genealogical, areal, phonological etc) between the source and pivot languages (Lauscher et al., 2020; Pires et al., 2019), sub-word overlap (Wu and Dredze, 2019), tokenizer quality (Rust et al., 2021) etc. Through Performance Prediction as we will see in next section, we seek to estimate the performance of an MMLMs on different languages based on these factors.

We would also like to point out that there are other problems with multilingual benchmarking as well. Recent multi-task multilingual benchmarks like X-GLUE (Liang et al., 2020), XTREME (Hu et al., 2020) and XTREME-R (Ruder et al., 2021) mainly provide training datasets for different tasks only in English and evaluate for zero-shot transfer to other languages. However, this standard of using English as a default pivot language was put in question by Turc et al. (2021), who showed empirically that German and Russian transfer more effectively to a set of diverse target languages. We shall see in the coming sections that the Performance Prediction approach can also be useful in identifying the best pivots for a target language.

3 Performance Prediction for Multilingual Evaluation

We define Performance Prediction as the task of predicting performance of a machine learning model

[2] https://www.ethnologue.com/guides/largest-families

on different configurations of training and test data. Consider a multilingual model $\mathcal{M}$ pre-trained on a set of languages $\mathcal{L}$, and a task $\mathfrak{T}$ containing training datasets $\mathcal{D}_{tr}^{p}$ in languages $p \in \mathcal{P}$ such that $\mathcal{P} \subset \mathcal{L}$ and test datasets $\mathcal{D}_{te}^{t}$ in languages $t \in \mathcal{T}$ such that $\mathcal{T} \subset \mathcal{L}$. Following Amini et al. (2009), we assume that both $\mathcal{D}_{tr}^{p}$ and $\mathcal{D}_{te}^{t}$ are the subsets of a multi-view dataset $\mathcal{D}$ where each sample $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}$ has multiple views (defined in terms of languages) of the same object i.e. $(\mathbf{x}, \mathbf{y}) \overset{\text{def}}{=} \{(x^l, y^l) | \forall l \in \mathcal{L}\}$ all of which are not observed.

A training configuration for fine-tuning $\mathcal{M}$ is given by the tuple (Π, Δ_{tr}^{Π}), where $\Pi \subseteq \mathcal{P}$ and $\Delta_{tr}^{\Pi} = \bigcup\limits_{p \in \Pi} \mathcal{D}_{tr}^{p}$. The performance on the test set $\mathcal{D}_{te}^{t}$ for language $t \in \mathcal{T}$ when $\mathcal{M}$ is fine-tuned on (Π, Δ_{tr}^{Π}) is denoted as $s_{\mathcal{M}, \mathfrak{T}, t, \mathcal{D}_{te}^{t}, \Pi, \Delta_{tr}^{\Pi}}$ or s for clarity, given as:

$$s = g(\mathcal{M}, \mathfrak{T}, t, \mathcal{D}_{te}^{t}, \Pi, \Delta_{tr}^{\Pi}) \tag{1}$$

In performance prediction we formulate estimating g by a parametric function f_θ as a regression problem such that we can approximate s for various configurations with reasonable accuracy, given by

$$s \approx f_\theta([\phi(t); \phi(\Pi); \phi(\Pi, t); \phi(\Delta_{tr}^{\Pi})]) \tag{2}$$

where $\phi(.)$ denotes the features representation of a given entity. Following Xia et al. (2020), we do not consider any features specific to $\mathcal{M}$ to focus more on how the performance varies for a given model with different data and language configurations. Since the languages for which we are trying to predict the performance might not have any data (labelled or unlabelled available), we also skip features for $\mathcal{D}_{te}^{t}$ from the equation. Note, we do consider coupled features for training and test languages i.e. $\phi(\Pi, t)$ as the interaction between the two has been shown to be a strong indicator of the performance of such models (Lauscher et al., 2020; Wu and Dredze, 2019).

Different training setups for multilingual models can be seen as special cases of our formulation. For zero-shot transfer we set $\Pi = \{p\}$, such that $p \neq t$. This reduces the performance prediction problem to the one described in Lauscher et al. (2020).

$$s_{zs} \approx f_\theta([\phi(t); \phi(p); \phi(p, t); \phi(\Delta_{tr}^{\{p\}})]) \tag{3}$$

There are many ways to represent the feature representations $\phi(.)$ that have been explored in previous work, including pre-training data size, typological relatedness between the pivot and target languages and more. For a complete list of features that we use in our experiments, refer to Table 1.

Type	Features	Reference
$\phi(t)$	Pre-training Size of t	Srinivasan et al. (2021); Lauscher et al. (2020)
	Tokenizer Quality for t	Rust et al. (2021)
$\phi(\Pi)$	Pre-training size of every $p \in \Pi$	
$\phi(\Pi, t)$	Subword Overlap between p and t for $p \in \Pi$	Lin et al. (2019); Xia et al. (2020); Srinivasan et al. (2021)
	Relatedness between lang2vec (Littell et al., 2017) features	Lin et al. (2019); Xia et al. (2020); Lauscher et al. (2020); Srinivasan et al. (2021)
$\phi(\Delta_{tr}^{\Pi})$	Training size $\lvert\mathcal{D}_{tr}^{p}\rvert$ of each language $p \in \Pi$	(Lin et al., 2019; Xia et al., 2020; Srinivasan et al., 2021)

Table 1: Features used to represent the languages and datasets used. For more details refer to Section A.2 in Appendix.

4 Case Study

To demonstrate the effectiveness of Performance Prediction in estimating the performance on different languages, we evaluate the approach on classification tasks i.e. PAWS-X and XNLI, and two Question Answering tasks XQUAD and TyDiQA-GoldP. We choose these tasks as their labels are transferable via translation, so we can compare our method with the automatic translation based approach. TyDiQA-GoldP has test sets for different languages created independently to combat the *translationese* problem (Clark et al., 2020b), while the other three have English test sets manually translated to the other languages.

4.1 Experimental Setup

For all the three tasks we try to estimate zero-shot performance of a fine-tuned mBERT model i.e. s_{zs} on different languages. For PAWS-X, XNLI and

Task	Baseline	Translate	Performance Predictors	
			XGBoost	Group Lasso
PAWS-X	7.18	3.85	5.46	**3.06**
XNLI	5.32	**2.70**	3.36	3.93
XQUAD	6.89	**3.42**	5.41	4.53
TyDiQA-GoldP	7.82	7.77	5.04	**4.73**

Table 2: Mean Absolute Errors (MAE) (scaled by 100 for readability) on the the three tasks for different methods of estimating performance.

XQUAD we have training data present only in English i.e. $\Pi = \{en\}$ always, but TyDiQA-GoldP contains training sets in 9 different languages and we predict transfer from all of those. To train Performance Prediction models we use the performance data for mBERT provided in Hu et al. (2020) as well as train our own models when required and evaluate the performance on test dataset of different languages. The performance prediction models are evaluated using a leave one out strategy also called *Leave One Language Out* (LOLO) as used in Lauscher et al. (2020); Srinivasan et al. (2021), where we use the performance data of target languages in the set $\mathcal{T} - \{t\}$ to predict the performance on a language t and do this for all $t \in \mathcal{T}$.

4.2 Methods

We compare the following methods for estimating the performance:

1. Average Score Baseline: In this method, to estimate the performance on a target language t we simply take a mean of the model's performance on the remaining $\mathcal{T} - \{t\}$ languages. Although conceptually simple, this is an unbiased estimate for the expected performance of the MMLM on different languages.

2. Translate: To estimate the performance on language t with this method, we automatically translate the test data in one of the languages $t' \in \mathcal{T} - \{t\}$,[3] to the target language t and evaluate the fine-tuned MMLM on the translated data. The performance on this pseudo-test set is used as the estimate of the actual performance. We use the Azure Translator[4] to translate the test sets.

3. Performance Predictors: We consider two different regression models to estimate the performance in our experiments.

i) **XGBoost**: We use the popular Tree Boosting algorithm XGBoost for solving the regression problem, which has been previously shown to achieve impressive results on the task (Xia et al., 2020; Srinivasan et al., 2021).

ii) **Group Lasso**: Group Lasso (Yuan and Lin, 2006) is a multi-task linear regression model that uses an l_1/l_q norm as a regularization term to ensure common sparsity patterns among the regression weights of different tasks. In our experiments, we use the performance data for all the tasks in the XTREME-R (Ruder et al., 2021) benchmark to train group lasso models.

4.3 Results

The average LOLO errors for the four tasks and the four methods are given in Table 2. As we can see both Translated baseline and Performance Predictors can obtain much lower errors compared to the Average Score Baseline on PAWS-X, XNLI and XQUAD tasks. Group Lasso outperforms all the other methods on PAWS-X dataset while for XNLI and XQUAD datasets though, the Translate method outperforms the two performance predictor models.

On TyDiQA-GoldP dataset , which had its test sets for different languages created independently without any translation, we see that the performance of Translate method drops with errors close to those obtained using the Average Score Baseline. While this behaviour is expected since the translated test sets and actual test sets now differ from each other, it still puts the reliability of the performance on translated data compared to the real data into question. Both XGBoost and Group Lasso though, obtain consistent improvements over the Baseline for TyDiQA-GoldP as well.

Figure 3 provides a breakdown of the errors for each language included in TyDiQA-GoldP bench-

[3] for our experiments we use $t' = p$ i.e. we use test data in pivot language which is often English to translate to t

[4] https://azure.microsoft.com/en-us/services/cognitive-services/translator/

mark, and again we can see that the Performance Predictors can outperform the Translate method almost all the languages except Telugu (te). Similar plots for the other tasks can be found in Figure 5 of Appendix.

4.4 Pivot Selection

Another benefit of using Performance Prediction models is that we can use them to select training configurations like training (pivot) languages or amount of training data to achieve desired performance. For our case study we demonstrate the application of our predictors towards selecting the best pivot language for each of the 100 languages supported by mBERT that maximizes the predicted performance on the language. The optimization problem can be defined as:

$$p^*(l) = \arg\max_{p \in \mathcal{P}} f_\theta([\phi(l); \phi(p); \phi(p,l); \phi(\Delta_{tr}^{\{p\}})]) \quad (4)$$

Where $p^*(l)$ denotes the pivot language that results in the best predicted performance on language $l \in \mathcal{L}$. Since, $\mathcal{P} = \{en\}$ only for PAWS-X, XQUAD and XNLI i.e. training data is available only in English, we run this experiment on TyDiQA-GoldP dataset which has training data available in 9 languages i.e. $\mathcal{P} = \{ar, bn, es, fi, id, ko, ru, sw, te\}$. We solve the optimization problem exactly by evaluating Equation 4 for all (p, l) pairs using a linear search and we use XGBoost Regressor as f_θ.

The results of this exercise are summarized in Figure 4. We see carefully selecting the best pivot for each language leads to substantially higher estimated performances instead of using the same language as pivot for all the languages. We also see that languages like Finnish, Indonesian, Arabic and Russian have higher average predicted performance across all the supported languages compared to English. This observation is also in line with Turc et al. (2021) observation that English might not always be the best pivot language for zero-shot transfer.

5 Conclusion

In this paper we discussed how the current state of benchmarking multilingual models is fundamentally limited by the amount of languages supported by the existing benchmarks, and proposed Performance Prediction as a potential solution to address the problem. Based on the discussion we summarize our findings through three key takeaways

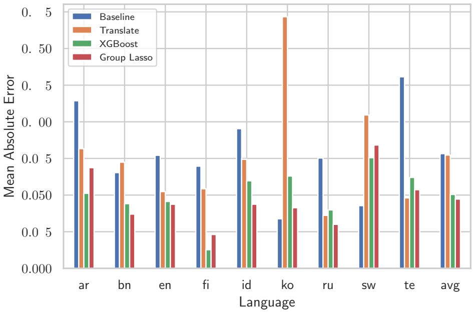


Figure 3: Language Wise Errors (LOLO setting) for predicting performances on the TyDiQA-GoldP dataset.

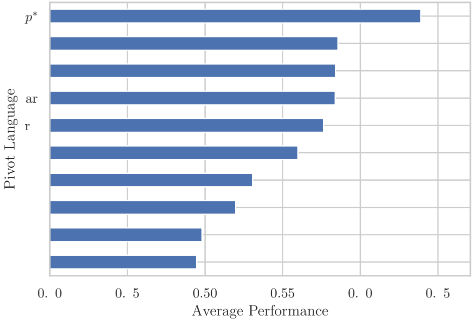


Figure 4: Average Performance on the 100 languages supported by mBERT for each of the 9 pivot languages for which training data is available in TyDiQA-GoldP.

1. Training performance prediction models on the existing evaluation data available for a benchmark can be a simple yet effective solution in estimating the MMLM's performance on a larger set of supported languages, which can often lead to much closer estimates compared to using the expected value estimate obtained from the existing languages.
2. One should be careful in using translated data to evaluate a model's performance on a language. Our experiments suggest that the performance measures estimated from the translated data can miscalculate the actual performance on the real world data for a language.
3. Performance Prediction can not only be effective for benchmarking on a larger set of languages but can also aid in selecting training strategies to maximize the performance of the MMLM on a given language which can be valuable towards building more accurate multilingual models.

Finally, there are a number of ways in which the current performance prediction methods can be improved for a more reliable estimation. Both Xia et al. (2020); Srinivasan et al. (2021) observed that these models can struggle to generalize on languages or configurations that have features that are remarkably different from the training data. Multi-task learning as hinted by Lin et al. (2019) and our experiments with Group Lasso can be a possible way to address this issue. The current methods also do not make use of model specific features for estimating the performance. Tran et al. (2019); Nguyen et al. (2020); You et al. (2021) explore certain measures like entropy values, maximum evidence derived from a pre-trained model to estimate the transferability of the learned representations. It can be worth exploring if such measures can be helpful in providing more accurate predictions.

References

Massih R. Amini, Nicolas Usunier, and Cyril Goutte. 2009. Learning from multiple partially observed views - an application to multilingual text categorization. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of ACL 2020*.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the ACL 2019*.

Jan A. Botha, Zifei Shan, and Daniel Gillick. 2020. Entity Linking in 100 Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7833–7845, Online. Association for Computational Linguistics.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020a. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. In *Transactions of the Association of Computational Linguistics*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020b. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP 2018*, pages 2475–2485.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. 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. Association for Computational Linguistics.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating Cross-lingual Extractive Question Answering. In *Proceedings of ACL 2020*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

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. 2019. 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. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Cuong V. Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. 2020. Leep: A new measure to evaluate transferability of learned representations.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of ACL 2017*, pages 1946–1958.

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

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. LAReQA: Language-agnostic answer retrieval from a multilingual pool. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5919–5930, Online. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Anirudh Srinivasan, Gauri Kholkar, Rahul Kejriwal, Tanuja Ganu, Sandipan Dandapat, Sunayana Sitaram, Balakrishnan Santhanam, Somak Aditya, Kalika Bali, and Monojit Choudhury. 2022. Litmus predictor: An ai assistant for building reliable, high-performing and fair multilingual nlp systems. In *Thirty-sixth AAAI Conference on Artificial Intelligence*. AAAI. System Demonstration.

Anirudh Srinivasan, Sunayana Sitaram, Tanuja Ganu, Sandipan Dandapat, Kalika Bali, and Monojit Choudhury. 2021. Predicting the performance of multilingual nlp models. *arXiv preprint arXiv:2110.08875*.

Anh T. Tran, Cuong V. Nguyen, and Tal Hassner. 2019. Transferability and hardness of supervised classification tasks.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

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

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of EMNLP 2019*, pages 3685–3690.

Zihuiwen Ye, Pengfei Liu, Jinlan Fu, and Graham Neubig. 2021. Towards more fine-grained and reliable NLP performance prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3703–3714, Online. Association for Computational Linguistics.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. 2021. Logme: Practical assessment of pre-trained models for transfer learning.

Ming Yuan and Yi Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

A Appendix

Table 3 contains the information about the tasks considered in the survey for Section 2. The language-wise errors for tasks other than TyDiQA-GoldP can be found in Figure 5.

A.1 Training Details

Performance Prediction Models

1. XGBoost: For training XGBoost regressor for the performance prediction, we use 100 estimators with a maximum depth of 10 and a learning rate of 0.1.

2. Group Lasso: We use a regularization strength of 0.005 for the l_1/l_2 norm term in the objective function, and use the implementation provided in the MuTaR software package [5].

Translate Baseline: We use the Azure Translator[6] to translate the data in pivot language to target languages. For classification tasks XNLI and PAWS-X, the labels can be directly transferred across the translations. For QA tasks XQUAD and TyDiQA we use the approach described in Hu et al. (2020) to obtain the answer span in the translated test which involves enclosing the answer span in the original text within <b> </b> tags to recover the answer in the translation.

A.2 Features Description

1. Pre-training Size of a Language: The amount of data in a language l that was used to pre-train the MMLM.

2. Tokenizer Quality: We use the two metrics defined by Rust et al. (2021) to measure the quality of a multilingual tokenizer on a target language t. The first metric is **Fertility** which is equal to the average number of sub-words produced per tokenized word and the other is **Percentage Continued Words** which measures how often the tokenizer chooses to continue a word across at least two tokens.

3. Subword Overlap: The subword overlap between a pivot and target language is defined as the fraction of sub-words that are common in the vocabulary of the two languages. Let V_p and V_t be the subword vocabularies of p and t. The subword overlap is then defined as :

$$o_{sw}(p,t) = \frac{|V_p \cap V_t|}{|V_p \cup V_t|} \tag{5}$$

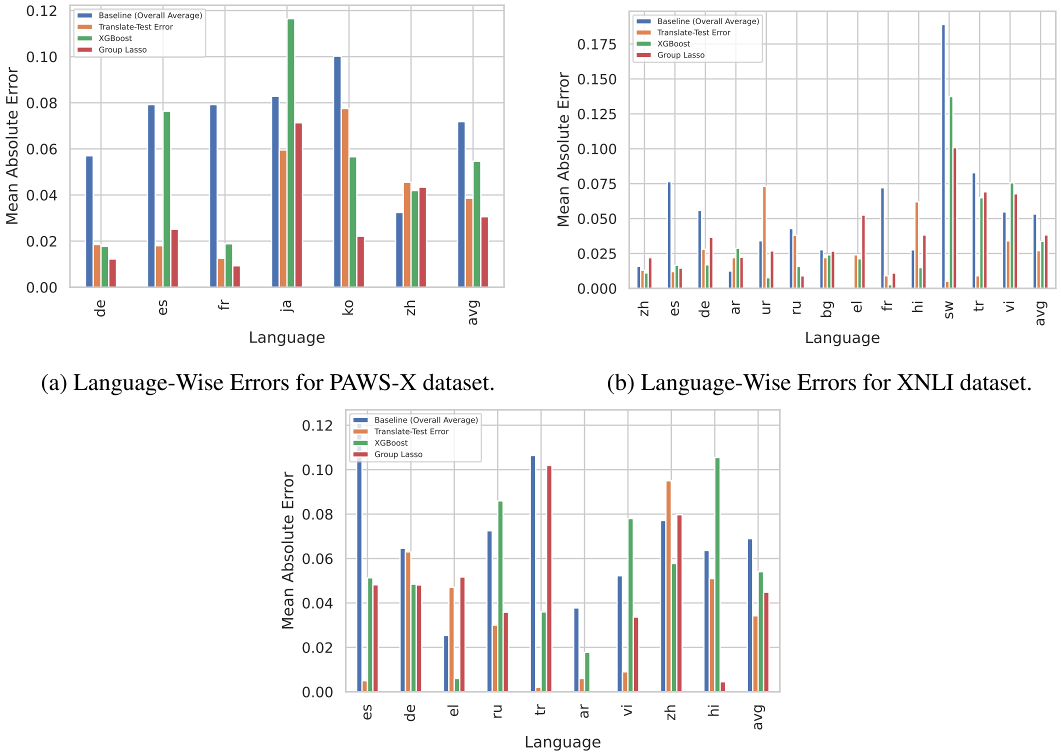
4. Relatedness between Lang2Vec features: Following Lin et al. (2019) and Lauscher et al. (2020), we compute the typological relatedness between p and t from the linguistic features provided by the URIEL project (Littell et al., 2017). We use syntactic ($s_{syn}(p,t)$), phonological similarity ($s_{pho}(p,t)$), genetic similarity ($s_{gen}(p,t)$) and geographic distance ($d_{geo}(p,t)$). For details, please see Littell et al. (2017)

[5] https://github.com/hichamjanati/mutar

[6] https://azure.microsoft.com/en-us/services/cognitive- services/translator/

	Type	Release Year	Number of Languages	Number of Language Families
UDPOS	Structure Prediction	2015	57	13
WikiANN	Structure Prediction	2017	100	15
XNLI	Classification	2018	15	7
XCOPA	Classification	2020	10	10
XQUAD	Question Answering	2020	11	6
MLQA	Question Answering	2020	7	4
TyDiQA	Question Answering	2020	11	9
MewsliX	Retrieval	2020	11	5
LAReQA	Retrieval	2020	11	6
PAWSX	Sentence Classification	2019	7	4
BUCC	Retrieval	2016	4	2
MLDoc	Classification	2018	8	3
QALD-9	Question Answering	2022	9	2
xSID	Classification	2021	11	6
WikiNEuRal	Structure Prediction	2021	8	1
WikiLingua	Summarization	2020	18	9
XL-BEL	Retrieval	2021	10	7
Tatoeba	Retrieval	2019	73	14

Table 3: The list of tasks surveyed for the discussion in Section 2.



(a) Language-Wise Errors for PAWS-X dataset.

(b) Language-Wise Errors for XNLI dataset.

(c) Language-Wise Errors for XQUAD dataset.

Figure 5