Predicting Machine Translation Performance on Low-Resource Languages: The Role of Domain Similarity

Eric Khiu[*], **Hasti Toossi**[†], **David Anugraha**[†], **Jinyu Liu**[†], **Jiaxu Li**[†], **Juan Armando Parra Flores**[¶], **Leandro Arcos Roman**[§], **A. Seza Doğruöz**[#], **En-Shiun Annie Lee**[†,‡]

[*] University of Michigan, USA [†] University of Toronto, Canada

[¶] Centro de Investigación en Matemáticas, Mexico [§] Amherst College, USA

[#] LT3, IDLab, Universiteit Gent, Belgium [‡] Ontario Tech University, Canada

khiu.eric@gmail.com as.dogruoz@ugent.be annie.lee@cs.toronto.edu

Abstract

Fine-tuning and testing a multilingual large language model is expensive and challenging for low-resource languages (LRLs). While previous studies have predicted the performance of natural language processing (NLP) tasks using machine learning methods, they primarily focus on high-resource languages, overlooking LRLs and shifts across domains. Focusing on LRLs, we investigate three factors: the size of the fine-tuning corpus, the domain similarity between fine-tuning and testing corpora, and the language similarity between source and target languages. We employ classical regression models to assess how these factors impact the model's performance. Our results indicate that domain similarity has the most critical impact on predicting the performance of Machine Translation models.

1 Introduction

Fine-tuning large language models for natural language processing (NLP) tasks across varying languages, tasks, and domains is a resource-intensive and environmentally harmful process. (Xia et al., 2020). This challenge is especially magnified for low-resource languages (LRLs). However, knowing how well a language model performs on a particular language can be useful information, such as improving the accuracy of quality estimation (QE) models (Zouhar et al., 2023). Therefore, there is a need to estimate the performance of these models for LRLs without conducting time-consuming and computationally expensive model pre-training and fine-tuning.

Existing approaches for predicting the performance of models for NLP tasks have shown promise using linear regression and gradient-boosting trees (Birch et al., 2008; Xia et al., 2020; Srinivasan et al., 2021; Ye et al., 2021). These studies have considered data size, typological features, and language similarity as factors contributing to the model performance. However, most of these studies are conducted for high-resource languages (HRLs) (e.g., Romance and Germanic families) thus limiting their applicability to LRLs. Furthermore, performance drops in NLP tasks have been observed due to domain shift (Elsahar and Gallé, 2019). However, this factor is not explicitly considered in the existing works that predict the performance of language models.

Based on the aforementioned limitations in the literature, we considered three factors for the Machine Translation (MT) performance prediction for LRLs using classical regression models. These factors are the size of the fine-tuning corpus, the domain similarity between fine-tuning and testing corpora, and the language similarity between source and target languages.

Then, we tested the statistical reliability of these regression models and evaluated them based on their prediction accuracy. We selected those with relatively high accuracy for each factor and explored how data partitioning (described in § 2) affects the quality of fit using these preferred models. Additionally, we analyzed the importance of the factors by ranking them based on their correlation with the MT performance, their weights in multifactor regression models, and their importance in multifactor models using the Random Forest Regressor.

Our contributions are as follows: 1) we developed a statistically rigorous method for performance prediction that can be repeated on any combination of LRLs, NLP tasks, and LLMs; 2) we specifically evaluated the impact of various factors on the performance of MT models; 3) we provided domain-specific and language-specific interpretations based on the performance of the regression models.

Findings of the Association for Computational Linguistics: EACL 2024, pages 1474–1486

2 Model and Data

Our data is collected from experiments of a prior study (Nayak et al., 2023) on fine-tuning and testing different corpora and target languages using the multilingual large language model mBART (Table 1). Each experiment consists of performance measured by spBLEU, with the source language (always English (EN)), the target language, l, the fine-tuning corpus, t and its size, s, and the testing corpus, τ.

Language Model and Evaluation Metric mBART is a pre-trained multilingual sequence-to-sequence model that is built based on the encoder-decoder Transformer architecture (Vaswani et al., 2017). Lee et al. (2022) has shown that mBART outperforms mT5, another multilingual large language model, especially on LRLs. Lee et al. (2022) also suggested the use of spBLEU as the evaluation metric for LRLs because it is a sentence-level metric that is more robust to the lack of reference translations than corpus-level metrics like BLEU. Although the size has been found to impact model loss rather than performance, Ghorbani et al. (2021) has demonstrated a negative linear relationship between performance and model loss.

Languages We covered five South Asian languages that are all considered low-resource other than Hindi (HI) (Joshi et al., 2020), (Table 2)[1]; Sinhala (SI) and Tamil TA are the official languages of Sri Lanka and Hindi (SI), Gujarati (GU), and Kannada (KA) are three of the many official languages of India. Kannada (KA) is unseen during mBART's pre-training. Note that we only considered the EN-XX direction because it often performs better than the XX-EN direction (Johnson et al., 2017; Lee et al., 2022). This mitigates our regression models from skewing excessively toward the low spBLEU extreme.

Corpora We had two fine-tuning corpora for each language. The first fine-tuning corpus is either an administrative (*Government*; SI,TA) or a news (PMIndia; HI, GU, KA) corpus. The second fine-tuning corpus is a religious (*Bible*) corpus. Due to limited availability, we scrapped the Bible corpus for SI from a different website[2]. For testing corpora, on top of the administrative/ news corpus and religious corpus, we also had an open-domain corpus (FLORES). Also due to limited availability, we used a slightly different corpus, FLORES-V1 instead of FLORES-101 for SI. For complete details of the corpora, see Appendix A.1). We define the experiments where the fine-tuning and testing corpora are from the same domain as *in-domain* experiments, and *out-domain* otherwise. To ensure that MT systems perform consistently across corpora of varying sizes, we extracted fixed-size fine-tuning sets from each corpus as in Table 1, based on the available amount of parallel text that we could sample from. All testing corpora are about 1k tokens.

Data Partitioning In our modeling, we split our data by grouping them according to their experimental settings (fine-tuning corpus, testing corpus, target language). We refer these groups of experiments as *partitions*. For instance, the "KA partition" refers to the first three columns in Table 1, while the "Fined-tuned-on-Bible partition" refers to the last three rows in Table 1. We refer the ways of partitioning the data as *partitioning schemes*, which differs by the factor that we model, as in Table 4.[3]

3 Factors and Featurization of Factors

We consider three potential factors that impact the performance score of the MT models: 1) the size of fine-tuning corpus, 2) the domain similarity between fine-tuning and testing corpora, and 3) the language similarity between source and target language. We represent these factors as feature variable(s) used as predictor(s) in the regression models described in the next section. These predictors are: ϕ_s = size feature variable; ϕ_d = domain feature variable; ϕ_l = language feature variable.

3.1 Fine-Tuning Corpus Size

It has been observed that the cross-entropy loss of MT models behaves as a power-law with respect to the amount of fine-tuning data (Gordon et al., 2021; Ghorbani et al., 2021; Kaplan et al., 2020). This suggests that the size of fine-tuning corpora is an important factor to consider in our study. We define the size factor, denoted as $\phi_s = \tilde{s}$, as the normalized count of sentence pairs in the fine-tuning

[1]The classification in Joshi et al. (2020) is outdated. (SI) must be at least Joshi's class 3 because it is used to train mBART. According to their definitions, all the languages in our study fall are at least class 2.

[2]Sinhala: https://www.wordproject.org/bibles/si/index.htm; and others: https://ebible.org/download.php

[3]Partitions with less than 10 data points are too small and thus not discussed.

Fine-Tuning Corpus	Size	Target Language and Testing Corpus														
		Kannada (KA)			Gujarati (GU)			Hindi (HI)			Sinhala (SI)			Tamil (TA)		
		FLORES	Bible	PMI	FLORES	Bible	PMI	FLORES	Bible	PMI	FLORES*	Bible†	Gov	FLORES	Bible	Gov
Gov/PMI	1k	2.2	0.3	12.0	7.8	2.3	22.6	6.6	1.0	19.7	3.8	0.2	21.7	2.6	0.3	19.7
	10k	11.8	1.5	30.7	16.6	4.0	34.2	14.5	3.0	32.4	9.2	0.9	41.7	7.1	0.8	34.8
	25k	14.2	1.7	34.3	19.9	4.8	37.9	17.0	3.5	35.5	11.3	1.2	47.0	9.0	1.3	38.2
	50k	NA	NA	NA	NA	NA	NA	19.0	3.4	36.7	12.3	1.5	49.5	11.3	1.6	40.8
Bible	1k	0.5	12.3	0.3	2.2	12.9	1.8	1.5	18.6	1.0	0.8	21.6	0.4	0.8	16.3	0.3
	10k	1.8	24.0	0.8	4.1	23.9	2.6	2.5	28.1	1.8	1.7	34.2	0.8	1.6	26.9	0.7
	25k	2.2	28.1	1.0	4.2	28.5	2.9	2.8	32.3	1.8	1.9	38.5	0.9	2.0	31.4	0.8

Table 1: MT Performance in spBLEU by fine-tuning mBART on different combinations of fine-tuning corpus, size of fine-tuning corpus, target language, and testing corpus.
* We used FLORES-V1 instead of FLROES-101 for SI due to availability.
† The bible corpus for SI is scrapped from a different website due to availability.

Language	Family	Script	Joshi Class	mBART Token	d_{geo}	d_{gen}	d_{syn}	d_{pho}	d_{inv}	d_{fea}
Kannada (KA)	Dravidian	Kannada	1	-	0.40	1.00	0.64	0.35	0.47	0.50
Gujarati (GU)	Indo Aryan	Gujarati	1	140M	0.30	0.90	0.68	0.57	0.48	0.60
Hindi (HI)	Indo Aryan	Devanagari	4	1715M	0.40	0.90	0.59	0.34	0.47	0.50
Sinhala (SI)	Indo Aryan	Sinhala	1	243M	0.40	0.90	0.78	0.41	0.50	0.60
Tamil (TA)	Dravidian	Tamil	3	595M	0.40	1.00	0.71	0.57	0.50	0.60

Table 2: Properties about the languages in our study and their lang2vec distances from English.

corpus. We achieve this normalization by employing a minimum-maximum scaling method, which constrains it to a range of $0 \leq \tilde{s} \leq 1$. This standardization aligns with the normalization applied to other features in our study.

3.2 Domain similarity

It has been discovered that the performance of language models faces significant drops when they encounter unfamiliar vocabulary and writing style (Blitzer, 2008; Jia and Liang, 2017; Calapodescu et al., 2019; Elsahar and Gallé, 2019). We refer to this situation as *domain shift* where *domain* is a "distribution over language characterizing a given topic or genre" (Gururangan et al., 2020). In our case, domain shift happens when the testing corpus is from a domain different from the fine-tuning corpus. This motivates us to consider domain similarity between fine-tuning and testing corpora as one factor affecting the performance of MT models.

Previous studies have proposed various methods to measure and mitigate domain divergence in MT models (Kashyap et al., 2021; Pillutla et al., 2021; Nayak et al., 2023; Lee et al., 2022). Kashyap et al. (2021) showed that information-theoretic measures such as Kullback–Leibler (KL) divergence, Jensen–Shannon divergence (JSD), and higher-order domain discriminator (e.g., Proxy A-distance (PAD)) capture good correlation with performance drop of MT models. Our study favors entropy methods, particularly JSD over KL divergence and PAD, for its symmetric property and relative simplicity. We refer to the domain feature, ϕ_d, as the JSD between fine-tuning and testing corpora, that is, $\phi_d = j = JSD(t, \tau)$. (see Appendix A.2 for complete details on JSD calculation).

3.3 Language similarity

Language similarity between source and target languages is important in translating from one language to another because it can help to leverage the cross-lingual transfer and multilinguality of the language model while exploiting parallel data from related language pairs (Lee, 2022; Gaschi et al., 2023; Philippy et al., 2023). This can be particularly promising for LRLs with insufficient quantities of high-quality parallel data (Goyal et al., 2020).

To measure language similarity, we utilize six distance features queried from URIEL Typological Database using lang2vec (Littell et al., 2017). The distance features are geographical distance, d_{geo}, genetic distance, d_{gen}, syntatic distance, d_{syn}, phonological distance, d_{pho}, inventory distance, d_{inv}, and featural distance, d_{fea} (Table 2, see Appendix A.3 for details). In our study, we refer to the language feature, ϕ_l, as any combination of the

six distance features.

4 Methodology

In this section, we outline our methodology for modeling and evaluating spBLEU predictions using factors mentioned previously, including the exploration of different regression models and their statistical reliability. We also examine the importance of individual features through correlation and feature importance analyses.

4.1 Modeling and Evaluation

Each model is defined by a predictor function f, which predicts a spBLEU value given a feature value x or a vector of feature values $\mathbf{x} = [x_1, ..., x_n]^\top$ of an experiment. Table 3 catalogues the predictor functions employed. Our selection includes straightforward mathematical functions such as linear, polynomial, and logarithmic types. This choice is grounded in the exploratory nature of our research and the classic use of these functions in regression analysis. It is important to note that in polynomial regressions, interaction variables (for instance, $x_i x_j, i \neq j$) are omitted in multifactor models. This exclusion is deliberate, as it allows us to focus on the impact of individual factors. The intricate interdependencies among these factors are comprehensively addressed through weight analysis (see § 4.3) in the multifactor linear regression model.

Name	Definition
Linear	$f_{\text{lnr}}(\mathbf{x}) = \beta_0 + \sum_j \beta_j x_j$
Quadratic	$f_{\text{poly}_2}(\mathbf{x}) = \beta_0 + \sum_j \left[\beta_{1j} x_j + \beta_{2j} x_j^2\right]$
Cubic	$f_{\text{poly}_3}(\mathbf{x}) = \beta_0 + \sum_j \left[\beta_{1j} x_j + \beta_{2j} x_j^2 + \beta_{3j} x_j^3\right]$
Logarithmic	$f_{\log}(\mathbf{x}) = \beta_0 + \sum_j \beta_j \log x_j$
Scaling Law	$f_{\text{SL}}(\tilde{s}) = \beta_0 (\tilde{s}^{-1} + \beta_1)^{\beta_2}$ (only used for size)

Table 3: The predictor functions explored in our study.

In order to understand the impact of individual factors, we explored predictor functions with one factor at a time as an input variable[4]. In addition, data partitioning mentioned in § 2 allowed us to minimize differences between experiments, except for the modeled factor. This approach provides insights into the relationships between individual factors and experimental settings.

[4]Specifically for size, scaling law was used as an additional predictor function as scaling law as supported by multiple studies (Gordon et al., 2021; Ghorbani et al., 2021; Kaplan et al., 2020).

For further exploration, the same predictor functions were explored using multiple features as multi-factor input variables. This approach allows for a more robust predictor function that captures the interactions between multiple factors, which had been postulated from the partitioning in single-factor modeling. The investigated multi-factor combinations included size and JSD, all six language features, and size, JSD, and all six language features.

To evaluate the prediction accuracy of our regression models, we used root-mean-square error (RMSE) as a metric for ranking models. The RMSE was determined by averaging the RMSE values obtained from each partition's k-fold cross-validation folds ($k = 10$).

4.2 Statistical Assessment on Regression Residuals

Residuals reflect the discrepancy between our model's predicted spBLEU and the true spBLEU for any given experiment. Residuals can provide a quantitative measure of our model's accuracy and how our model's predictions deviate from the true spBLEU, offering insights on any issues with the model's robustness and overall reliability. We verified two model assumptions described in Bates and Watts (1988), namely, normality and homoscedasticity of residuals. The normality of residuals is verified using D'Agnostino-Pearson test (Pearson et al., 1977), whereas the homoscedasticity is observed from the plots.

4.3 Ranking Feature Importance

To assess the correlation between each feature and spBLEU as well as their importance as predictors in our regression models, we ranked the features by the following three analyses:

(I) Pearson's Correlation Analysis To measure the strength and direction of the linear relationship between each feature and spBLEU, we calculated the Pearson Correlation coefficient along with the statistical significance p-value for the correlation.

(II) Weight analysis In addition to pairwise relationships measured by Pearson's Correlation Analysis, we also analyzed the unique contribution of each feature while considering the interdependencies among them by ranking the features by their weight in the multifactor linear regression model.

(III) Random Forest To assess the importance of each factor in our modeling using various regression models, we used Random Forest to identify the most important features in the multifactor models. See Appendix B for optimal hyperparameters settings used in our study.

5 Results

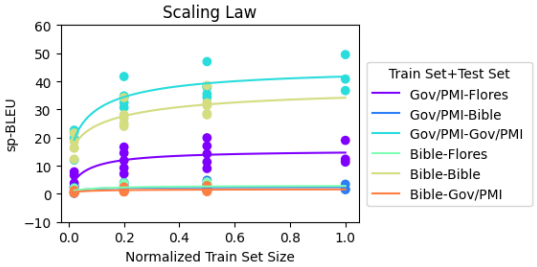
In this section, we discuss the performance of our regression models based on their RMSE in k-fold cross-validation (Table 4). In § 5.1, we extensively discuss the regression models that work well, along with their statistical reliability. Then, in § 5.2, we analyze the residuals' distribution of those models on specific partitions and provide our domain-specific and language-specific interpretations of the observations. Lastly, in § 5.3, we compare the correlation between each feature and spBLEU, as well as their importance in multifactor models, which gives us insights into the impact of various factors on the performance of MT models.

5.1 Prediction Accuracy of Factors

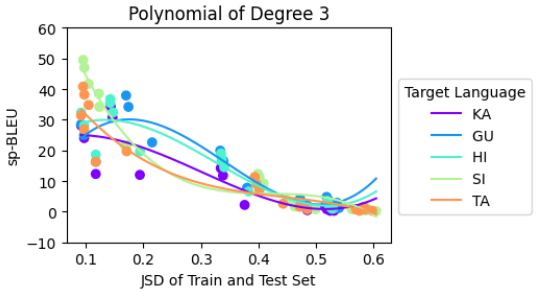
To explore the impact of each factor on spBLEU, we performed regression based on subsets of factors. The prediction accuracy of each regression model was measured in RMSE from k-fold cross-validation.

Regression using size feature In the case of predictor functions that take the size feature as a predictor, we observed that the partitioning scheme has a more significant impact on the RMSE than the predictor functions. For instance, the RMSE is significantly lower when partitioning by fine-tuning and testing corpora (Table 4). Such a trend could be attributed to the concentration of data points when mBART is tested in-domain and out-domain (Figure 1a). Consequently, separating the in-domain and out-domain experiments (i.e., partitioning by both fine-tuning and testing corpora) results in a notably lower RMSE. On the best partitioning scheme, the scaling law model has the lowest RMSE (Figure 1a, RMSE = 2.2998). This result is consistent with the current literature, which asserts that encoder-decoder Transformers used for MT exhibit a scaling law relationship between the volume of training data and model performance. (Gordon et al., 2021; Ghorbani et al., 2021; Kaplan et al., 2020).

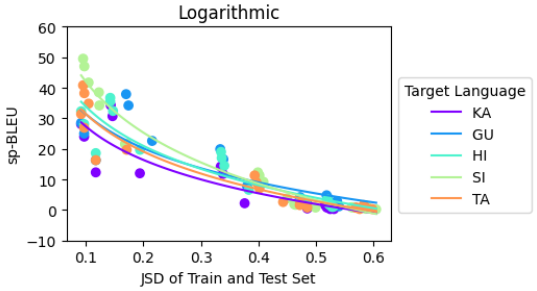
When modeling with scaling law, the residuals follow normal distribution on all partitions, as in Table 5a. However, the model is heteroscedastic for partitions involving the Bible corpus that are out-domain. This suggests that translation involving out-of-domain data (particularly Bible corpus) may exhibit highly variable performance. Consequently, it implies that the Bible corpus is better suited for the in-domain corpora rather than out-domain corpora.



(a) Regression plot using scaling law on size, $f_{\text{SL}}(\tilde{s})$; partitioned by both fine-tuning and testing corpora.



(b) Regression using polynomial (deg 3) regression on JSD, $f_{\text{poly}_3}(j)$; partitioned by language.



(c) Regression plot using logarithmic regression on JSD, $f_{\log}(j)$; partitioned by language.

Figure 1: Regression plots using best predictor functions for size and domain on best partitioning schemes.

Regression using domain similarity For predictor functions that take JSD as the predictor, polynomial regression with degree 3 has the lowest RMSE (Figure 1b, RMSE = 4.1202). Since polynomial regression models have a higher chance of being overfitted as their degree increases, we also consider the best performing non-polynomial model using JSD, i.e., the logarithmic regression model (Figure 1c, RMSE = 4.9355). Regarding their sta-

Predictor Function	Feature Variable(s)* and partitioning scheme								
	ϕ_s only					ϕ_d only		ϕ_s, ϕ_d	ϕ_s, ϕ_d, ϕ_l
	None	Fine-tune	Test	Lang	Fine-tune, test	None	Lang	None	None
Linear	13.2388	12.9270	11.1404	13.0014	2.9682	5.6433	5.0782	4.8766	4.5786
Polynomial-2	13.2092	12.8183	11.1218	13.0414	2.4561	5.4633	4.5698	4.6604	4.3840
Polynomial-3	13.1706	12.7914	22.4824	13.0601	2.3335	**5.4141**	**4.1202**	**4.4509**	**4.2168**
Logarithmic	13.1543	12.7835	11.3084	**12.8578**	2.3077	5.6315	4.9247	4.9502	4.6815
Scaling Law	13.1541	**12.7828**	11.1960	12.8929	**2.2998**	NA	NA	NA	NA

Table 4: Average Error Measurement† for Various Prediction Methods and Schemes.
* Feature variable(s) used as predictor(s) in the regression models: ϕ_s = size feature variable; ϕ_d = domain feature variable; ϕ_l = language feature variable.
† Measured by average RMSE from k-fold cross validation: **Bold** = function with lowest RMSE on this combination of feature variable(s) and partitioning scheme; underline = partitioning scheme with lowest RMSE using this combination of feature variable(s) and predictor function.

tistical reliability, the polynomial regression with degree 3 failed normality test on HI partition while the logarithmic regression failed normality test on TA partition, suggesting specific transformation per language on JSD is needed, otherwise more datapoints is required for the above to ensure model reliability.

We also noticed that models with size as the predictor have higher RMSE than those with JSD as the predictor. This difference can be attributed to the fact that there are only four unique size values[5]. Unless we have small enough partitions that contain fewer data points for a fixed size value, for instance, in the fine-tuning-test partition, size as a factor will obtain a lower RMSE.

We also observed that partitioning by language does not lead to a significant improvement in RMSE of the models on either size or JSD. This indicates that there is no substantial difference in spBLEU when mBART is tested on various languages, which can be attributed to the limited diversity in our languages. Furthermore, this may suggest a weak correlation between language features and spBLEU as described in Table 6.

Regression using multiple factors We evaluated two additional regression models with multiple factors to examine how these factors interact with each other in predicting spBLEU scores. Table 4 includes RMSE of multifactor models with ϕ_s and ϕ_d as predictors, and multifactor models with ϕ_s, ϕ_d, and ϕ_l (all lang2vec distances in Table 2) as predictors.

Relative to single-factor models that take only ϕ_d without partitioning, we observed that including ϕ_s and ϕ_l does improve the RMSE. However, the improvement is insignificant, further suggesting the high importance of domain similarity in the prediction relative to other factors considered in this study.

[5]For future work, we are collecting more sample points using low-cost transformers.

5.2 Residuals by Partition

To observe how our models performs on different partitions, we created boxplots of residuals when modeling data on each partition using the predictor functions. Using the best predictor function for size (scaling law) with the best partitioning scheme (by both fine-tuning and testing corpora), we noticed that the mean and variance of the residuals were lower for out-domain partitions (gov-gov and bible-bible, Figure 2a). This suggests that our model predicts better for out-domain partitions, which could be explained by the difference in the range of raw spBLEU when mBART is tested on in-domain and out-domain experiments ($[6.5, 49.5]$ for in-domain, $[0.2, 19.9]$ for out-domain).

Figure 2b presents how well the scaling law works for different languages. We noticed that the SI partition has relatively high residual mean and variance, implying that the performance of mBART on Sinhala is harder to predict with respect to the size of the fine-tuning corpus. This could be due to the use of different versions of the Bible corpus and FLORES corpus for SI, resulting in a higher range of spBLEU in this partition ([0.2, 49.5], Table 1) and hence harder to predict. However, this phenomenon is not observed in Figure 2c when the feature variable is JSD. This implies that using JSD as the predictor yields a more stable prediction for SI because it is not affected by using different fine-tuning corpora.

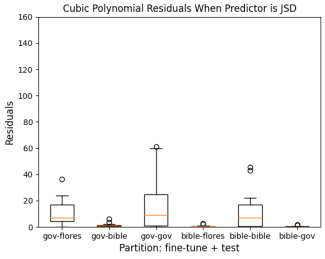
Fine-tuning – test	Normality	Homoscedastic?
bible-bible	0.3996	Yes
bible-FLORES	0.1380	**No**
bible-gov	0.2570	**No**
gov-bible	0.2534	**No**
gov-FLORES	0.2623	Yes
gov-gov	0.6127	No

(a) $f_{\text{SL}(\tilde{s})}$ on each train-test partition.

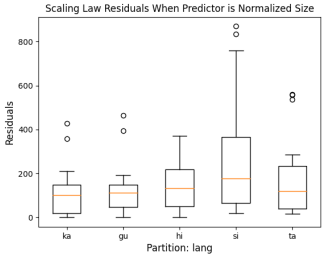
	$f_{\text{poly3}}(j)$		$f_{\log}(j)$	
Language	Normality	Homoscedastic?	Normality	Homoscedastic?
KA	0.1578	Yes	0.2155	Yes
GU	0.0563	Yes	0.2027	Yes
HI	**0.0129**	Yes	0.7290	Yes
SI	0.6021	Yes	0.2702	Yes
TA	0.0500	Yes	**0.0299**	Yes

(b) $f_{\text{poly3}}(j)$ and $f_{\log}(j)$ for each language partition.

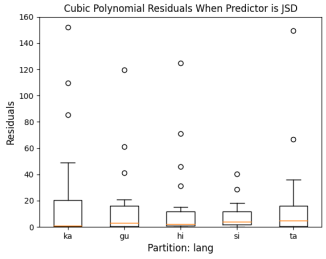
Table 5: Statistical Assessment on Normality and Homoscedasticity for size and JSD on best partitioning schmes respectively. For normality, **bold** = residuals are not normally distributed ($p < 0.05$).



(a) Residuals from $f_{\text{SL}}(\tilde{s})$; partitioned by fine-tuning and testing corpora.



(b) Residuals from $f_{\text{SL}}(\tilde{s})$; partitioned by language.



(c) Residuals from $f_{\text{poly}_3}(j)$; partitioned by language.

Figure 2: Boxplots of residuals using best predictor functions for size and domain on some partitioning schemes.

5.3 Feature Rankings

In order to assess the impact of the features in predicting spBLEU, Table 6 provided Pearson correlation coefficient and the statistical significance measured in p-value. We also include weights for each feature in the best multifactor linear regression model computed and their feature importance based on the best-performing Random Forest Regressor.

In Pearson's Correlation Analysis ranking, JSD stands out with a strong and statistically significant correlation to spBLEU (Table 6), suggesting a strong linear relationship between JSD and spBLEU. It also ranks highest in both weight analysis and Random Forest feature importance analysis, further illustrating its importance in predicting spBLEU (Table 6). This finding brings hope for developing a reliable model to understand the relationship between domain similarity and performance in MT tasks.

Surprisingly, all six language features show low correlations with spBLEU. The high similarity amongst our South Asian languages could be a factor, resulting in a similar distance from EN in Table 2. It suggests that the language features are not as significant as other features, like size and domain, for use as predictors in our regression models.

Feature Variable	Pearson Correlation Coefficient	Statistical Significance (p-value)	Weight Analysis	Random Forest (%)
j	-0.9176 [1]	8.47×10^{-71}	-68.5404 [1]	88.393 [1]
$\tilde{s}$	0.2468 [2]	0.0010	19.1317 [3=]	7.805 [2]
d_{gen}	-0.0863 [3]	0.2574	-25.7118 [2]	0.365 [5]
d_{syn}	0.0365 [4]	0.6325	3.6204 [7]	2.267 [3]
d_{inv}	0.0239 [6]	0.7542	13.0297 [5]	0.782 [4]
d_{fea}	0.0337 [5]	0.6585	19.1317 [3=]	0.079 [8]
d_{geo}	0.0025 [7]	0.9738	7.1308 [6]	0.147 [7]
d_{pho}	-0.0076 [8]	0.9104	-1.1780 [8]	0.161 [6]

Table 6: Feature importance rankings by Pearson's correlation analysis (along with its statistical significance), weight in linear regression model, and Random Forest feature importance analysis. Rankings in brackets.

6 Discussion

In this study, we revealed that domain similarity plays an important role in MT. In other words, it significantly affects the performance of MT models. All three feature rankings in § 5.3, as depicted in Table 6, underscore the significance of domain similarity in predicting spBLEU. The relationship between JSD and spBLEU is best modeled by polynomial regression of degree 3 in terms of k-fold RMSE, whereas the best non-polynomial model was logarithmic regression. Both models are relatively reliable in terms of the normality and homoscedasticity of the residuals.

Recognizing the importance of domain similar-

ity in MT, we also observed how it affects the predictability of spBLEU when modeling with the scaling law, which uses size as a predictor. The separation of in-domain and out-domain data improves the RMSE due to the distinct clustering of in-domain and out-domain data points. Additionally, we found that the performance of MT models on out-domain partitions is easier to predict. In other words, the prediction models are more confident that the spBLEU values are low when the range of spBLEU values is small. However, despite the lower variance in the residuals of the scaling law on out-domain partitions, the residuals exhibit heteroscedasticity in most of the out-domain partitions when using the scaling law for modeling.

Furthermore, the FLORES-v1 dataset for Sinhala includes data from OpenSubtitles, which are mainly transcripts of spoken data (Guzmán et al. (2019); Lison et al. (2018)). It should be noted that these transcripts may exhibit varying degrees of reliability, as they lack a control mechanism for verifying the translation accuracy. In addition, spoken Sinhala has different syntactical rules of written Sinhala (De Silva, 2019)), which means that there is variation in our Sinhala corpus (e.g., Bible and government documents corpora) as well. This would likely result in a lower translation score across FLORES-v1 and out-domain corpus. However, the JSD score can predict some of these differences in language caused by domain shift, similar to partitioning out by fine-tuning and test datasets. This explains why our model's predictive performance improved under these conditions.

Additionally, the Sri Lanka constitution states that "Sinhala shall be the language of administration and be used for the maintenance of public records and the transaction of all business" for most regions (Sri Lanka Const. art. XXII, § 1). Tamil, also an official language of Sri Lanka, would instead be translated. This difference in language choice could also explain why Sinhala outperforms Tamil in government-related in-domain documents and why domain similarity is such a powerful predictor in these cases.

Furthermore, we have detected heteroscedasticity in various models. For JSD, the data points will be heteroscedastic due to the inherent high domain divergence, resulting in experiments with very low spBLEU. In contrast, low domain divergence is highly variable, as other factors, such as language and fine-tuning set size, can impact the MT performance. The observation that JSD does not guarantee good model performance in single-factor regression motivates us to consider alternative techniques. The alternative techniques should be more robust or include additional variables to capture variations during low-JSD predictions. Additionally, we observed from the boxplots of residuals that residuals are skewed towards low spBLEU.

7 Conclusion

In our research, we conducted a comprehensive analysis focusing on three key factors (the size of the fine-tuning corpus, domain similarity between the fine-tuning and testing corpora, and the linguistic similarity between the source and target languages) affecting performance prediction of the MT for five South Asian languages. We find that domain similarity exerts the most significant influence on performance, surpassing even the impact of fine-tuning the corpus size. Additionally, the background of the corpora and language being translated emerged as a crucial factor in predicting performance and stability. Lastly, we verify that our approach to ascertain predictive factors for LRLs' performance is statistically rigorous. This approach enables performance prediction without the need for fine-tuning and testing resource-intensive and costly language models, ultimately fostering greater accessibility and equity for LRLs.

Limitations

The most prominent limitation of our study is the amount of data to fine-tune our regression models. As we observed that our models are generally biased towards experiments with low spBLEU and we could include more experiments with larger fine-tuning corpus size, or perhaps at constant interval between 1k and 100k tokens. There could also be a need to balance the amount of data from in-domain and out-domain.

The high degree of similarity between the languages in our data set rendered the effectiveness of language features from lang2vec as predictors. Due to the lack of LRL data in the URIEL library, lang2vec may not have sufficient data to provide approximation that accurately describe the LRL. Consequently, many languages might exhibit similar values for the same features, making it difficult to distinguish between them. This motivates us to consider incorporating experiments involving a

more diverse range of languages in future studies in order to thoroughly examine the impact of language similarity on MT. Additionally, apart from dataset-independent linguistic features, as suggested by Lin et al. (2019), we will explore dataset-dependent language features (e.g., Type-Token Ratio (TTR), word overlap, and subword overlap). Therefore, a more rigorous investigation into measuring language similarity is essential to identify suitable predictors for our task.

In addition, it is also important to consider additional factors that could potentially impact the performance of MT models, such as the use of pivot languages (Srinivasan et al., 2021) and the presence of noise (Gordon et al., 2021). Expanding our analysis to include data from different MT models and various evaluation metrics will help us assess the transferability of our prediction models across different MT models and evaluation metrics.

Acknowledgement

We extend our profound gratitude to the Fields Undergraduate Summer Research Program (FUSRP) for their invaluable support and the unique opportunity they provided for engaging in high-quality mathematical research. Our sincere thanks also go to Juan Armando Parra Flores and Leandro Arcos Roman, whose contributions through the FUSRP were instrumental in the success of our work.

Ethical Considerations

Equitability in Language Representation Given that our study revolves around LRLs, it is imperative to conscientiously acknowledge the imperative to foster equitable technological developments across varied linguistic communities. Our exploration into optimizing MT models for LRLs partially addresses this, but it's vital to consistently prioritize and amplify underrepresented languages in our future research and model development to prevent linguistic bias and facilitate digital inclusivity.

Data Bias and Representation Our regression models, as indicated in the limitations section, have potential biases towards experiments with low spBLEU, which may affect the robustness and fairness of our predictive models across various language datasets and use-cases. Ensuring unbiased and representative datasets is crucial not only for the accuracy of predictive models but also for avoiding the unintentional marginalization of certain linguistic features or dialects within the LRLs.

References

Douglas M. Bates and Donald G. Watts. 1988. *Nonlinear regression analysis and its applications*. Wiley series in probability and mathematical statistics. Applied probability and statistics. Wiley.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 745–754, Honolulu, Hawaii. Association for Computational Linguistics.

John Blitzer. 2008. *Domain adaptation of natural language processing systems*. Ph.D. thesis, University of Pennsylvania.

Ioan Calapodescu, Caroline Brun, Vassilina Nikoulina, and Salah Aït-Mokhtar. 2019. "sentiment aware map": exploration cartographique de points d'intérêt via l'analyse de sentiments au niveau des aspects (). In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume IV: Démonstrations*, pages 635–638.

Chris Collins and Richard Kayne. 2011. Syntactic structures of the world's languages.

Nisansa De Silva. 2019. Survey on publicly available sinhala natural language processing tools and research. *arXiv preprint arXiv:1906.02358*.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.3)*. Zenodo.

Hady Elsahar and Matthias Gallé. 2019. To annotate or not? predicting performance drop under domain shift. 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 2163–2173, Hong Kong, China. Association for Computational Linguistics.

Aloka Fernando, Surangika Ranathunga, and Gihan Dias. 2021. Data augmentation and terminology integration for domain-specific sinhala-english-tamil statistical machine translation.

Félix Gaschi, Patricio Cerda, Parisa Rastin, and Yannick Toussaint. 2023. Exploring the relationship between alignment and cross-lingual transfer in multilingual transformers. *arXiv preprint arXiv:2306.02790*.

Behrooz Ghorbani, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, and Colin Cherry. 2021. Scaling laws for neural machine translation.

Mitchell A Gordon, Kevin Duh, and Jared Kaplan. 2021. Data and parameter scaling laws for neural machine translation. In *Proceedings of the 2021 Conference*

on Empirical Methods in Natural Language Processing, pages 5915–5922, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Vikrant Goyal, Sourav Kumar, and Dipti Misra Sharma. 2020. Efficient neural machine translation for low-resource languages via exploiting related languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 162–168, Online. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. 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 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Barry Haddow and Faheem Kirefu. 2020. Pmindia – a collection of parallel corpora of languages of india.

Harald Hammarström, Robert Forkel, and Martin Haspelmath. 2018. Glottolog 3.0. *Max Planck Institute for the Science of Human History*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation.

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.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, and Roger Zimmermann. 2021. Domain divergences: a survey and empirical analysis.

En-Shiun Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Adelani, Ruisi Su, and Arya McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics.

En-Shiun Annie Lee. 2022. Improving translation capabilities of pre-trained multilingual sequence-to-sequence models for low-resource languages.

M. Paul Lewis, editor. 2009. *Ethnologue: Languages of the World, 16th editon*. SIL International.

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.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

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.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

Steven Moran, Daniel McCloy, and Richard Wright. 2014. Phoible online.

Shravan Nayak, Surangika Ranathunga, Sarubi Thillainathan, Rikki Hung, Anthony Rinaldi, Yining Wang, Jonah Mackey, Andrew Ho, and En-Shiun Annie Lee. 2023. Leveraging auxiliary domain parallel data in intermediate task fine-tuning for low-resource translation.

E. S. Pearson, R. B. D'Agostino, and K. O. Bowman. 1977. Tests for departure from normality: Comparison of powers. *Biometrika*, 64(2):231–246.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. *arXiv preprint arXiv:2305.16768*.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers.

Sri Lanka Const. art. XXII, § 1. Constitution of Sri Lanka (as amended in 2022). https://www.parliament.lk/files/pdf/constitution.pdf.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Mengzhou Xia, Antonios Anastasopoulos, Ruochen Xu, Yiming Yang, and Graham Neubigº. 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.

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.

Vilém Zouhar, Shehzaad Dhuliawala, Wangchunshu Zhou, Nico Daheim, Tom Kocmi, Yuchen Eleanor Jiang, and Mrinmaya Sachan. 2023. Poor man's quality estimation: Predicting reference-based MT metrics without the reference. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1311–1325, Dubrovnik, Croatia. Association for Computational Linguistics.

A Experimental Setup

A.1 Details of Corpora

Bible corpus (Bible)

The JHU Bible Corpus (McCarthy et al., 2020) contains Bible translations in over 1600 languages and serves as the only available parallel text for several low-resource languages. Due to the limited data available for our languages, we created a Bible corpus specifically for our experiments by scrapping Bible data from web[6] and aligned the sentences at verse level automatically. The resulting curated multi-way parallel corpus consists of 25k parallel sentences in KA, GU, HI, and TA. Note that SI was sourced from a different website, resulting in distinct content for this language.

FLORES corpus

FLORES-101 (Flores) (Goyal et al., 2022) is a corpus containing translations of English Wikipedia sentences into 101 different languages. The translations were done manually, and the corpus covers diverse topics and domains. For SI, we use FLORES-v1 (Guzmán et al., 2019) instead since it is not present in FLORES-101 .

Government corpus (Gov)

The government corpus (Gov) (Fernando et al., 2021) is a multi-way parallel corpus comprising Sinhala, Tamil, and English texts. The corpus is manually curated and includes data from various official Sri Lankan government sources, such as annual reports, committee reports, government institutional websites, procurement documents, and acts of the Parliament.

PMIndia corpus (PMI)

The PMIndia corpus (PMI) (Haddow and Kirefu, 2020) is a multi-way parallel corpus consisting of 13 Indian languages, along with English. The corpus has been curated from news updates taken from the Prime Minister of India's website.

A.2 Jensen-Shannon Divergence

Jensen-Shannon divergence (JSD) beteen two distributions P and Q is calculated using the formula

$$JSD(P||Q) = \frac{1}{2}KL(P||M) + \frac{1}{2}KL(Q||M)$$

[6]Sinhala: https://www.wordproject.org/bibles/si/index.htm; and others: https://ebible.org/download.php

where M is an equally weighted sum of the wo distributions and $KL(\cdot||\cdot)$ is the Kullback-Leibler divergence.

In preparation of this calculation, we first tokenized each corpus using the NLTK package[7], striped all stopwords, and transformed them into a (discrete) frequency distribution over all word tokens. Then, we convert all times and numbers into the tokens <TIME> and <NUMBER>, respectively. Finally, we compared the frequency distributions of each fine-tuning and test set using the formula above.

Note that JSD ranged from 0 to 1, with lower values indicating higher similarity between the two distributions.

A.3 Language Features

In this study, language feature refers to measures of similarity between two languages that are based on phylogenetic or typological properties established by linguistic study. The six language features from the URIEL database Littell et al. (2017) utilized in this study includes:

Geographic distance (d_{geo})

The orthodromic distance between the languages on the surface of the earth, divided by the antipodal distance. It is based primarily on language location descriptions in Glottolog (Hammarström et al., 2018).

Genetic distance (d_{gen})

The genealogical distance of the languages, derived from the hypothesized tree of language descent in Glottolog.

Inventory distance (d_{inv})

The cosine distance between the phonological feature vectors derived from the PHOIBLE database (Moran et al., 2014).

Syntactic distance (d_{syn})

The cosine distance between the syntactic structures feature vectors of the languages (Collins and Kayne, 2011), derived mostly from the WALS database (Dryer and Haspelmath, 2013).

[7]Documentation of NLTK package: https://www.nltk.org/

Phonological distance (d_{pho})

The cosine distance between the phonological feature vectors derived from the WALS and Ethnologue databases (Lewis, 2009).

Featural distance (d_{fea})

The cosine distance between feature vectors combining all 5 features mentioned above.

B Hyperparameters of Random Forest Regressor

We conducted grid search with k-fold cross-validation to find the optimal hyperparameter settings, including the number of decision trees in the ensemble (`n_estimators`), the maximum depth of each decision tree (`max_depth`), the minimum number of samples required to split an internal node (`min_samples_split`), the minimum number of samples required to be at a leaf node (`min_samples_leaf`), and whether bootstrap samples were used in building trees (`bootstrap`). The optimal hyperparameter settings are tabulated in Table 7, resulting in an RMSE of 3.29.

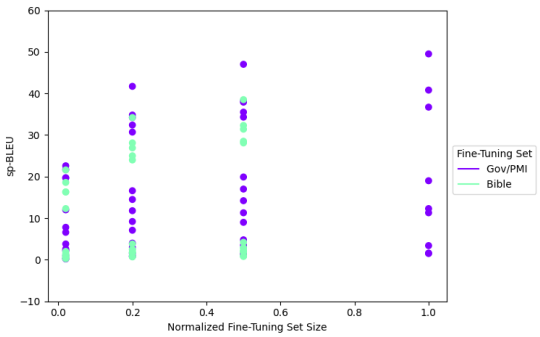
Hyperparameter	Values Searched	Optimal Setting
`n_estimators`	$\{n \mid n = 50 + 25k,\ 0 \le k \le 14\}$	100
`max_depth`	$\{n \mid n = 3 + 2k,\ 0 \le k \le 6\}$	9
`min_samples_split`	$\{2, 3, 4, 5\}$	1
`min_samples_leaf`	$\{1, 2, 3\}$	2
`bootstrap`	{TRUE, FALSE}	TRUE

Table 7: List of hyperparameters used in the optimization of the Random Forest Regressor using grid search.

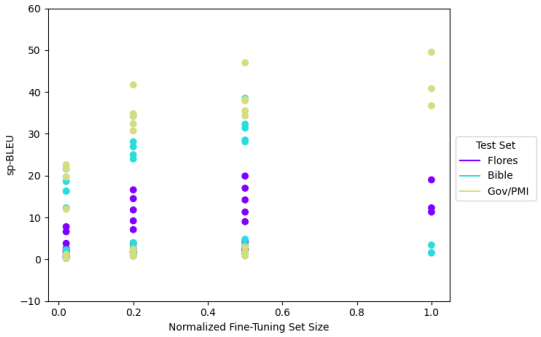
C Scatter Plots

In this section, we present the scatter plots of spBLEU with respect to size of fine-tuning corpora using different partitioning schemes.

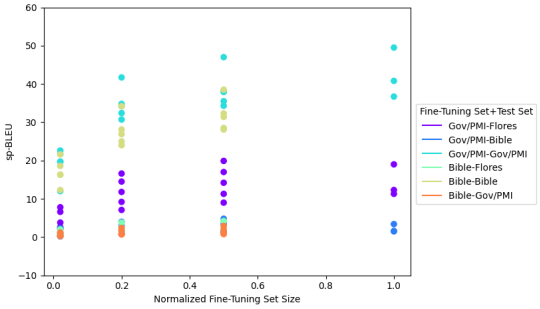
C.1 Factor = Size



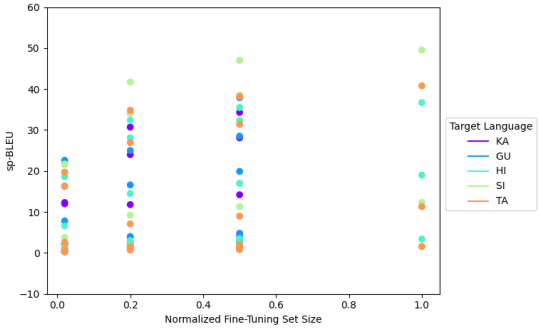
(a) Scatter Plot of spBLEU with respect to size, partitioned by fine-tuning corpora.



(b) Scatter Plot of spBLEU with respect to size of fine-tuning corpora, partitioned by testing corpora.



(c) Scatter Plot of spBLEU with respect to size, partitioned by both fine-tuning and testing corpora.



(d) Scatter Plot of spBLEU with respect to size, partitioned by target language.

Figure 3: Scatter Plots of spBLEU with respect to size using different partitioning schemes.

C.2 Factor = Domain Similarity

In this section, we present the scatter plot of spBLEU with respect to JSD, partitioned by target language.

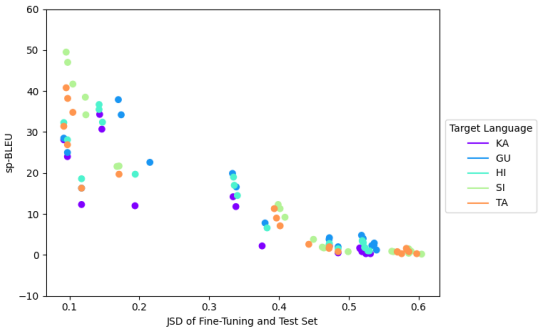


Figure 4: Scatter Plot of spBLEU with respect to JSD, partitioned by target language.