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, ID-Lab, Universiteit Gent, Belgium
[‡]Ontario Tech University, Canada

Outline

- Introduction
- Methodology
- Results and Discussion
- Conclusions

Outline

- Introduction
- Methodology
- Results and Discussion
- Conclusions

Problem Statement

- Background
  - Fine-tuning language models is expensive [1]
  - Even more challenging for low-resource languages (LRLs)
  - Performance data can be useful to optimize training cost or for other tasks (e.g., quality estimation (QE))
- Our goal: Model the performance of machine translation models **mathematically**
  - Some (potential) factors affecting the performance are measurable
  - Existing regression models [2, 3] exhibit promising capabilities, but:
    - Not much is done for LRLs
    - Varying in terms of statistical rigor

[1] 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.

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

[3] 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

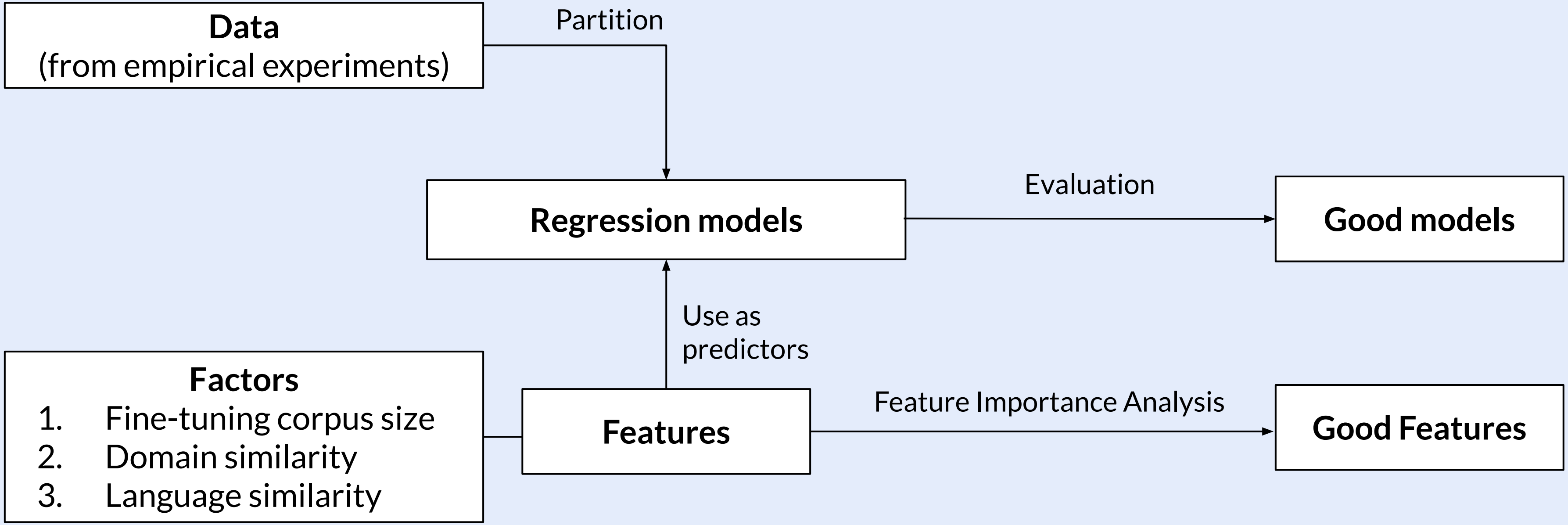
Research Questions

How do fine-tuning corpus size, domain similarity, and language similarity impact the performance of MT models in LRLs settings?

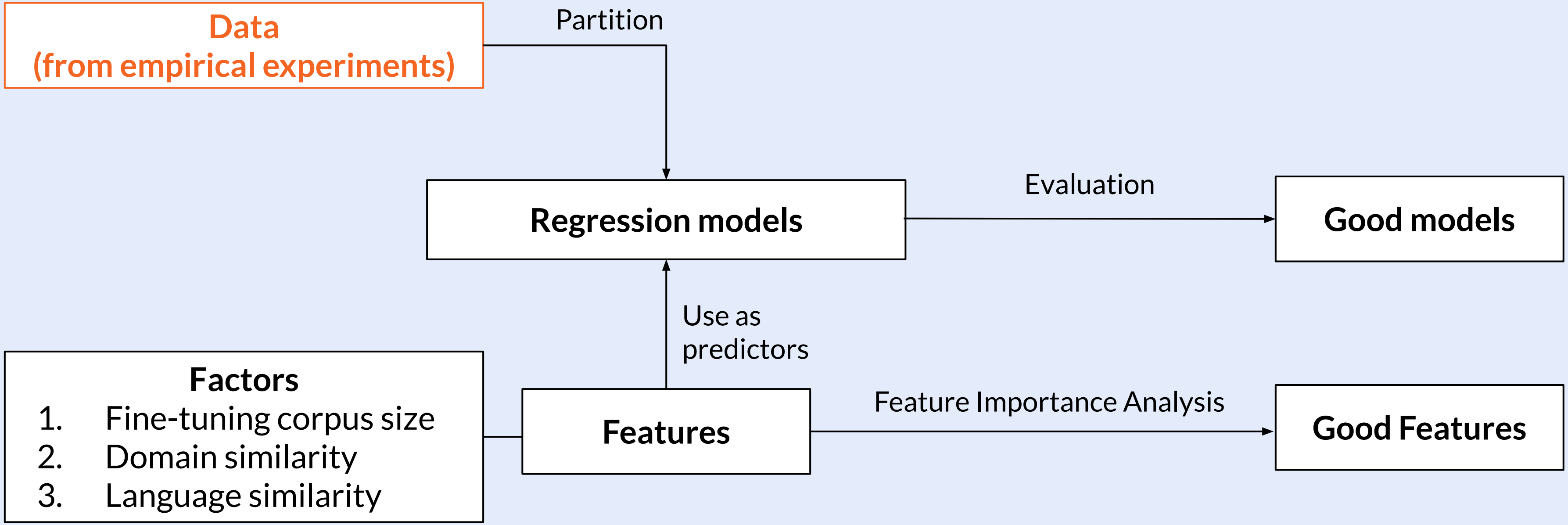
Outline

- Introduction
- Methodology
- Results and Discussion
- Conclusions

Methodology



Methodology



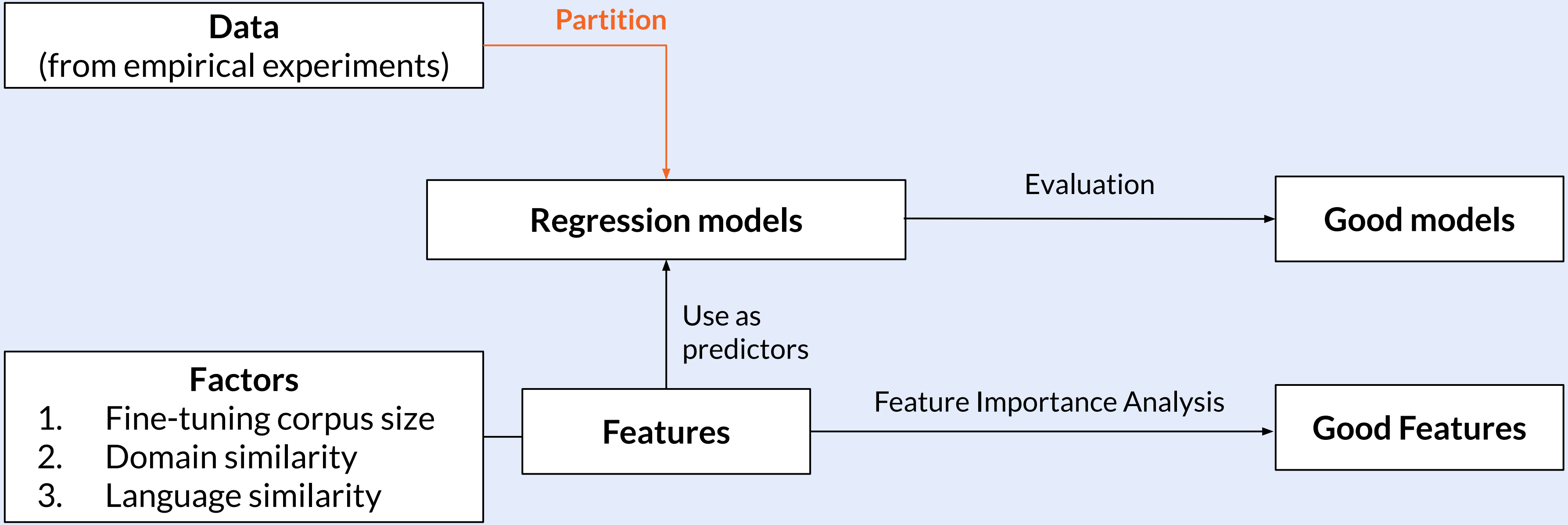
Data

- Empirical experiment data from Nayak et.al., 2023 [4]
- MT model: mBart
- Performance Metric: spBLEU

Fine-tune	Size	gu			hi			ka			si			ta		
		flores	bible	pmo	flores	bible	pmo	flores	bible	pmo	flores	bible	gov	flores	bible	gov
PMO/ Gov	1k	7.8	2.3	22.6	6.6	1.0	19.7	2.2	0.3	12.0	3.8	0.2	21.7	2.6	0.3	19.7
	10k	16.6	4.0	34.2	14.5	3.0	32.4	11.8	1.5	30.7	9.2	0.9	41.7	7.1	0.8	34.8
	25k	19.9	4.8	37.9	17.0	3.5	35.5	14.2	1.7	34.3	11.3	1.2	47.0	9.0	1.3	38.2
	50k				19.0	3.4	36.7				12.3	1.5	49.5	11.3	1.6	40.8
Bible	1k	2.0	16.3	1.2	1.5	18.6	1.0	0.5	12.3	0.3	0.8	21.6	0.4	0.8	16.3	0.3
	10k	3.8	25.0	2.4	2.5	28.1	1.8	1.8	24.0	0.8	1.7	34.2	0.8	1.6	26.9	0.7
	25k	4.2	28.5	2.9	2.8	32.3	1.8	2.2	28.1	1.0	1.9	38.5	0.9	2.0	31.4	0.8

[4] 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.

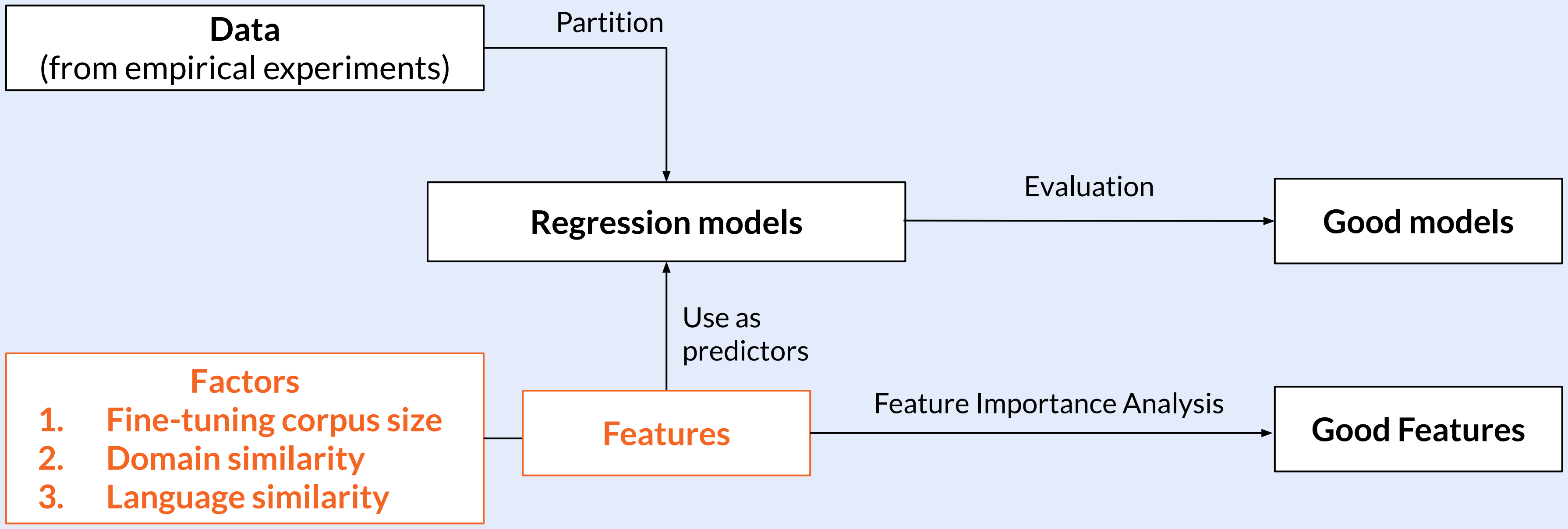
Methodology



Partitioning

- Group our data by experimental settings (fine-tuning corpus, testing corpus and target language)
- Each partition has its own coefficient values in the model
- The partitioning schemes differ by factors that we model

Methodology



Fine-tuning Corpus Size

- Rationale
  - Previous literatures [5, 6] suggest that the cross-entropy loss of MT models behaves as a power-law with respect to the amount of fine-tuning data
- Featurization:
  - We use the count of sentence pairs in fine-tuning corpora
  - Normalized using minimum-maximum scaling method

[5] 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.

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

Domain Similarity

- Rationale
  - Performance of LM drops when they encounter unfamiliar vocabulary and writing style
  - *Domain shift*: Testing corpus is from a different domain than the fine-tuning corpus
- Featurization
  - Kashyap et al. [7] showed that information-theoretic measures such as Kullback-Leibler (KL) divergence, Jensen-Shannon divergence (JSD) and higher-order discriminator (e.g., Proxy A-distance (PAD)) captures good correlation with performance drop
  - We use **JSD** for its symmetric property and relative simplicity

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

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

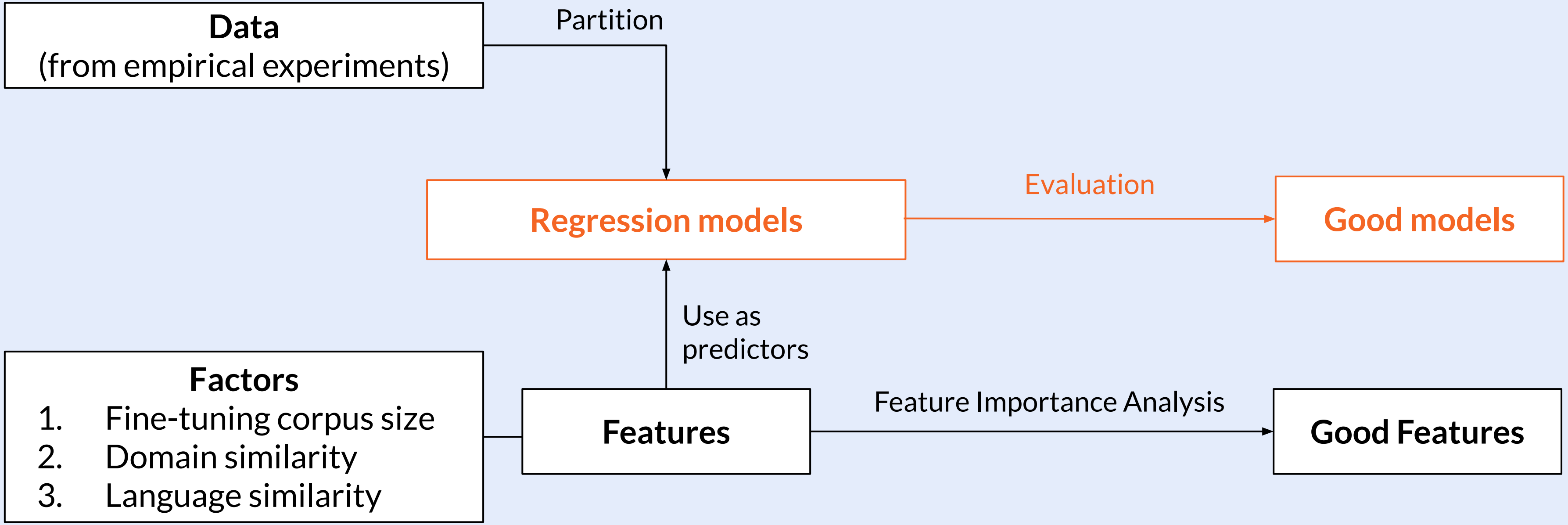
[7] Abhinav Ramesh Kashyap, Devamanyu Hazarika, MinYen Kan, and Roger Zimmermann. 2021. Domain divergences: a survey and empirical analysis.

Language Similarity

- Rationale
  - Language similarity can help to leverage cross-lingual transfer and multilinguality of the LM while exploiting parallel data from *related* language pairs
  - This can be promising for LRLs with insufficient high-quality parallel data
- Featurization
  - URIEL typological database [8], consisting of 6 distance measures across languages (geographical, genetic, syntactic, phonology, inventory, and featural distances)
  - We use the lang2vec to query URIEL for EN-XX language distances

[8] 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.

Methodology



Regression Models

- Performed regression analysis on each factors using classical mathematics functions

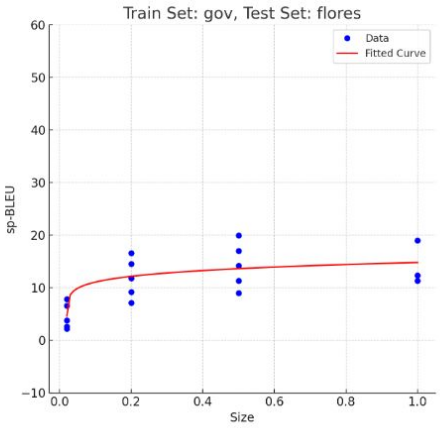
Name	Definition
Linear	$f_{\text{lnr}}(\mathbf{x}) = \beta_0 + \sum_j \beta_j x_j$
Quadratic	$f_{\text{poly}_2}(\mathbf{x}) = \beta_0 + \sum_j \left[\beta_{1j} x_j + \beta_{2j} x_j^2\right]$
Cubic	$f_{\text{poly}_3}(\mathbf{x}) = \beta_0 + \sum_j \left[\beta_{1j} x_j + \beta_{2j} x_j^2 + \beta_{3j} x_j^3\right]$
Logarithmic	$f_{\text{log}}(\mathbf{x}) = \beta_0 + \sum_j \beta_j \log x_j$
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.

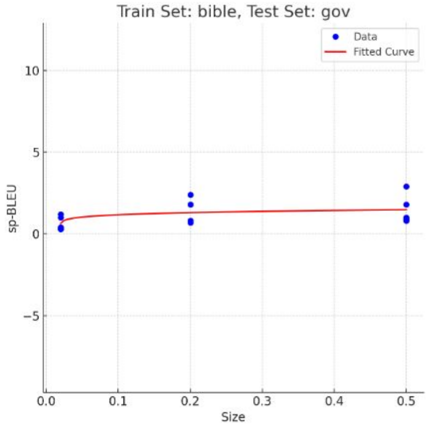
- Accuracy is measured using rooted mean square error (RMSE) over 10-fold cross validations

Statistical Assessment

- We assessed the statistical reliability of the regression models by normality and homoscedasticity of the residuals
  - **Normality** is assessed using D’Agnostino-Pearson test
  - **Homoscedasticity** is observed from the plots

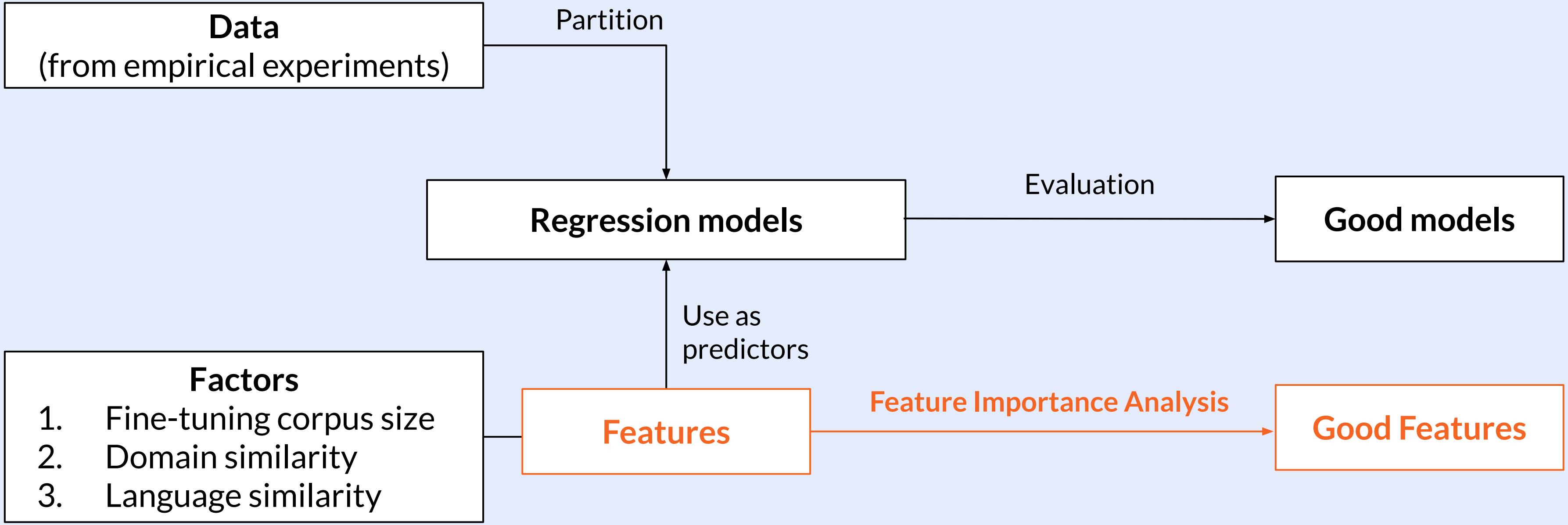


Example of homoscedastic (constant variance) model



Example of heteroscedastic (nonconstant variance) model

Methodology



Feature Importance Analysis

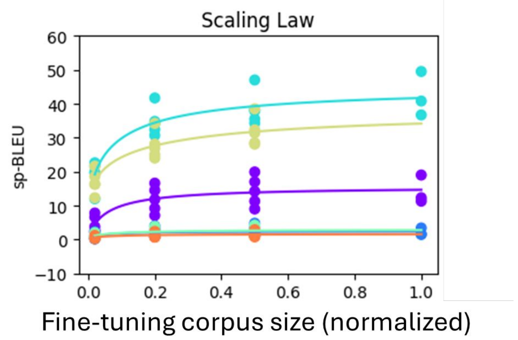
- **Pearson Correlation:** Measure strength of linear relationship
- **Weight Analysis:** Rank features by regression weight in multifactor linear model, considering interdependencies
- **Random Forest:** Identify key features via feature selection technique in multifactor models

Outline

- Introduction
- Methodology
- **Results and Discussion**
- Conclusions

Regression Using Size Feature

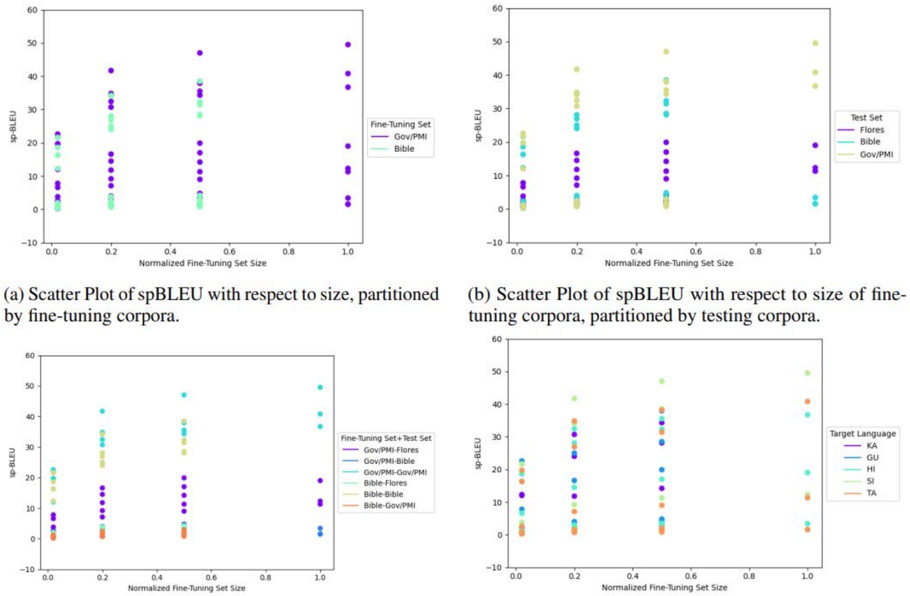
Predictor Function	Partitioning scheme				
	None	Fine-tuning	Testing	Lang	Fine-tuning and Testing
Linear	13.2388	12.9270	11.1404	13.0014	2.9682
Quadratic	13.2092	12.8183	11.1218	13.0414	2.4561
Cubic	13.1706	12.7914	22.4824	13.0601	2.3335
Logarithmic	13.1543	12.7855	11.3084	12.8578	2.3077
Scaling Law	13.1541	12.7828	11.1960	12.8929	**2.2998**



Fine-tuning–test	Normal?	Homoscedastic?
Gov/PMI-Flores	✅	✅
Gov/PMI-Bible	✅	❌
Gov/PMI-Gov/PMI	✅	❌
Bible-Flores	✅	❌
Bible-Bible	✅	✅
Bible-Gov/PMI	✅	❌

Effects of Domain Similarity on Scaling Law

- Partitioning by fine-tuning–testing pairs reduces “clustering” of data points to be modeled
- In/out-domain data separation enhances predictability in scaling law models that uses size as predictor



(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.

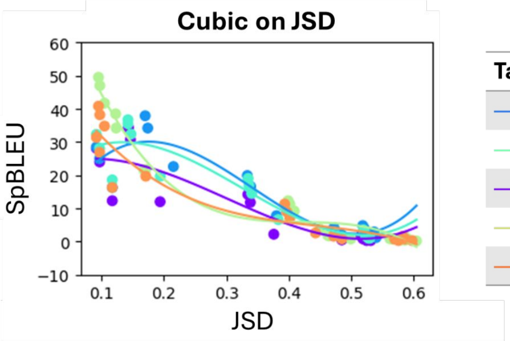
(c) Scatter Plot of spBLEU with respect to size, partitioned by both fine-tuning and testing corpora.

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

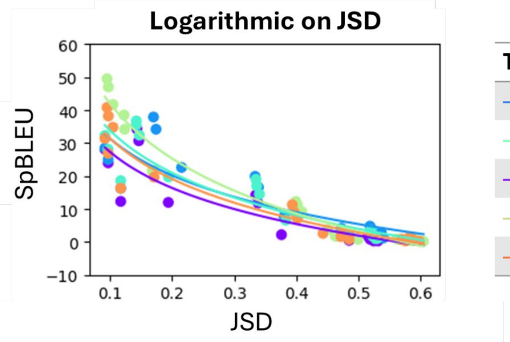
Figure 3: Scatter Plots of spBLEU with respect to size using different partitioning schemes.

Regression Using Domain Similarity

Predictor Function	Partitioning scheme	
	None	Language
Linear	5.6433	5.0782
Quadratic	5.4633	4.5698
Cubic	5.4141	**4.1202**
Logarithmic	5.6315	4.9247



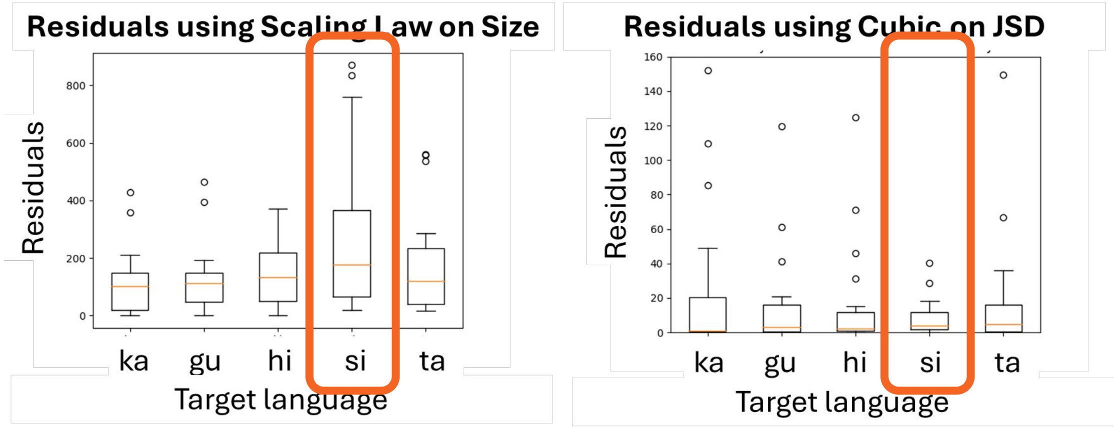
Target Language	Normal?	Homoscedastic?
Gujarati	✅	✅
Hindi	❌	✅
Kannada	✅	✅
Sinhala	✅	✅
Tamil	✅	✅



Target Language	Normal?	Homoscedastic?
Gujarati	✅	✅
Hindi	✅	✅
Kannada	✅	✅
Sinhala	✅	✅
Tamil	❌	✅

Statistical Reliability

- Using JSD as predictor yields a more reliable prediction in terms of normality and homoscedasticity of residuals



Impact of Language Similarity

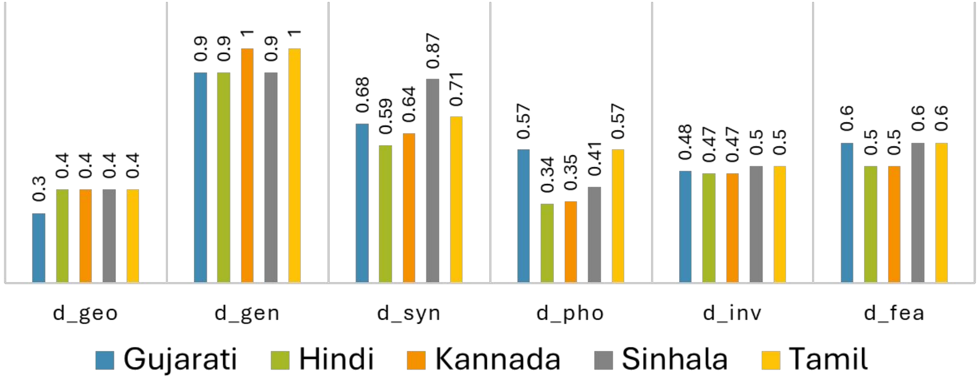
- Single-factor regression models on language features have high RMSE.

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

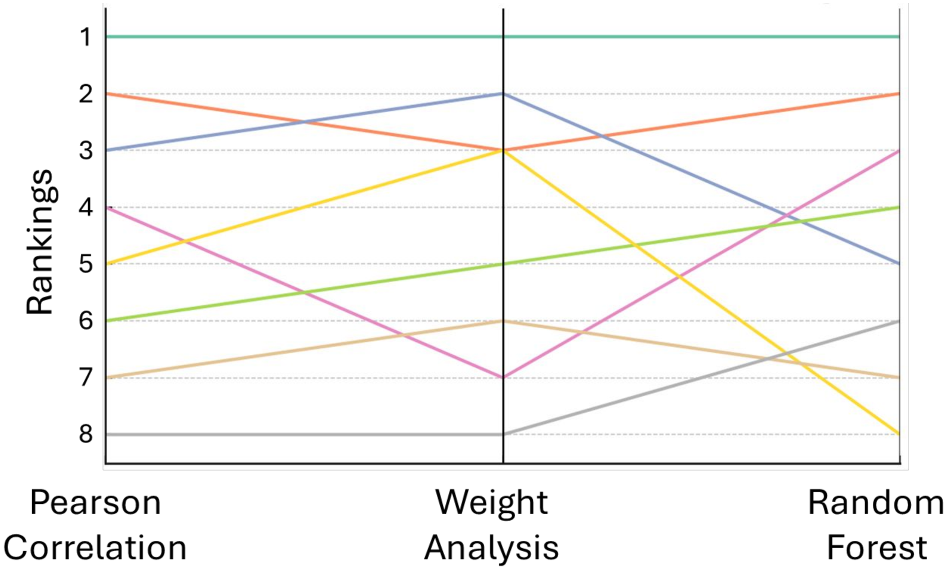
- **Observation:** Including language similarity does not improve the RMSE significantly, implying that it is less important than other factors

Comments on Language Similarity

- The high language similarity in our data renders the effectiveness of language features from lang2vec as predictors
- Insufficient LRL data in the URIEL database limits lang2vec's precision/approximations in describing LRLs
- Low feature discriminative power of LRLs' lang2vec features render the effectiveness of prediction models



Feature Importance Ranking



- JSD outranks all other features in all three rankings

Feature	Random Forest (%)
JSD	88.393
Size	7.805
d_{syn}	2.267
d_{inv}	0.782
d_{gen}	0.365
d_{pho}	0.161
d_{geo}	0.147
d_{fea}	0.079

Outline

- Introduction
- Methodology
- Results and Discussion
- Conclusions

Conclusions

- Domain similarity exerts the most significant impact on performance of MT models, surpassing even the impact of fine-tuning corpus size.
- Using domain similarity as predictor produces the best prediction model in terms of accuracy and statistical reliability.
- A more rigorous study on language similarity measurement is essential to identify suitable predictors for our task.

Thank you!

Link to our paper

Questions?

Contact:
erickhiu@umich.edu