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, Ghent University, Belgium ‡Ontario Tech University, Canada
erickhiu@umich.edu as.dogruoz@ugent.be annie.Lee@ontariotechu.ca

Motivation

Cost of Fine-Tuning Language Models

- High for diverse tasks, languages, and domains.
- Low resource languages (LRLs) lacks data and computing power.

Importance of Performance Data

- Useful for optimizing training cost or for other tasks (e.g., quality estimation (QE)).

Our contributions

- Analyzed the impact of fine-tuning corpus size, domain similarity, and language similarity on MT models for Indian Low Resource Languages (Gujarati, Hindi, Kannada, Sinhala, Tamil) through regression analysis.
- Provided domain-specific and language-specific interpretations based on the performance of regression models.

Methodology

Experimental Data

- Carried out MT experiments using mBart to translate from English to Gujarati, Hindi, Kannada, Sinhala, and Tamil with spBLEU as performance metric from Nayak et.al., (2023).
- Partitioned by fine-tuning corpus size, fine-tuning–testing corpora pair, and target language.

Factors Explored and Featurization

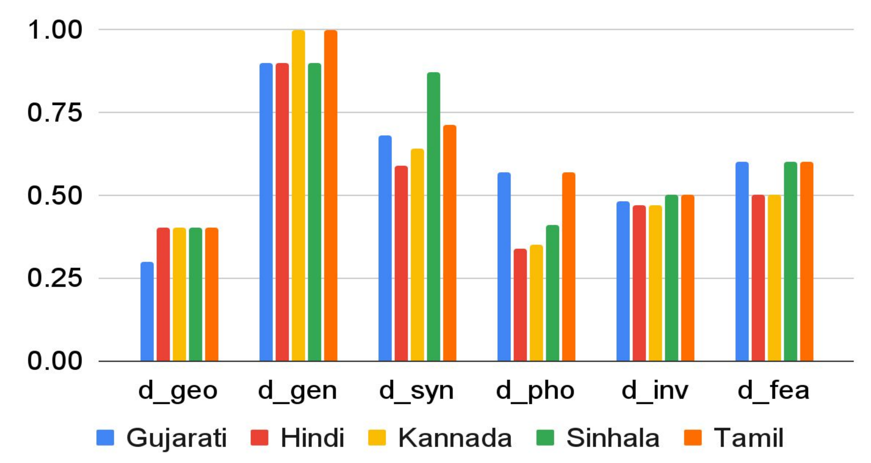
Size: Sentence pair counts in fine-tuning corpora

Domain Similarity: Jensen-Shannon divergence (JSD)

$$JSD(P||Q) = 0.5\,KL(P||M) + 0.5\,KL(Q||M)$$

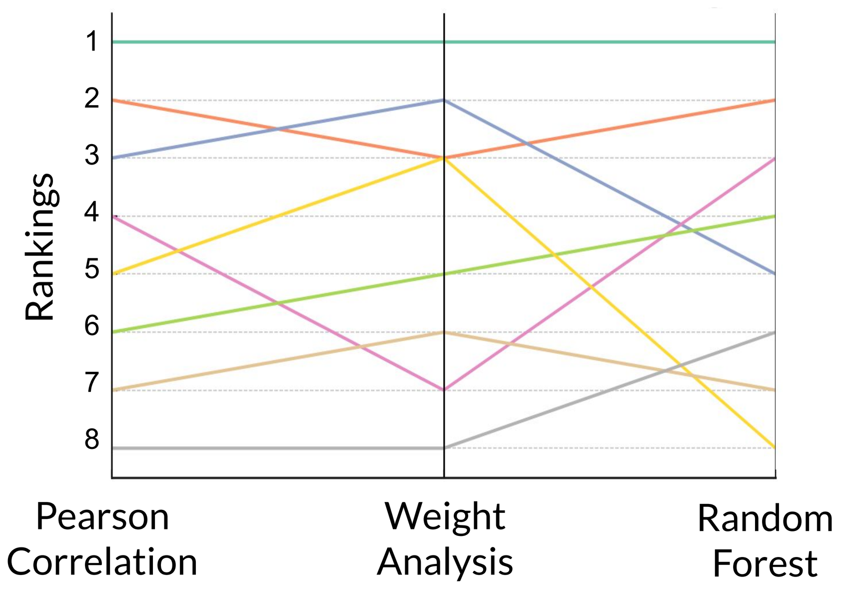
where M is an equally weighted sum of the two distributions and $KL(\cdot||\cdot)$ is the Kullback-Leibler divergence.

Language Similarity: Lang2vec language distances from English



Feature Importance

- JSD top-ranked all three feature importance rankings.
- Most language features ranked lower than JSD and size.

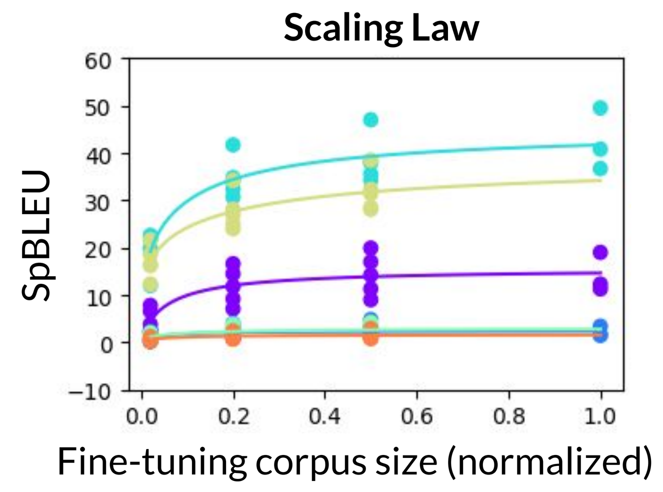


Feature	Random Forest (%)
JSD	88.393
Size	7.805
d_{syn}	2.267
d_{inv}	0.782
d_{gen}	0.365
d_{pho}	0.161
d_{geo}	0.147
d_{fea}	0.079

Regression Analysis

Size

Model: Scaling Law; partitioned by fine-tuning–testing corpora pair; RMSE* = 2.2998

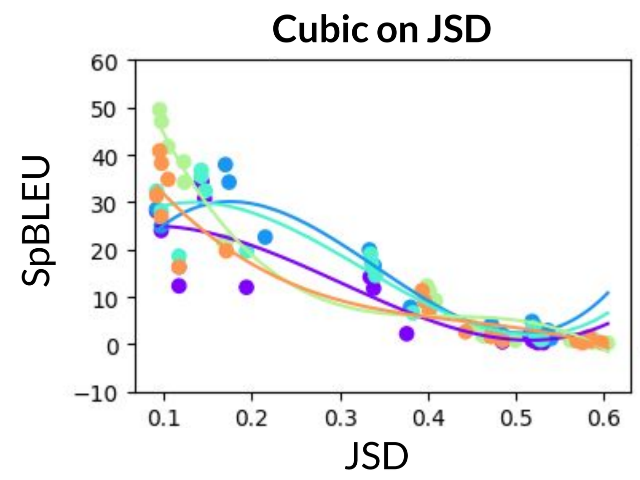


Fine-tuning–testing	Normal?	Homoscedastic?
Gov/PMI-Flores	✅	✅
Gov/PMI-Bible	✅	❌
Gov/PMI-Gov/PMI	✅	❌
Bible-Flores	✅	❌
Bible-Bible	✅	✅
Bible-Gov/PMI	✅	❌

*RMSE: Rooted mean square error

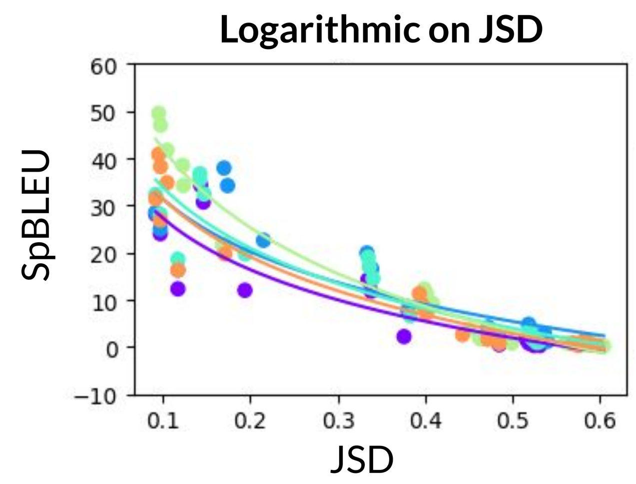
Domain Similarity

Model: Cubic; partitioned by target language; RMSE = 4.1202



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

Model: Logarithmic; partitioned by target language; RMSE = 4.9247



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

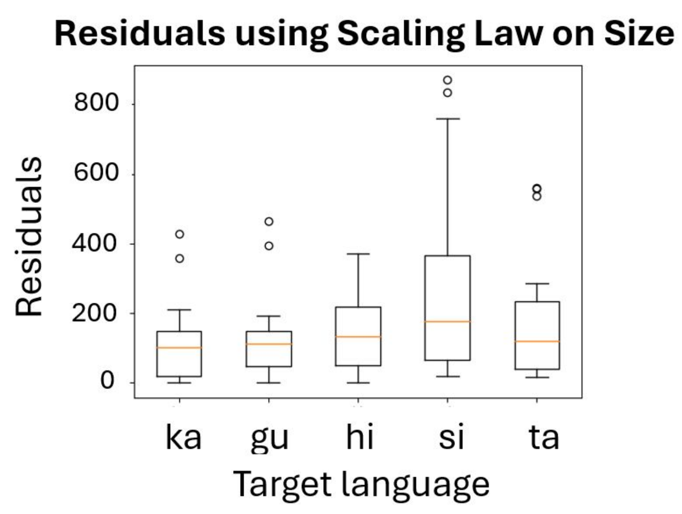
Language Similarity

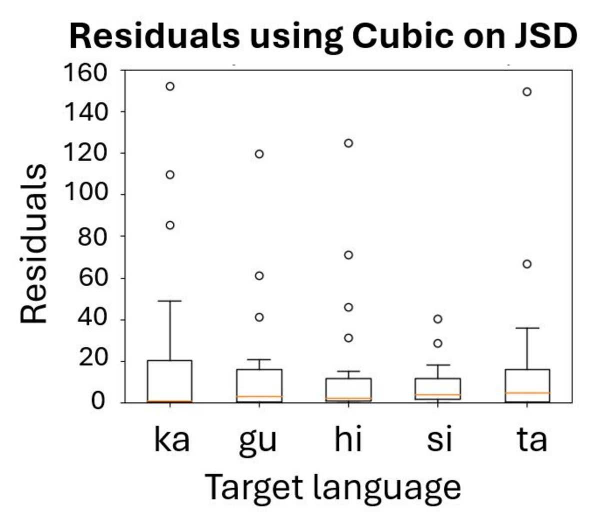
Model	Without language features	With language features
Linear	4.8766	4.5786
Quadratic	4.6604	4.3840
Cubic	4.4509	4.2168
Logarithmic	4.9502	4.6815

- Single-factor regression models on language features have high RMSE.
- Including language features in multifactors models do not significantly improve RMSE.
- Insufficient LRL data in the URIEL database limit lang2vec's precision/ approximations in describing LRLs.
- Low feature discriminative power of LRLs' lang2vec features render the effectiveness of using them as predictors.

Role of Domain Similarity

- Enhanced predictability of scaling law models with in/out-domain data separation (partitioned by fine-tuning–testing pair).
- Yielded a more reliable prediction in terms of normality and homoscedasticity of residuals.





Conclusion & Next Steps

- 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 in terms of accuracy and statistical reliability.
- Next Step: A more rigorous study on language similarity measurement to identify suitable predictors for our task.

Acknowledgement

Thanks to Shravan Nayak, Surangika Ranathunga, Sarubi Thillainathan for the experimental data. Grateful for the Fields Undergraduate Summer Research Program (FUSRP) for their support in this project.