Match the Script, Adapt if Multilingual: Analyzing the Effect of Multilingual Pretraining on Cross-lingual Transferability

Yoshinari Fujinuma,*[1] Jordan Boyd-Graber,[2] Katharina Kann[3]

[1]AWS AI Labs [2]University of Maryland [3]University of Coloardo Boulder

fujinumay@gmail.com jbg@umiacs.umd.edu katharina.kann@colorado.edu

*Work done while at University of Colorado Boulder

At a Glance

Pretrained multilingual models enable zero-shot learning even for unseen languages, and that performance can be further improved via adaptation prior to finetuning.
However, it is unclear how the number of pretraining languages influences a model's zero-shot learning for languages unseen during pretraining. To fill this gap, we ask the following research questions:

- How does the number of pretraining languages influence zero-shot performance on unseen target languages?
- Does the answer to that question change with model adaptation?
- Do the findings for our first question change if the languages used for pretraining are all related?

Our findings are

- *Without* model adaptation, surprisingly, increasing the number of pretraining languages yields better results up to adding related languages, after which performance plateaus.
- *With* model adaptation via continued pretraining, pretraining on a larger number of languages often gives further improvement, suggesting that model adaptation is crucial to exploit additional pretraining languages.

Experimental Setup

- Pretraining Corpus: CoNLL 2017 Wikipedia dump [1] downsampled to ≈ 200MB
- Transformer with same hyperparameters and vocabulary as XLM-R base
- Choice of pretraining languages
  - Diverse set of languages (Div-X)
  - Related set of languages (Rel-X)
- Downstream Tasks: POS, NER, NLI
- Task Dataset: XTREME [2]

For model adaptation on each target language:

- Continued pretraining with Masked Language Modeling [3]
- Adaptation Corpus: JHU Bible Corpus [4]

Pretraining Languages use in the first set of experiments are:

Div-2	EN, RU
Div-3	EN, RU, ZH
Div-4	EN, RU, ZH, AR
Div-5	EN, RU, ZH, AR, HI
Div-6	EN, RU, ZH, AR, HI, ES
Div-7	EN, RU, ZH, AR, HI, ES, EL
Div-8	EN, RU, ZH, AR, HI, ES, EL, FI
Div-9	EN, RU, ZH, AR, HI, ES, EL, FI, ID
Div-10	EN, RU, ZH, AR, HI, ES, EL, FI, ID, TR
Rel-2	EN, DE
Rel-3	EN, DE, SV
Rel-4	EN, DE, SV, NL
Rel-5	EN, DE, SV, NL, DA

Regression Analysis on RQ1

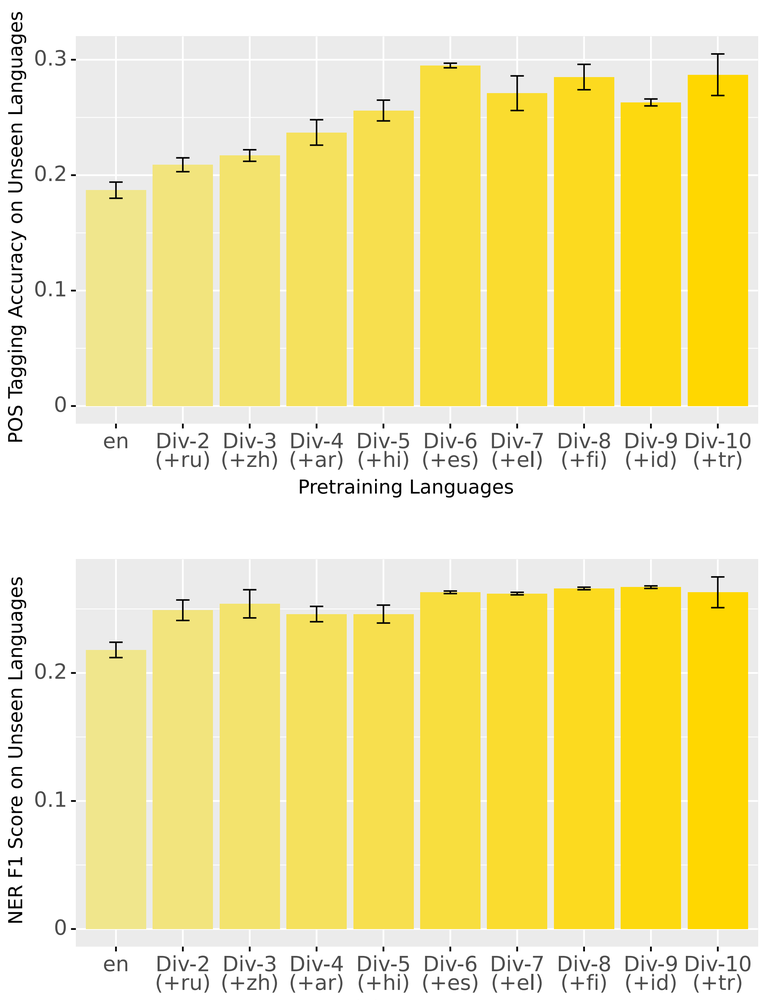
- Predict the POS tagging accuracy Y using X which are the features of pretraining and target languages. Typological features are converted to binary (1 if same, 0 if different).
- Script type match between pretraining languages and the target language is the most important one

Features	Coef.	p-value	CI
Script	**.061**	< .001	[.050, .073]
Family	.022	.004	[.007, .036]
Syntax	.001	.905	[-.016, .018]
Phonology	.021	< .001	[.009, .033]
# pretrain langs	.011	.044	[.000, .022]

Regression analysis on the POS tagging

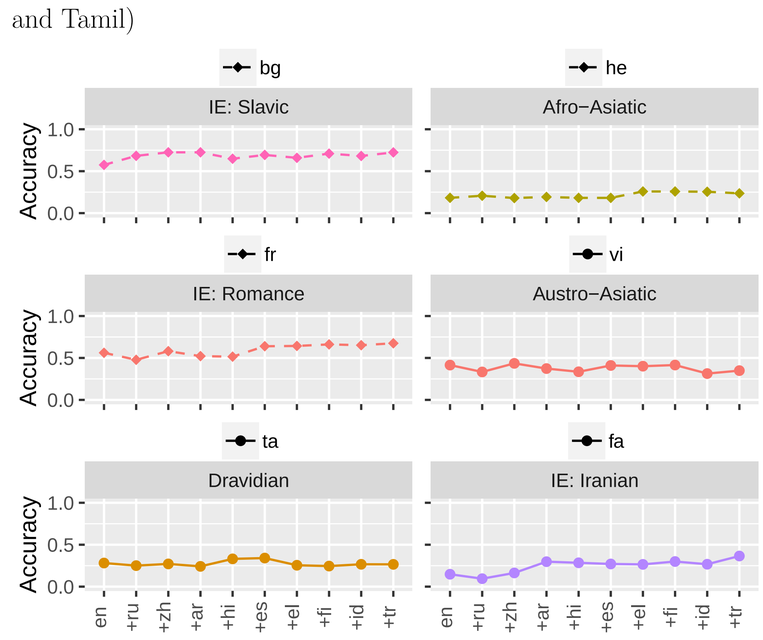
Results: Diverse Languages

- Average cross-lingual zero-shot accuracy increases up to some point



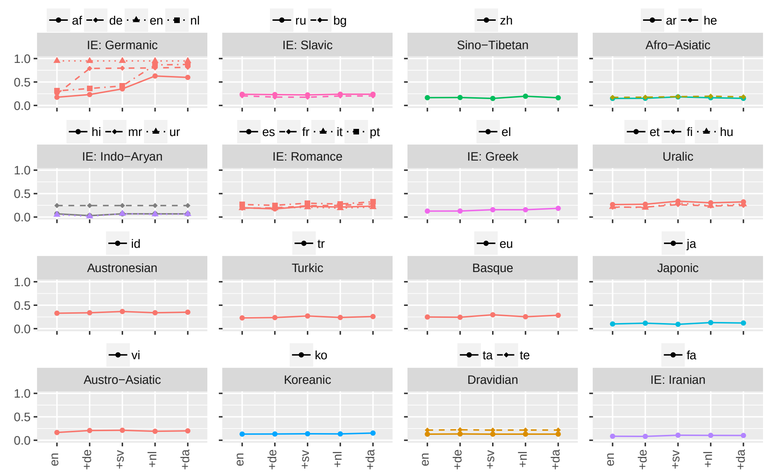
Results: Model Adaptation

- Trend 1: More languages are better (French and Farsi)
- Trend 2: More languages does not necessarily improve (Vietnamese and Tamil)



Results: Related Languages

- Limited cross-lingual transfer across language families



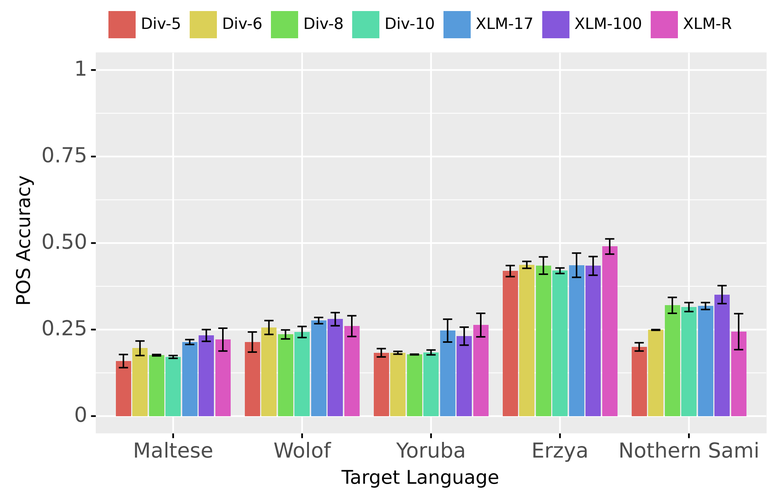
Results up to 100 Languages

Limitations in the First Set of Experiments

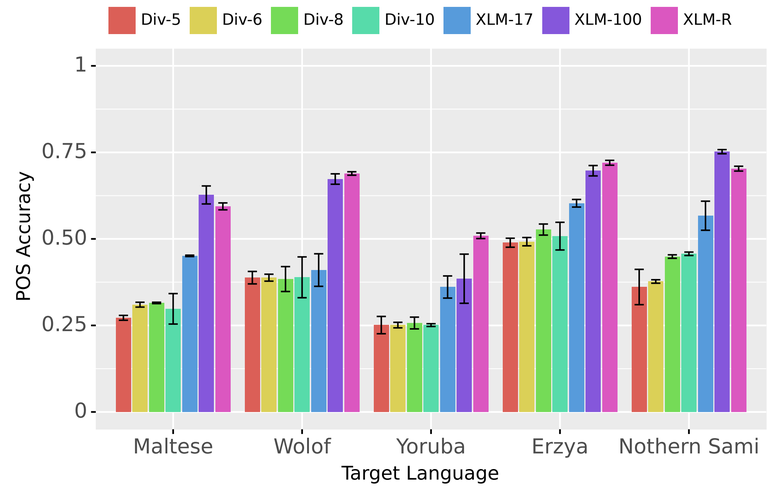
- Computationally intensive
  - Downsampled corpus per language (∼200MB)
  - Up to 10 pretraining languages
- Using XLM-R base vocabulary are not truely unseen for the target languages

Experiment Setup with Model Adaptation on Truely Unseen Languages

- Use the following pretrained models in addition to the pretrained language models up to 10 languages
  - XLM-17 (17 languages, pretrained on full Wikipedia) [5]
  - XLM-100 (100 languages, pretrained on full Wikipedia) [5]
  - XLM-R base (100 languages, pretrained on Common Crawl) [6]



Before Adaptation



After Adaptation

Conclusion

✓ Match the script between pretraining and target languages if not adapting multilingual models

✓ The more languages the better if adapting multilingual models

References

[1] F. Ginter, *et al.*, "CoNLL 2017 shared task - automatically annotated raw texts and word embeddings."

[2] J. Hu, *et al.*, "XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation," in *Proceedings of the International Conference of Machine Learning*.

[3] A. Ebrahimi *et al.*, "How to adapt your pretrained multilingual model to 1600 languages," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, 2021.

[4] A. D. McCarthy, *et al.*, "The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration," in *Proceedings of the Language Resources and Evaluation Conference*, 2020.

[5] G. Lample *et al.*, "Cross-lingual language model pretraining," in *Proceedings of Advances in Neural Information Processing Systems*.

[6] A. Conneau, *et al.*, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the Association for Computational Linguistics*, 2020.