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 Colorado Boulder

* Done while at University of Colorado Boulder

Introduction

- Situation: Traveling in a country where NLP resources are scarce.
- Pretrained multilingual language models are applicable to wide variety of languages under zero-shot transfer setting.

- What happens to languages unseen during pretraining?

Motivation

- Issues with multilingual pretraining for language models
  - Curse of multilinguality (Conneau et al., 2020)
  - Negative interference (Wang et al., 2020)

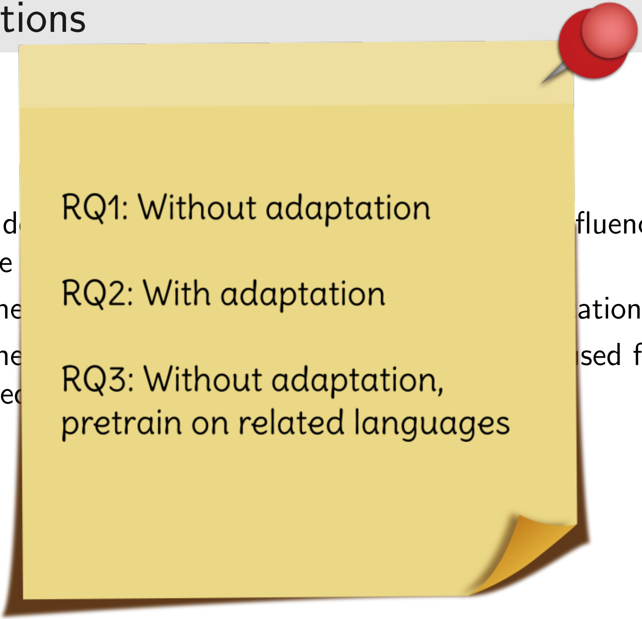
- Are the above phenomenon also an issue for languages unseen during pretraining?
- What happens if we adapt multilingual models to target languages?

Research Questions

- RQ1: How does the number of pretraining languages influence zero-shot performance on unseen target languages?
- RQ2: Do the findings of RQ1 change with model adaptation?
- RQ3: Do the findings of RQ1 change if the languages used for pretraining are all related?

Research Questions

- RQ1: How d fluence zero-shot performance
- RQ2: Do the ation?
- RQ3: Do the sed for pretraining are all relatec



Experiment Setup: Pretraining Languages

- Diverse set of languages (Div-X): up to ten languages from diverse language families for RQ1 and RQ2
- Related set of languages (Rel-X): up to five Germanic Languages for RQ3

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

Experiment Setup: Model, Data and Tasks

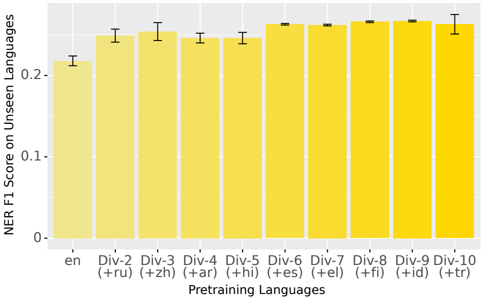
- Transformer with the same architecture and vocabulary as XLM-R base
- Pretraining Corpus: CoNLL 2017 Wikipedia dump (Ginter et al., 2017)
  - Downsampled to $\approx$ 200MB (to the smallest pretraining language)
- Task Dataset: XTREME (Hu et al., 2020)

For RQ2 (with model adaptation):

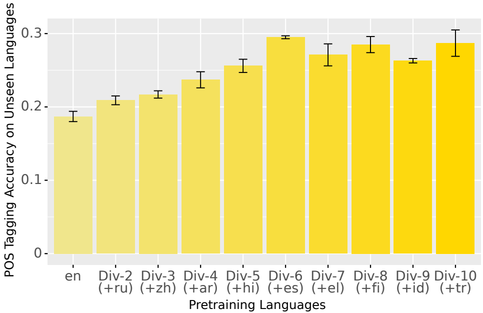
- Continued pretraining with Masked Language Modeling (Ebrahimi and Kann, 2021).
- Adaptation Corpus: JHU Bible Corpus (McCarthy et al., 2020)

RQ1: Results Without Model Adaptation

- Cross-lingual zero-shot accuracy on using diverse set of pretraining languages
-



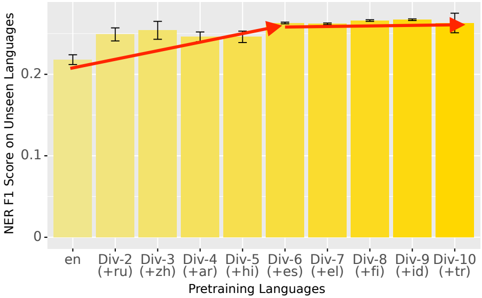
NER Results



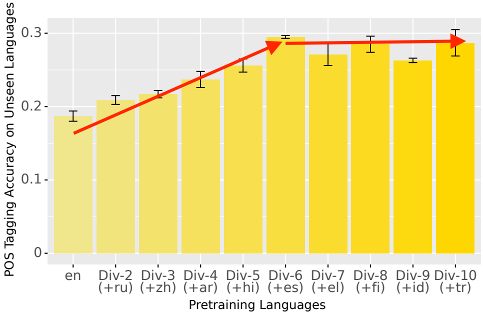
POS Tagging Results

RQ1: Results Without Model Adaptation

- Cross-lingual zero-shot accuracy on using diverse set of pretraining languages
- Average cross-lingual zero-shot accuracy increases up to a certain point



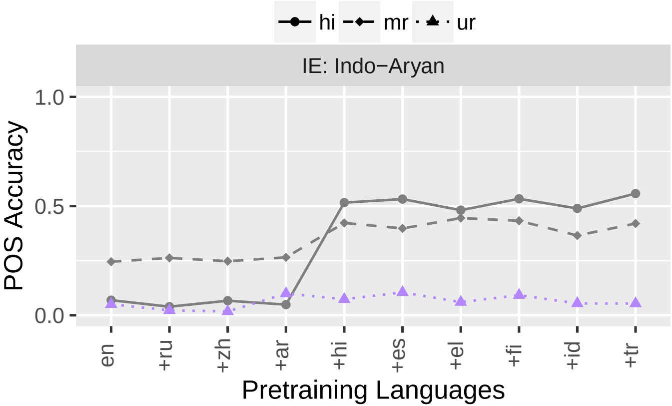
NER Results



POS Tagging Results

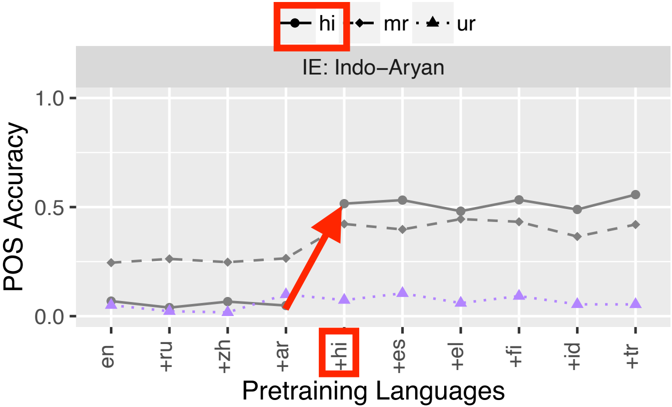
RQ1: Results Without Model Adaptation on Each Language

-
-



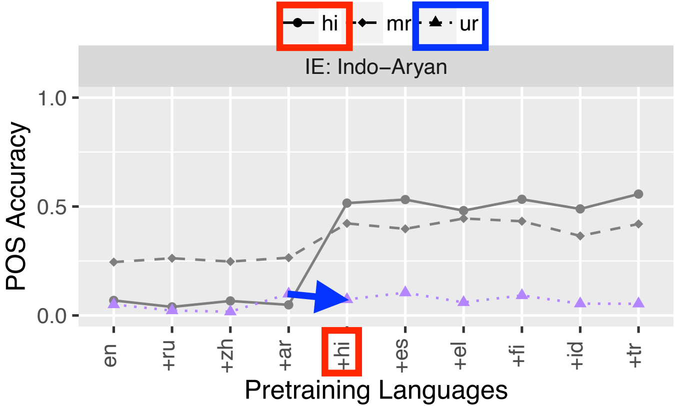
RQ1: Results Without Model Adaptation on Each Language

- Large increase if adding a pretraining language from the same language
-



RQ1: Results Without Model Adaptation on Each Language

- Large increase if adding a pretraining language from the same language
- But no significant gain e.g., in Urdu after adding Hindi



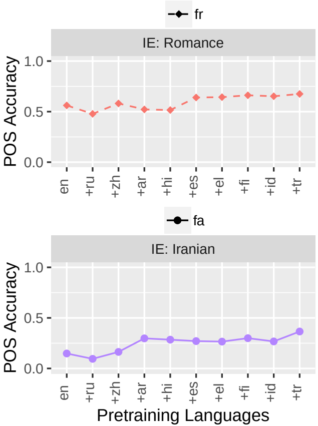
RQ1: Results Without Model Adaptation: Regression Analysis

- Predict the POS tagging accuracy Y using X which are the features of pretraining and target languages
  - Same or different script / family
  - Syntax and phonology features are from URIEL (Littell et al., 2017)
- Script match between pretraining and target languages is the most important one

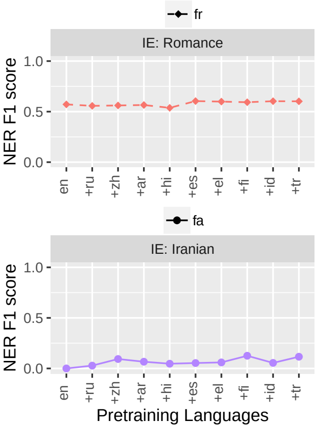
Features	**Coef.**	**p-value**	**CI**
Script	**.061**	$< .001$	[.050, .073]
Family	.022	$.004$	[.007, .036]
Syntax	.001	$.905$	[-.016, .018]
Phonology	.021	$< .001$	[.009, .033]
# pretrain langs	.011	$.044$	[.000, .022]

RQ2: Results With Model Adaptation

- Trend 1: More languages are better (French and Farsi)



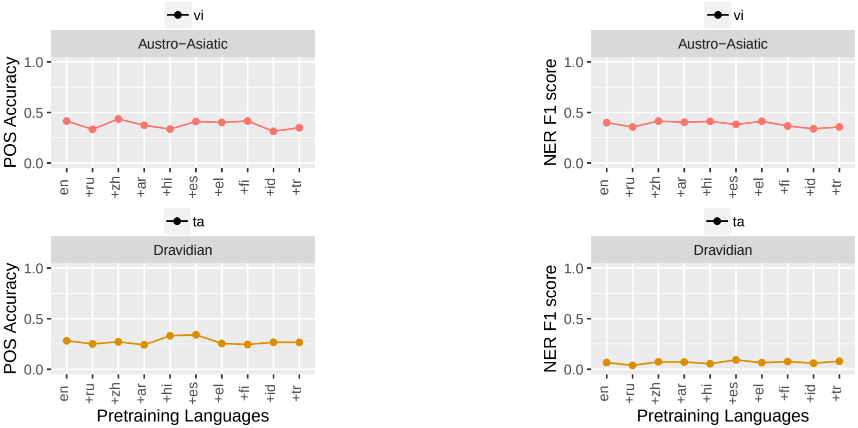
POS Tagging Results



NER Results

RQ2: Results With Model Adaptation

- Trend 2: More languages does not necessarily improve (Vietnamese and Tamil)

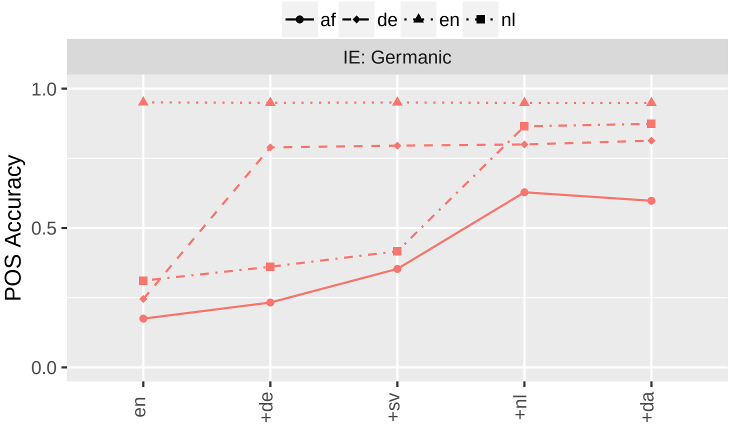


POS Tagging Results

NER Results

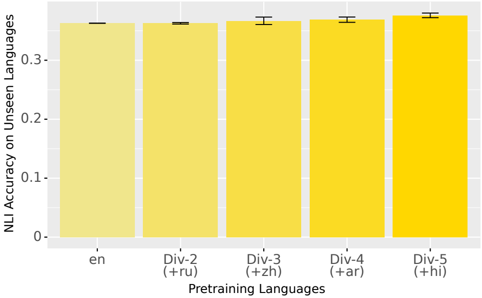
RQ3: Results w/o Adaptation, Pretrain on Related Languages

- Within the same language family: Better cross-lingual transfer
- Limited cross-lingual transfer across language families (please see the paper for details)

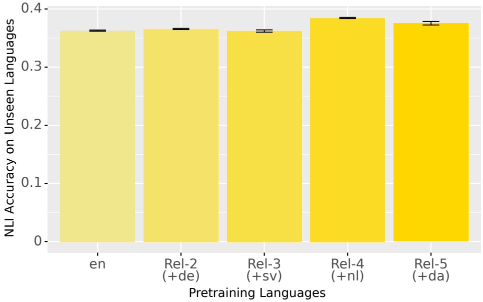


NLI Results

- Trend is less clear on NLI results
- Reason: NLI requires larger pretraining corpus (Lauscher et al., 2020)



Pretrained on Diverse Languages



Pretrained on Germanic Languages

Limitations in the Previous Experiments

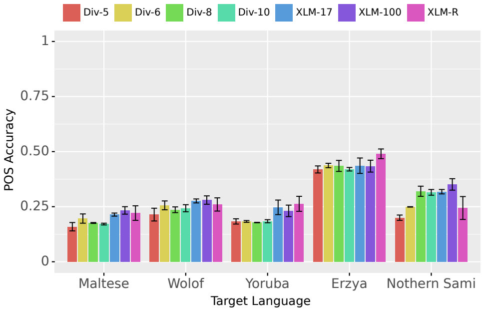
- Small scale due to pretraining being computationally intensive
  - Downsampled corpus per pretraining language ($\sim$200MB)
  - Up to 10 languages
- Using XLM-R base vocabulary are not truely unseen for the target languages

Scaling up to 100+ Languages

- Use more pretrained models in addition to the pretrained language models up to 10 languages
  - XLM-17 (Lample and Conneau, 2019)
    - 17 languages, pretrained on full Wikipedia
  - XLM-100 (Lample and Conneau, 2019)
    - 100 languages, pretrained on full Wikipedia
  - XLM-R base (Conneau et al., 2020)
    - 100 languages, pretrained on Common Crawl
- Target languages unseen when building vocabulary for XLM-17, XLM-100, or XLM-R
  - Maltese, Wolof, Yoruba, Erzya, and Northern Sami

Results up to 100 Languages

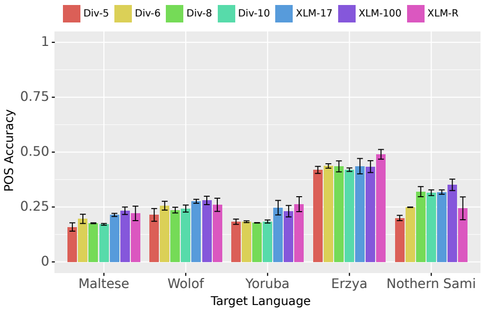
- More languages does not necessarily improve if not adapting



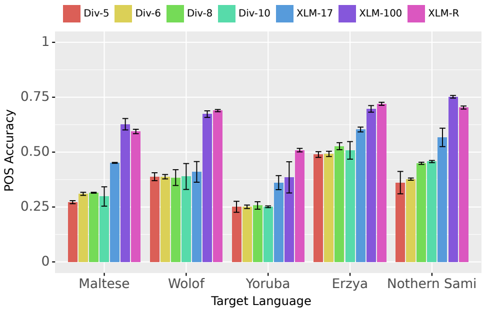
Before Adaptation

Results up to 100 Languages

- More languages does not necessarily improve if not adapting
- More languages the better when adapted



Before Adaptation



After Adaptation

Conclusion and Future Work

If pretraining a new multilingual model and applying to unseen languages:

- If not adapting, small set of diverse pretraining languages is likely sufficient
  - Match the script and family
- If adapting, at least train on 100 languages or possibly more

Future Work:

- Different vocabulary (Muller et al., 2021; Mielke et al., 2021)
- Recent models e.g., mT5 (Xue et al., 2021)
- Beyond simple NLP tasks (e.g., generation tasks)

References I

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *ACL*.

Ebrahimi, A. and Kann, K. (2021). How to adapt your pretrained multilingual model to 1600 languages. In *ACL-IJCNLP*.

Ginter, F., Hajič, J., Luotolahti, J., Straka, M., and Zeman, D. (2017). CoNLL 2017 shared task - automatically annotated raw texts and word embeddings. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *ICML*.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. In *NeurIPS*.

Lauscher, A., Ravishankar, V., Vulić, I., and Glavaš, G. (2020). From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *EMNLP*.

Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C., and Levin, L. (2017). Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *EACL*.

McCarthy, A. D., Wicks, R., Lewis, D., Mueller, A., Wu, W., Adams, O., Nicolai, G., Post, M., and Yarowsky, D. (2020). The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *LREC*.

References II

Mielke, S. J., Alyafeai, Z., Salesky, E., Raffel, C., Dey, M., Gallé, M., Raja, A., Si, C., Lee, W. Y., Sagot, B., and Tan, S. (2021). Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. *CoRR*, abs/2112.10508.

Muller, B., Anastasopoulos, A., Sagot, B., and Seddah, D. (2021). When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *NAACL-HLT*.

Wang, Z., Lipton, Z. C., and Tsvetkov, Y. (2020). On negative interference in multilingual models: Findings and a meta-learning treatment. In *EMNLP*.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer.