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

Yoshinari Fujinuma*
AWS AI Labs
Amazon.com
fujinumay@gmail.com

Jordan Boyd-Graber
UMIACS, CS, LSC, iSchool
University of Maryland
jbg@umiacs.umd.edu

Katharina Kann
Computer Science
University of Colorado Boulder
katharina.kann@colorado.edu

Abstract

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: (1) How does the number of pretraining languages influence zero-shot performance on unseen target languages? (2) Does the answer to that question change with model adaptation? (3) Do the findings for our first question change if the languages used for pretraining are all related? Our experiments on pretraining with *related* languages indicate that choosing a diverse set of languages is crucial. *Without* model adaptation, surprisingly, increasing the number of pretraining languages yields better results up to adding related languages, after which performance plateaus. In contrast, *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.[1]

1 Introduction

Pretrained multilingual language models (Devlin et al., 2019; Conneau et al., 2020) are now a standard approach for cross-lingual transfer in natural language processing (NLP). However, there are multiple, potentially related issues on pretraining multilingual models. Conneau et al. (2020) find the "curse of multilinguality": for a fixed model size, zero-shot performance on target languages seen during pretraining increases with additional pretraining languages only until a certain point, after which performance decreases. Wang et al. (2020b) also report "negative interference", where monolingual models achieve better results than multilingual models, both on subsets of high- and low-resource languages. However, those findings are limited to target languages seen during pretraining.

Current multilingual models cover only a small subset of the world's languages. Furthermore, due to data sparsity, monolingual pretrained models are not likely to obtain good results for many low-resource languages. In those cases, multilingual models can zero-shot learn for unseen languages with an above-chance performance, which can be further improved via model adaptation with target-language text (Wang et al., 2020a), even for limited amounts (Ebrahimi and Kann, 2021). However, it is poorly understood how the number of pretraining languages influences performance in those cases. Does the "curse of multilinguality" or "negative interference" also impact performance on unseen target languages? And, if we want a model to be applicable to as many unseen languages as possible, how many languages should it be trained on?

Specifically, we ask the following research questions: (1) How does pretraining on an increasing number of languages impact zero-shot performance on unseen target languages? (2) Does the effect of the number of pretraining languages change with model adaptation to target languages? (3) Does the answer to the first research question change if the pretraining languages are all related to each other?

We pretrain a variety of monolingual and multilingual models, which we then finetune on English and apply to three zero-shot cross-lingual downstream tasks in unseen target languages: part-of-speech (POS) tagging, named entity recognition (NER), and natural language inference (NLI). Experimental results suggest that choosing a diverse set of pretraining languages is crucial for effective transfer. Without model adaptation, increasing the number of pretraining languages im-

*This work was done while the first author was a student at University of Colorado Boulder.

[1]All code used in this paper is available at https://github.com/akkikiki/multilingual_zeroshot_analysis.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1500 - 1512

proves accuracy on unrelated unseen target languages at first and plateaus thereafter. Last, with model adaptation, additional pretraining languages beyond English generally help.

We are aware of the intense computational cost of pretraining and its environmental impact (Strubell et al., 2019). Thus, our experiments in Section 4 are on a relatively small scale with a fixed computational budget for each model and on relatively simple NLP tasks (POS tagging, NER, and NLI), but validate our most central findings in Section 5 on large publicly available pretrained models.

2 Cross-lingual Transfer via Pretraining

Pretrained multilingual models are a straightforward cross-lingual transfer approach: a model pretrained on multiple languages is then fine-tuned on target-task data in the *source* language. Subsequently, the model is applied to target-task data in the *target* language. Most commonly, the target language is part of the model's pretraining data. However, cross-lingual transfer is possible even if this is not the case, though performance tends to be lower. This paper extends prior work exploring the cross-lingual transfer abilities of pretrained models for *seen* target languages depending on the number of pretraining languages to *unseen* target languages. We now transfer via pretrained multilingual models and introduce the models and methods vetted in our experiments.

2.1 Background and Methods

Pretrained Language Models Contextual representations such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) are not just useful for monolingual representations. Multilingual BERT (Devlin et al., 2019, mBERT), XLM (Lample and Conneau, 2019), and XLM-RoBERTa (Conneau et al., 2020, XLM-R) have surprisingly high cross-lingual transfer performance compared to the previous best practice: static cross-lingual word embeddings (Pires et al., 2019; Wu and Dredze, 2019). Multilingual models are also practical—why have hundreds of separate models for each language when you could do better with just one? Furthermore, Wu and Dredze (2020) report that models pretrained on 100+ languages are better than bilingual or monolingual language models in zero-shot cross-lingual transfer.

Model Adaptation to Unseen Languages Adapting pretrained multilingual models such as mBERT and XLM-R to unseen languages is one way to use such models beyond the languages covered during pretraining time. Several methods for adapting pretrained multilingual language models to unseen languages have been proposed, including continuing masked language model (MLM) training (Chau et al., 2020; Müller et al., 2020), optionally adding Adapter modules (Pfeiffer et al., 2020), or extending the vocabulary of the pretrained models (Artetxe et al., 2020; Wang et al., 2020a). However, such adaptation methods assume the existence of sufficient monolingual corpora in the target languages. Some spoken languages, dialects, or extinct languages lack monolingual corpora to conduct model adaptation, which motivates us to look into languages unseen during pretraining. We leave investigation on the effect of target language-specific processing, e.g., transliteration into Latin scripts (Muller et al., 2021), for future work.

2.2 Research Questions

A single pretrained model that can be applied to any language, including those unseen during pretraining, is both more efficient and more practical than pretraining one model per language. Moreover, it is the only practical option for unknown target languages or for languages without enough resources for pretraining. Thus, models that can be applied or at least easily adapted to unseen languages are an important research focus. This work addresses the following research questions (RQ), using English as the source language for finetuning.

RQ1*: How does the number of pretraining languages influence zero-shot cross-lingual transfer of simple NLP tasks on unseen target languages?*

We first explore how many languages a model should be pretrained on if the target language is unknown at test time or has too limited monolingual resources for model adaptation. On one hand, we hypothesize that increasing the number of pretraining languages will improve performance, as the model sees a more diverse set of scripts and linguistic phenomena. Also, the more pretraining languages, the better chance of having a related language to the target language. However, multilingual training can cause interference: other languages could distract from English, the finetuning source language, and thus, lower performance.

RQ2: *How does the answer to RQ1 change with model adaptation to the target language?*

This question is concerned with settings in which we have enough monolingual data to adapt a pretrained model to the target language. Like our hypothesis for RQ1, we expect that having seen more pretraining languages should make adaptation to unseen target languages easier. However, another possibility is that adapting the model makes any languages other than the finetuning source language unnecessary; performance stays the same or decreases when adding more pretraining languages.

RQ3: *Do the answers to RQ1 change if all pretraining languages are related to each other?*

We use a diverse set of pretraining languages when exploring RQ1, since we expect that to be maximally beneficial. However, the results might change depending on the exact languages. Thus, as a case study, we repeat all experiments using a set of closely related languages. On the one hand, we hypothesize that benefits due to adding more pretraining languages (if any) will be smaller with related languages, as we reduce the diversity of linguistic phenomena in the pretraining data. However, on the other hand, if English is all we use during fine-tuning, performance might increase with related languages, as this will approximate training on more English data more closely.

3 Experimental Setup

Pretraining Corpora All our models are pretrained on the CoNLL 2017 Wikipedia dump (Ginter et al., 2017). To use equal amounts of data for all pretraining languages, we downsample all Wikipedia datasets to an equal number of sequences. We standardize to the smallest corpus, Hindi. The resulting pretraining corpus size is around 200MB per language.[2] We hold out 1K sequences with around 512 tokens per sequence after preprocessing as a development set to track the models' performance during pretraining.

Corpora for Model Adaptation For model adaptation (RQ2), we select unseen target languages contained in both XNLI (Conneau et al., 2018b) and Universal Dependencies 2.5 (Nivre et al., 2019): Farsi (FA), Hebrew (HE), French (FR), Vietnamese (VI), Tamil (TA), and Bulgarian (BG). Model adaptation is typically done for low-resource languages not seen during pretraining because monolingual corpora are too small (Wang et al., 2020a). Therefore, we use the Johns Hopkins University Bible corpus by McCarthy et al. (2020) following Ebrahimi and Kann (2021).[3]

Langs	Tasks
Seen languages	
English (EN)	POS, NER, NLI
Russian (RU)	POS, NER, NLI
Arabic (AR)	POS, NER, NLI
Chinese (ZH)	POS, NER, NLI
Hindi (HI)	POS, NER, NLI
Spanish (ES)	POS, NER, NLI
Greek (EL)	POS, NER, NLI
Finnish (FI)	POS, NER
Indonesian (ID)	POS, NER
Turkish (TR)	POS, NER, NLI
German (DE)	POS, NER, NLI
Dutch (NL)	POS, NER, NLI
Swedish (SV)	-
Danish (DA)	-
Unseen languages	
Bulgarian (BG)	POS, NER, NLI
French (FR)	POS, NER, NLI
Urdu (UR)	POS, NER, NLI
Africaans (AF)	POS, NER
Estonian (ET)	POS, NER
Basque (EU)	POS, NER
Farsi (FA)	POS, NER
Hebrew (HE)	POS, NER
Hungarian (HU)	POS, NER
Italian (IT)	POS, NER
Japanese (JA)	POS, NER
Korean (KO)	POS, NER
Marathi (MR)	POS, NER
Portuguese (PT)	POS, NER
Vietnamese (VI)	POS, NER
Tamil (TA)	POS, NER
Telugu (TE)	POS, NER
Swahili (SW)	NLI
Thai (TH)	NLI

Table 1: Languages used in our experiments.

Tasks We evaluate our pretrained models on the following downstream tasks from the XTREME dataset (Hu et al., 2020): POS tagging and NLI. For the former, we select 29 languages from Universal Dependencies v2.5 (Nivre et al., 2019). For the latter, we use all fifteen languages in XNLI (Conneau et al., 2018b). We follow the default train, validation, and test split in XTREME.

Models and Hyperparameters Following Conneau et al. (2020)'s XLM-R Base model, we train transformers (Vaswani et al., 2017) with 12 layers, 768 units, 12 attention heads, and a maximum of 512 tokens per sequence. To accommodate all

[2] Micheli et al. (2020) show that corpora of at least 100MB are reasonable for pretraining.

[3] In cases where multiple versions of the Bible are available in the target language, we select the largest one.

Model	Pretraining Languages
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

Table 2: Pretraining languages used for the models in our experiments: models are trained on a diverse set (Div-X) and related pretraining languages (Rel-X), with different numbers of pretraining languages.

languages and facilitate comparability between all pretraining setups, we use XLM-R's vocabulary and the SentencePiece (Kudo and Richardson, 2018) tokenizer by Conneau et al. (2020).

We use masked language modeling (MLM) as our pretraining objective and, like Devlin et al. (2019), mask 15% of the tokens. We pretrain all models for 150K steps, using Adam W (Loshchilov and Hutter, 2019) with a learning rate of 1×10^{-4} and a batch size of two on either NVIDIA RTX2080Ti or GTX1080Ti 12GB, on which it approximately took four days to train each model. When pretraining, we preprocess sentences together to generate sequences of approximately 512 tokens. For continued pretraining, we use a learning rate of 2×10^{-5} and train for forty epochs, otherwise following the setup for pretraining. For finetuning, we use a learning rate of 2×10^{-5} and train for an additional ten epochs for both POS tagging and NER, and an additional five epochs for NLI, following Hu et al. (2020).

Languages Table 1 shows the languages used in our experiments. English is part of the pretraining data of all models. It is also the finetuning source language for all tasks, following Hu et al. (2020). We use two different sets of pretraining languages: "Diverse (Div)" and "Related (Rel)" (Table 2). We mainly focus on pretraining on up to five languages, except for POS tagging where the trend is not clear and we further experiment on up to ten.

For POS tagging and NER, we regard seventeen of the twenty-nine languages available in XTREME as *unseen*, while the remaining twelve languages are pretraining languages for at least one model.

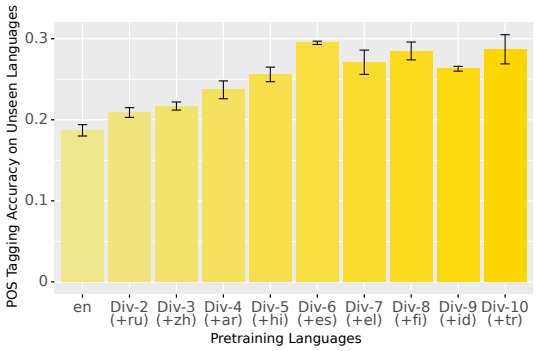


Figure 1: POS tagging accuracy after pretraining on a diverse set of up to 10 languages and finetuning on English. The accuracy improves until six languages on the given target languages.

For NLI, six languages are *seen* and the rest are *unseen*. The order in which we add pretraining languages follows the size of their original CoNLL 2017 Wikipedia dumps, with larger sizes being added first.

4 Results

We now present experimental results for each RQ.

4.1 Findings for RQ1

POS Tagging Figure 1 shows the POS tagging accuracy averaged over the 17 languages unseen during pretraining. On average, models pretrained on multiple languages have higher accuracy on unseen languages than the model pretrained exclusively on English, showing that the model benefits from a more diverse set of pretraining data. However, the average accuracy only increases up to six languages. This indicates that our initial hypothesis "the more languages the better" might not be true.

Figure 2 provides a more detailed picture, showing the accuracy for different numbers of pretraining languages for all seen and unseen target languages. As expected, accuracy jumps when a language itself is added as a pretraining language. Furthermore, accuracy rises if a pretraining language from the same language family as a target language is added: for example, the accuracy of Marathi goes up by 9.3% after adding Hindi during pretraining, and the accuracy of Bulgarian increases by 31.2% after adding Russian. This shows that related languages are indeed beneficial for transfer learning. Also, (partially) sharing the same script with a pretraining language (e.g., ES and ET, AR and FA) helps with zero-shot cross-lingual transfer even for languages which are not from the same

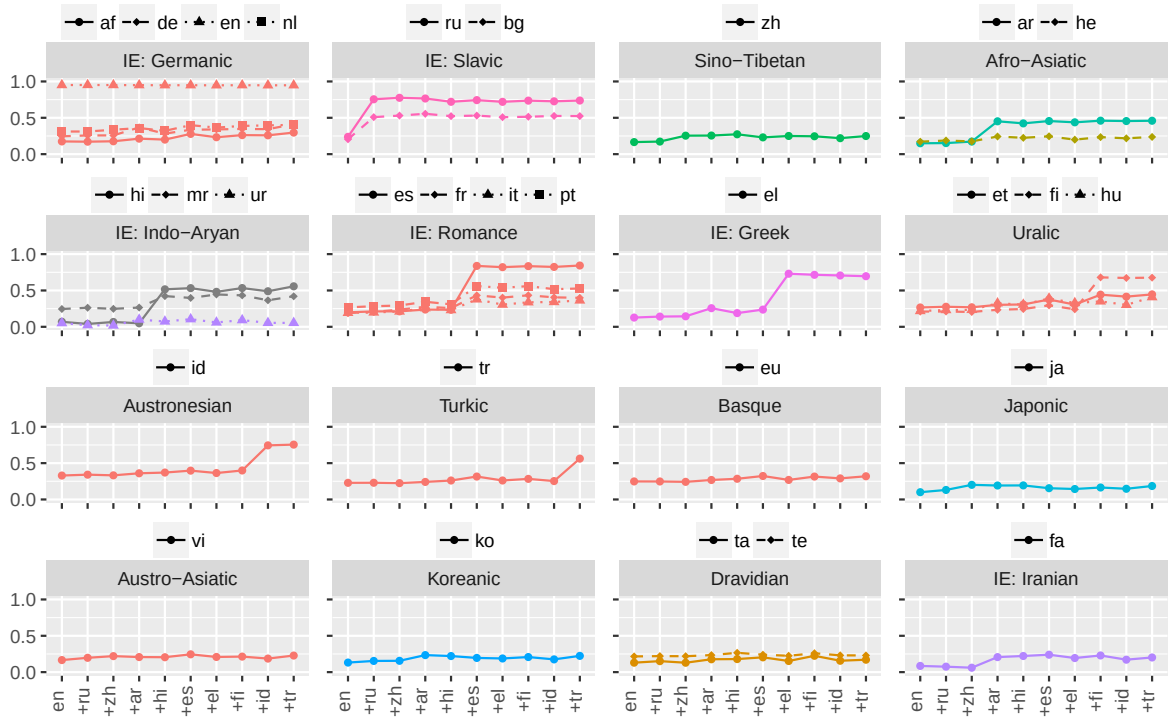


Figure 2: POS tagging accuracy using models pretrained on a diverse set of languages (EN, RU, ZH, AR, HI, ES, EL, FI, ID, TR) grouped by families of target languages, with Indo-European (IE) languages further divided into subgroups following XTREME. The colors represent the script type of the languages. The accuracy gain is larger when a pretraining language from the same family or using the same script is added.

family. These results are consistent with the outcome of Müller et al. (2020) and partially support the hypothesis by Pires et al. (2019) that shared scripts are effective on unseen languages.

But how important are the scripts compared to other features? To quantify the importance of it, we conduct a linear regression analysis on the POS tagging result. Table 3 shows the linear regression analysis results using typological features among target and pretraining languages. For the script and family features, we follow Xu et al. (2019) and encoded them into binary values set to one if a language with the same script or from the same family is included as one of the pretraining languages. For syntax and phonology features, we derive those vectors from the URIEL database using lang2vec (Littell et al., 2017) following Lauscher et al. (2020). We take the maximum cosine similarity between the target language and any of the pretraining languages. Table 3 further confirms that having a pretraining language which shares the same script contributes the most to positive cross-lingual transfer.

We sadly cannot give a definitive optimal number of pretraining languages. One consistent finding is that, for the large majority of languages, using only English yields the worst results for unseen languages. However, adding pretraining languages does not necessarily improve accuracy (Figure 1). This indicates that, while we want more than one pretraining language, using a smaller number than the 100 commonly used pretraining languages is likely sufficient unless we expect them to be closely related to one of the potential target languages.

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]

Table 3: Regression analysis on the POS tagging accuracy with coefficients (Coef.), p-value, and 95% confidence interval (CI). A large coefficient with a low p-value indicates that the feature significantly contributes to better cross-lingual transfer, which shows that the same script is the most important feature.

NER Our NER results show a similar trend. Therefore, we only report the average performance in the main part of this paper (Figure 3), and full

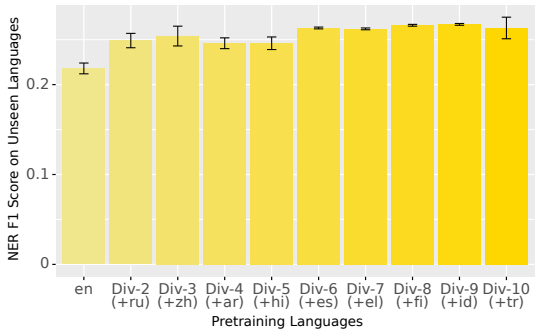


Figure 3: NER F1 score after pretraining on a diverse set of up to 10 languages and finetuning on English.

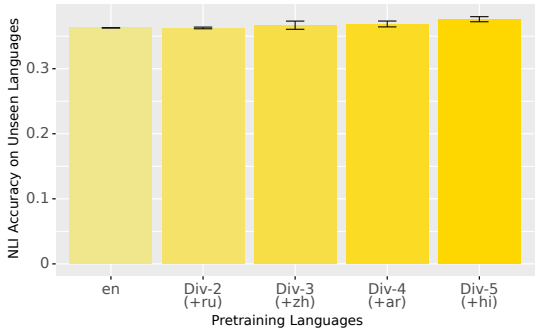


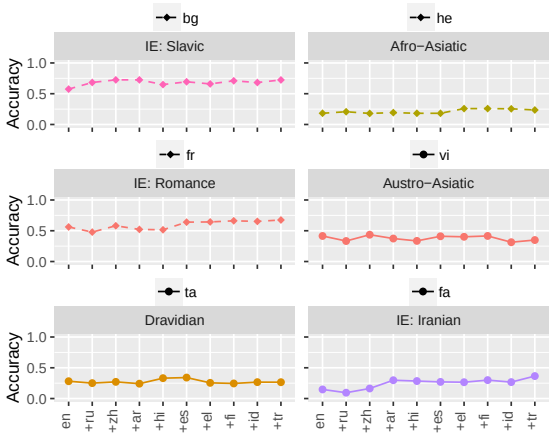
Figure 4: XNLI accuracy after pretraining on a diverse set and finetuning on English.

details are available in Appendix A. For NER, transfer to unseen languages is more limited, likely due to the small subset of tokens which are labeled as entities when compared to POS tags.

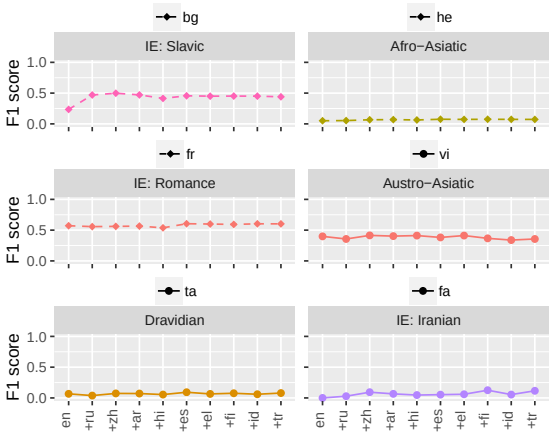
NLI Our NLI results in Figure 4 show a similar trend: accuracy on unseen languages plateaus at a relatively small number of pretraining languages. Specifically, Div-4 has the highest accuracy for 8 target languages, while Div-5 is best only for two target languages. Accuracy again increases with related languages, such as an improvement of 3.7% accuracy for Bulgarian after adding Russian as a pretraining language. Full results are available in Appendix B.

4.2 Findings for RQ2

POS Tagging Figure 5a shows the POS tagging results for six languages after adaptation of the pretrained models via continued pretraining. As expected, accuracy is overall higher than in Figure 2. Importantly, there are accuracy gains in Farsi when adding Turkish (+9.8%) and in Hebrew when adding Greek (+7.7%), which are not observed before adapting models. We further investigate it in Section 5.



(a) POS tagging accuracy.



(b) NER F1 scores.

Figure 5: Results after continued training on the Bible of each target language. The continued training gives limited improvement on NER for most languages when compared to POS tagging.

NER NER results in Figure 5b show similarities between POS tagging (e.g., improvement on Bulgarian after adding Russian). However, there is limited improvement on Farsi after adding Arabic despite partially shared scripts between the two languages. This indicates that the effect of adding related pretraining languages is partially task-dependent.

NLI For NLI, accuracy increases slightly after adding a second pretraining language. Results for two to five pretraining languages are similar for all target languages and, for Greek and Turkish, still similar to the English-only model. This indicates that, similar to our findings for POS tagging, a few pretraining languages could be sufficient for model adaptation. Full results are available in Appendix B. Finally, our NLI results are low overall. This is likely due to the size of the pretraining corpus being one of the top correlated features for NLI (Lauscher

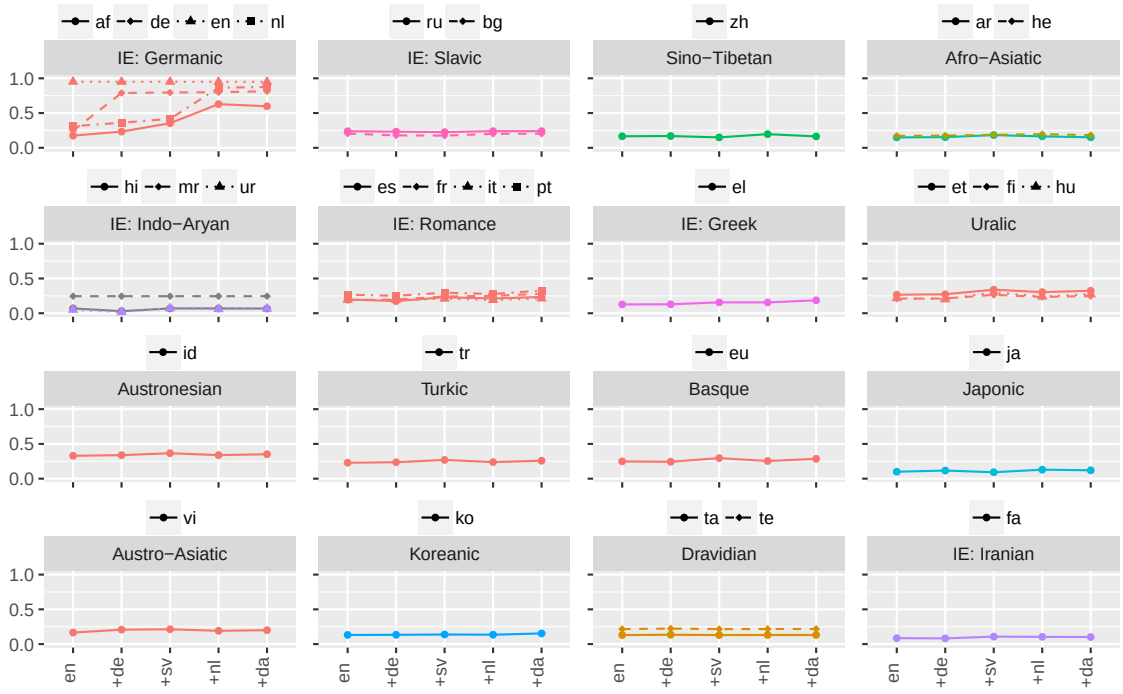


Figure 6: POS tagging accuracy using related pretraining languages (EN, DE, SV, NL, DA) grouped by families of target languages, with Indo-European (IE) languages further divided into subgroups following the XTREME dataset. A change in accuracy can mainly be observed for Germanic, Romance, and Uralic languages due to only using pretraining languages from the Germanic language family.

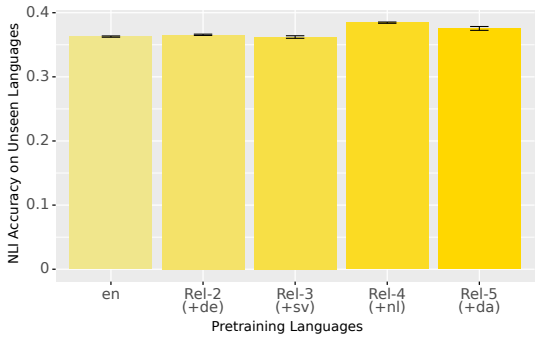


Figure 7: XNLI accuracy on 10 unseen languages after pretraining on a set of related languages and finetuning on English.

et al., 2020), unlike for POS tagging (Hu et al., 2020).

4.3 Findings for RQ3

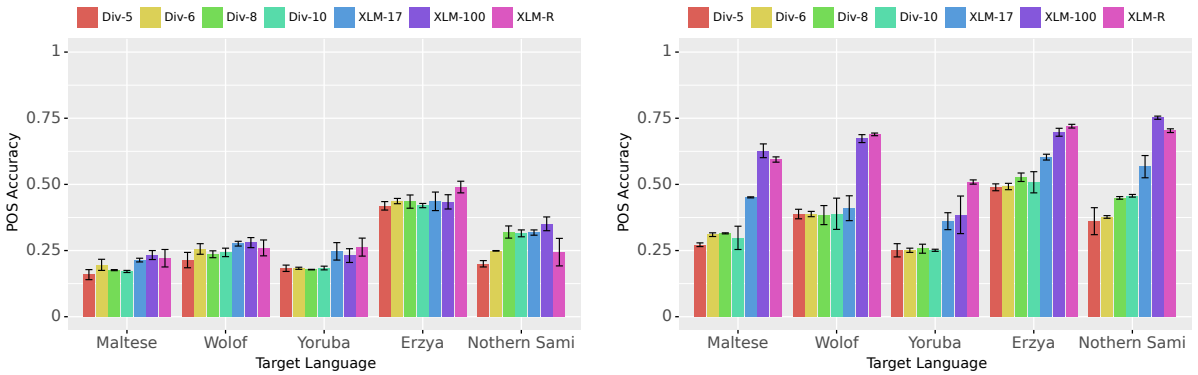
POS Tagging In contrast to RQ1, POS tagging accuracy changes for most languages are limited when increasing the number of pretraining languages (Figure 6). The unseen languages on which we observe gains belong to the Germanic, Romance, and Uralic language families, which are relatively (as compared to the other language families) close to English. The accuracy on languages from other language families changes by $< 10\%$, which is smaller than the change for a diverse set of pretraining languages. This indicates that the models pretrained on similar languages struggle to transfer to unrelated languages.

NER F1 scores of EN, Rel-2, Rel-3, Rel-4, and Rel-5 are .218, .219, .227, .236, and .237 respectively. Compared to Div-X, pretraining on related languages also improves up to adding five languages. However, these models bring a smaller improvement, similar to POS tagging.

NLI Figure 7 shows a similar trend for NLI: when adding related pretraining languages, accuracy on languages far from English either does not change much or decreases. In fact, for nine out of thirteen unseen target languages, Rel-5 is the worst.

5 More Pretraining Languages

Our main takeaways from the last section are: (RQ1) without model adaptation, increasing the number of pretraining languages does not improve accuracy on unrelated unseen target languages; (RQ2) model adaptation largely helps exploiting models pretrained on more languages; and (RQ3)



(a) Before model adaptation.

(b) After model adaptation.

Figure 8: POS tagging accuracy of our models pretrained on a diverse set of languages, XLM-17, XLM-100, and XLM-R after finetuning on English. The models before adaptation are roughly on par regardless of the number of pretraining languages, and the models after adaptation are more affected by related pretraining languages.

when using more than one pretraining language, diversity is important.

However, there are limitations in the experimental settings in Section 4. We assume the following: (1) relatively small pretraining corpora; (2) the target languages are included when building the model's vocabulary; (3) fixed computational resources; and (4) only up to ten pretraining languages. We now explore if our findings for RQ1 and RQ2 hold without such limitations. For this, we use two publicly available pretrained XLM models (Lample and Conneau, 2019), which have been pretrained on full size Wikipedia in 17 (XLM-17) and 100 (XLM-100) languages, and XLM-R base model trained on a larger Common Crawl corpus (Conneau et al., 2020) in 100 languages. We conduct a case study on low-resource languages unseen for all models, including unseen vocabularies: Maltese (MT), Wolof (WO), Yoruba (YO), Erzya (MYV), and Northern Sami (SME). All pretraining languages used in Div-X are included in XLM-17 except for Finnish, and all 17 pretraining languages for XLM-17 are a subset of the pretraining languages for XLM-100. We report the averages with standard deviations from three random seeds.

5.1 Results

RQ1 For models without adaptation, accuracy does not improve for increasing numbers of source languages (Figure 8a). Indeed, the accuracy on both XLM-17 and XLM-100 are on par even though the former uses 17 pretraining languages and the latter uses 100. One exception is Northern Sami (Uralic language with Latin script) due to XLM-17 not seeing any Uralic languages, but XLM-100 does during pretraining. When further comparing Div-10 and XLM-17, increase in accuracy by additional pretraining languages is limited. Erzya remains constant from five to 100 languages (except for XLM-R), even when increasing the pretraining corpus size from downsampled (Div-X) to full Wikipedia (XLM-17 and XLM-100).

RQ2 For the models with adaptation (Figure 8b), there is a significant gap between XLM-17 and XLM-100. This confirms our findings in the last section: more pretraining languages is beneficial if the pretrained models are adapted to the target languages. Thus, a possible explanation is that one or more of XLM-100's pretraining languages is similar to our target languages and such languages can only be exploited through continued pretraining (e.g., Ukrainian included in XLM-100 but not in Div-X). Therefore, having the model see more languages during pretraining is better when the models can be adapted to each target language.

6 Related Work

Static Cross-lingual Word Embeddings Static cross-lingual word embeddings (Mikolov et al., 2013; Conneau et al., 2018a) embed and align words from multiple languages for downstream NLP tasks (Lample et al., 2018; Gu et al., 2018), including a massive one trained on 50+ languages (Ammar et al., 2016). Static cross-lingual embedding methods can be classified into two groups: supervised and unsupervised. Supervised methods use bilingual lexica as the cross-lingual supervision signal. On the other hand, pretrained multilingual language models and unsupervised

cross-lingual embeddings are similar because they do not use a bilingual lexicon. Lin et al. (2019) explore the selection of transfer language using both data-independent (e.g., typological) features, and data-dependent features (e.g., lexical overlap). Their work is on static supervised cross-lingual word embeddings, whereas this paper explores pretrained language models.

Analysis of Pretrained Multilingual Models on Seen Languages Starting from Pires et al. (2019), analysis of the cross-lingual transferability of pretrained multilingual language models has been a topic of interest. Pires et al. (2019) hypothesize that cross-lingual transfer occurs due to shared tokens across languages, but Artetxe et al. (2020) show that cross-lingual transfer can be successful even among languages without shared scripts. Other work investigates the relationship between zero-shot cross-lingual learning and typological features (Lauscher et al., 2020), encoding language-specific features (Libovický et al., 2020), and mBERT's multilinguality (Dufter and Schütze, 2020). However, the majority of analyses have either been limited to large public models (e.g., mBERT, XLM-R), to up to two pretraining languages (K et al., 2020; Wu and Dredze, 2020), or to target languages seen during pretraining. One exception is the concurrent work by de Vries et al. (2022) on analyzing the choice of language for the task-specific training data on unseen languages. Here, we analyze the ability of models to benefit from an increasing number of pretraining languages.

7 Conclusion

This paper explores the effect which pretraining on different numbers of languages has on unseen target languages after finetuning on English. We find: (1) if not adapting the pretrained multilingual language models to target languages, a set of diverse pretraining languages which covers the script and family of unseen target languages (e.g., 17 languages used for XLM-17) is likely sufficient; and (2) if adapting the pretrained multilingual language model to target languages, then one should pretrain on as many languages as possible up to at least 100.

Future directions include analyzing the effect of multilingual pretraining from different perspectives such as different pretraining tasks and architectures, e.g., mT5 (Xue et al., 2021), and more complex tasks beyond classification or sequence tagging.

Acknowledgements

We sincerely thank the reviewers for their constructive and detailed feedback. We also thank the members of University of Colorado Boulder's NALA group, especially Abteen Ebrahimi for providing the code and Stéphane Aroca-Ouellette for giving feedback on an early draft. Boyd-Graber is supported by ODNI, IARPA, via the BETTER Program contract 2019-19051600005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *Computing Research Repository*, arXiv:1602.01925. Version 2.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the Association for Computational Linguistics*.

Ethan C. Chau, Lucy H. Lin, and Noah A. Smith. 2020. Parsing with multilingual BERT, a small corpus, and a small treebank. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the Association for Computational Linguistics*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018a. Word translation without parallel data. In *Proceedings of the International Conference on Learning Representations*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of Empirical Methods in Natural Language Processing*.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. When being unseen from mBERT is just the

beginning: Handling new languages with multilingual language models. In *Proceedings of the Association for Computational Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of Empirical Methods in Natural Language Processing*.

Abteen Ebrahimi and Katharina Kann. 2021. 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*.

Filip Ginter, Jan Hajič, Juhani Luotolahti, Milan Straka, and Daniel Zeman. 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.

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the International Conference of Machine Learning*.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: An empirical study. In *Proceedings of the International Conference on Learning Representations*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of Empirical Methods in Natural Language Processing*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Proceedings of Advances in Neural Information Processing Systems*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the International Conference on Learning Representations*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of Empirical Methods in Natural Language Processing*.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

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. In *Proceedings of the Association for Computational Linguistics*.

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 European Chapter of the Association for Computational Linguistics*.

Ilya Loshchilov and Frank Hutter. 2019. Fixing weight decay regularization in adam. In *Proceedings of the International Conference on Learning Representations*.

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 Language Resources and Evaluation Conference*.

Vincent Micheli, Martin d'Hoffschmidt, and François Fleuret. 2020. On the importance of pre-training data volume for compact language models. In *Proceedings of Empirical Methods in Natural Language Processing*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *Computing Research Repository*, arXiv:1309.4168. Version 1.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Benjamin Müller, Antonis Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2020. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. *CoRR*, abs/2010.12858.

Joakim Nivre, Mitchell Abrams, Željko Agić, and et al. 2019. Universal dependencies 2.5. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of Empirical Methods in Natural Language Processing*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the Association for Computational Linguistics*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the Association for Computational Linguistics*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems*.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020a. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020b. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of Empirical Methods in Natural Language Processing*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*.

Kun Xu, Liwei Wang, Mo Yu, Yansong Feng, Yan Song, Zhiguo Wang, and Dong Yu. 2019. Cross-lingual knowledge graph alignment via graph matching neural network. In *Proceedings of the Association for Computational Linguistics*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer.

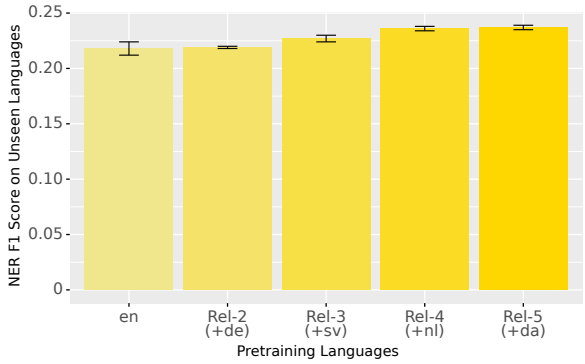


Figure 9: NER F1 score using related pretraining languages (EN, DE, SV, NL, DA)

Pretrain	EL	VI	TR	FR
EN	.351	.367	.365	.395
Div-2 (+ru)	.360	.411	.372	.436
Div-3 (+zh)	.353	.386	.368	.403
Div-4 (+ar)	.362	.395	.374	.438
Div-5 (+hi)	.358	.389	.376	.418

Table 4: NLI accuracy after pretraining on a diverse set of up to 5 languages, continued pretraining on the target-language Bible, and finetuning on English.

A NER Results

We show additional experimental results on NER in Figures 9 and 10.

B NLI Results

Tables 5 and 6 shows the results without model adaptation, and Table 4 shows the full results with model adaptation.

C Notes on the Experimental Setup for Model Adaptation

Following are the additional notes on the setup of the model adaptation:

- No vocabulary augmentation is conducted unlike Wang et al. (2020a). We use XLM-R's vocabulary throughout all experiments in this paper.
- The Bible is used instead of Wikipedia for the continued pretraining or model adaptation to minimize the corpus size and contents inconsistency across languages.

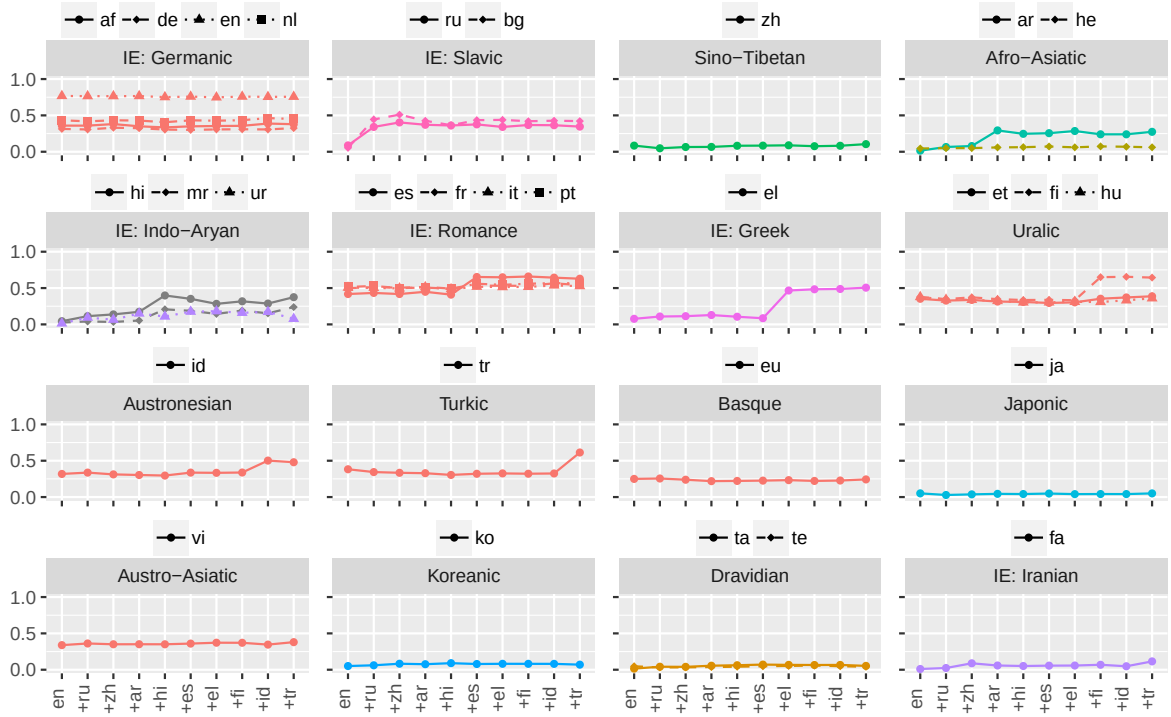


Figure 10: NER F1 score on diverse pretraining languages (EN, RU, ZH, AR, HI, ES, EL, FI, ID, TR) grouped by families of target languages, with Indo-European (IE) languages further divided into subgroups following XTREME. The accuracy gain is significant for seen pretraining languages, and also the languages from the same family of the pretraining languages when added.

Pretrain	en	ru	zh	ar	hi	bg	de	el	es	fr	sw	th	tr	ur	vi
EN	.731	.343	.340	.339	.345	.347	.375	.346	.404	.381	.366	.350	.358	.347	.354
Div-2	.725	.457	.336	.341	.342	.384	.373	.346	.421	.382	.364	.342	.354	.338	.352
Div-3	.738	.500	.485	.336	.338	.389	.374	.341	.412	.382	.354	.340	.345	.339	.345
Div-4	.718	.452	.467	.460	.350	.418	.398	.352	.439	.417	.379	.351	.369	.361	.361
Div-5	.717	.466	.484	.460	.462	.426	.382	.346	.443	.386	.370	.348	.356	.349	.349

Table 5: NLI accuracy on diverse pretraining languages over five seen (EN,RU,ZH,AR,HI) and 10 unseen languages.

Pretrain	en	de	ru	zh	ar	hi	bg	el	es	fr	sw	th	tr	ur	vi
EN	.731	.375	.343	.340	.339	.345	.347	.346	.404	.381	.366	.350	.358	.347	.354
Rel-2	.733	.536	.363	.350	.357	.361	.359	.367	.422	.384	.374	.360	.381	.363	.369
Rel-3	.721	.535	.351	.349	.350	.355	.350	.352	.434	.420	.383	.357	.382	.348	.370
Rel-4	.710	.493	.350	.336	.348	.355	.354	.349	.433	.409	.368	.360	.373	.347	.363
Rel-5	.726	.527	.339	.335	.335	.342	.343	.342	.430	.415	.376	.339	.372	.335	.347

Table 6: NLI accuracy on the 13 unseen languages using the models pretrained on related languages (EN, DE, SV, NL, DA), incrementally added one language at a time up to five languages.