Make the Best of Cross-lingual Transfer: Evidence from POS Tagging with over 100 Languages

Anonymous ACL submission

Abstract

Cross-lingual transfer learning with large multilingual pre-trained models can be an effective approach for low-resource languages with no labeled training data. Existing evaluations of cross-lingual generalisability of large pre-trained models use datasets with English training data, and test data in a selection of target languages. We explore a more extensive transfer learning setup with 65 different source languages and 105 target languages for part-of-speech tagging. Through our analysis, we show that pre-training of both source and target language, as well as matching language families, writing systems, word order systems, and lexical-phonetic distance significantly impact cross-lingual performance.

1 Introduction

At present, for a large majority of natural language processing tasks, the most successful approach is fine-tuning pre-trained models with task-specific labelled data. Unfortunately, for many languages, and especially low-resource languages, such task-specific labelled data is often not available. A potential solution is cross-lingual fine-tuning of multilingual pre-trained language models (Conneau et al., 2020; Devlin et al., 2018), using available data from some source language to model the phenomenon in a different target language for which labelled data does not exist.

Cross-lingual generalisability of large pre-trained language models is often evaluated by fine-tuning multilingual models on English data and testing them on unseen languages (Conneau et al., 2018; Artetxe et al., 2020; Lewis et al., 2020; Hu et al., 2020). Of course, this approach is influenced by the availability of English training data for given tasks, but also then comes with the implicit assumption that English is a representative source language. This, however, may not be true in practice. Specifically, depending on the task, aspects of similarity between source and target language may be relevant for cross-lingual transfer performance (de Vries et al., 2021). If similarity between source and target language impacts performance, cross-lingual transfer should not be assessed using *only* a single predetermined source language, especially if training sets in multiple languages are available.

Furthermore, target test languages are generally selected based on data availability for the evaluated tasks, but availability may not result in a representative subset of the world's languages. The XTreme benchmark collection (Hu et al., 2020), for example, attempts to alleviate this problem by including a varied selection of languages from different language families. This collection contains token classification, text classification, question answering and retrieval tasks in 40 languages. The language selection does, however, obfuscate the fact that for most non-Indo-European and low-resource languages no data is available for semantically rich tasks such as question answering. This imbalance regarding tasks in this type of collections may consequently inflate the perceived performance for these languages.

In this work, we aim to shed light on what factors make a language a good source and/or target language for cross-lingual transfer when fine-tuning a large multilingual model. We evaluate this via part-of-speech (POS) tagging data, as this is the only task for which high-quality data is available in a large number of languages, including low-resource languages from different language families. Also, high cross-lingual POS tagging performance may be seen as a precondition for more semantically complex tasks, as a base understanding of syntactic structure in both the source and target language is necessary for any meaningful natural language processing task.

Contributions This paper is a case-study of cross-lingual transfer learning with part-of-speech tagging. We explore the limits and contributing factors to successful cross-lingual transfer and part-of-speech tagging in particular. Among others, we evaluate the effects of (matching) language families, (matching) writing systems, and pre-training on cross-lingual training. Moreover, we provide insights that can help to estimate performance when one tries to transfer to a low-resource language with little or no annotated data. Source code will be released on Github, and 65 fine-tuned models will be shared via the Hugging Face Model Hub.

2 Approach

We fine-tune a pre-trained model for the task of part-of-speech tagging using different languages in training and testing. Every combination of source and target language yields an accuracy score, with a large matrix of accuracies as a result. Monolingual, or within-language performance is the accuracy where the source and target language are the same. Overall cross-lingual source or target accuracies can be calculated per column or row in the accuracy matrix, excluding the monolingual accuracy. Such accuracies give an overall indication of (i) how suitable a given language is as source for cross-lingual POS tagging, and (ii) how easy or difficult it is to POS-tag a given target language when monolingual training data isn't available.

Predictors Through a mixed-effects regression analysis, with source and target language (family) as random-effect factors, we assess which variables determine a "good" source language. The variables we consider are whether or not the language family is shared between source and target language, the writing systems (and writing system types) of both languages and whether or not these match, the subject-object-verb (SOV) word order of both languages and whether or not these match, and whether or not a (source or target) language was included in pre-training. Additionally, we add the (lexical-phonetic) LDND measure (Wichmann et al., 2010) on the basis of the 40-item word lists from the ASJP database (Wichmann et al., 2010) as a quantitative similarity measure comparing source and target language. Finally, we also consider the size of the training set of the source language as a predictor. We analyze results both from a quantitative and a qualitative viewpoint.

Task data We use the POS tag data from Universal Dependencies 2.8 (Zeman et al., 2021). It contains manually annotated data for 114 languages; among these all have test data and 75 languages have training data. We exclude three mixed-code languages, one sign language, three languages with fewer than 10 test samples and two languages that do not have any word-level annotations. Moreover, we exclude training data for five languages that have fewer than 25 training samples. All other training datasets consist of at least 125 samples. As a result, we have 105 languages which can serve as target languages, of which 65 can also serve as source languages since they have training data.

Model The XLM-RoBERTa base model (Conneau et al., 2020) is used for our experiments.[1] XLM-RoBERTa is pre-trained on web crawled data from 100 languages (with the largest Wikipedia sizes). For our dataset, 53 of our 65 source languages and 58 of our 105 target languages were included in XLM-RoBERTa pre-training.

Data sampling Typical fine-tuning procedures train for a fixed number of epochs on the training data. However, there is a substantial amount of variation in the size of our source language datasets (127 to 163,106 sentences). In such a situation, choosing a fixed number of epochs might result in underfitting for the smaller languages and overfitting for the larger languages. Figure 1 shows that accuracies start decreasing with more than 10K samples, so we choose this threshold for further evaluation. Consequently, the 25 source languages with more than 10K training samples are randomly under-sampled, whereas the other 40 languages are over-sampled (i.e. multiple epochs). The four languages with more than 50K training samples (German, Czech, Russian and Turkish) achieve highest overall average accuracy with 1250, 20K, 1250 and 10K samples, respectively, showing that under-sampling can improve cross-lingual performance. Within-language performance does keep increasing with longer training, which indicates that longer training can cause source language overfitting.

Training procedure All models are trained with the same hyper-parameter settings. Specifically, the models are trained for 1,000 batches of 10 samples with a linearly decreasing learning rate starting at $5e-5$. We use 10% dropout between transformer layers and 10% self-attention dropout. These hy-

[1] Preliminary experiments have shown no performance gain with the large model variant, so out of practical and environmental considerations, we limit our experiments to this model.

perparameters were selected based on preliminary experiments with the English, Dutch, Armenian, Marathi and Chinese source languages. Models for different source languages were trained with the same random seed.

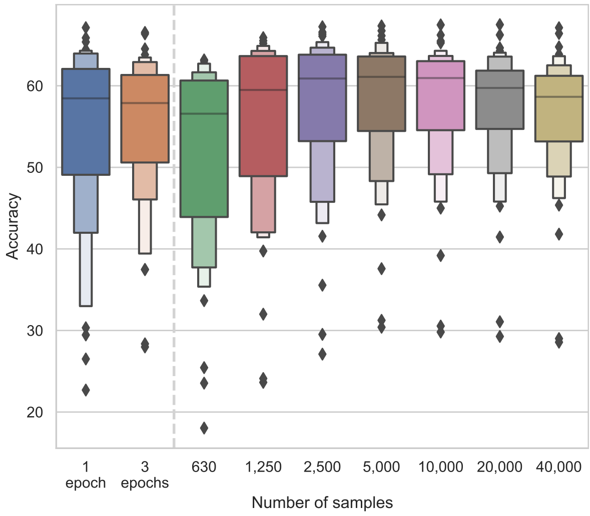

Figure 1: Accuracy distributions for different sampling strategies. Median and mean overall POS-tagging accuracy starts decreasing with more than 10K training samples.

Predictor	Coef.	Std. Err.
(Intercept)	42.2	3.3
Target pre-trained	19.2	2.5
LDND distance	−12.7	1.0
Both pre-trained	7.4	7.4
Same family	6.8	6.8
Source pre-trained	5.6	2.0
Same writing system type	3.6	0.4
Same writing system	1.4	0.3
Same SOV word order	1.3	0.2

Table 1: Coefficients and standard errors of predictors in the final mixed-effects regression model with Accuracy as the dependent variable. All predictors were significant at the $p < 0.01$ level. LDND distances were scaled between 0 (minimum) and 1 (maximum). The predictors are sorted in order of decreasing importance.

3 Results

Figure 2 illustrates the test accuracies for every combination of source and target language. The heat map shows that the model achieves relatively high performance for cases where the source and target language is the same (outlined in black). While for many languages same-language training is the only way to achieve high performance (for example Maltese), there are also many target languages for which high performance is observed when training on several other languages (for example Russian). Indeed, within-language performance tends to be high with a mean accuracy of 94.1% ($\sigma = 4.5$). However, there is a substantial amount of variation for cross-lingual accuracies with an overall mean of 57.4% ($\sigma = 22.4$). This shows that cross-lingual training does not universally yield good performance.

We evaluate several predictors for inclusion (see Section 2) by adding them to a linear mixed-effects model with random intercepts for source language, target language, and target language family. No other random intercepts were found to improve the model (via model comparison). We ascertained that the predictors of the final model remained significant when the corresponding random slopes were included. These are not further reported, however. Fixed-effect predictors were included if they significantly ($p < 0.05$) improved the model fit as determined via (maximum likelihood) model comparison. Table 1 shows the predictors included in the final model. This mixed-effects regression model yields a conditional R^2 of 91.1% and a marginal R^2 of 47.1%. In other words, the included fixed effects explain 47.1% of variance, whereas the additional 44% is captured by the random effects (i.e. other language-related factors). Regarding the random-effects, the variance explained by the target language was more than three times as high as the variance explained by the source language, reflecting the fact that the POS accuracy is much stronger linked to the target language than to the source language. This is also visible in Figure 2, where the rows are much more variable than the columns.

4 Quantitative discussion

4.1 Pre-training

Table 1 shows that the best predictor for accuracy differences is whether the target language is included in pre-training or not, with an estimated 19.2% higher accuracy for target languages that were included. Similarly, performance is higher when the source language is included in pre-training, but with a much smaller effect (5.6%) as the target language. There is an additional increase of 7.4% in accuracy if both the source language

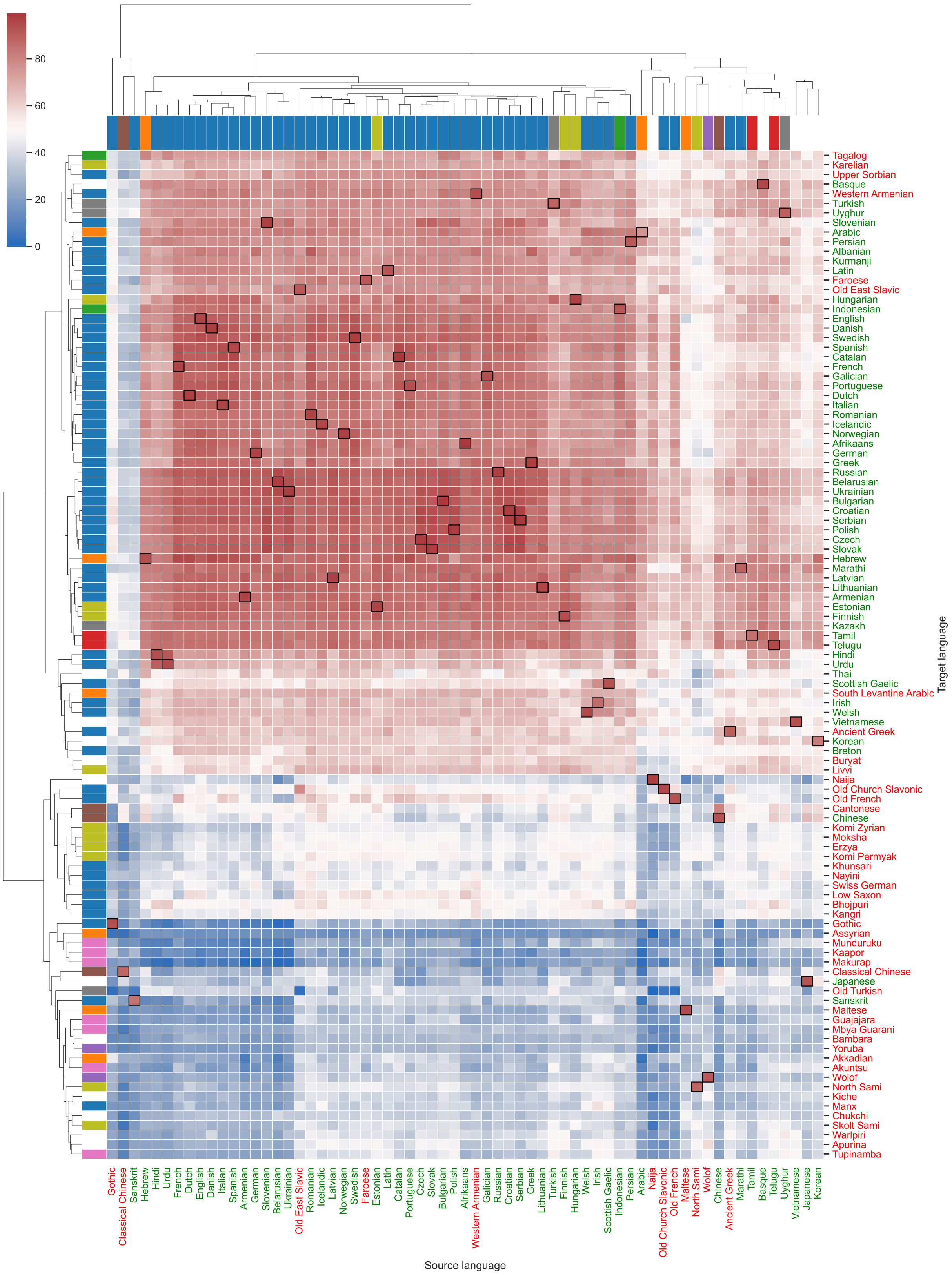

Figure 2: Universal Dependencies part-of-speech tagging accuracies for every combination of source (column) and target (row) languages by fine-tuning XLM-RoBERTa base on the source language. Language names printed in green were included in XLM-RoBERTa pre-training, whereas language names printed in red were not. Group colors in the dendrograms indicate different language families. Different shades of blue indicate different branches in the Indo-European language family. Dendrograms are based on hierarchical clusters using unweighted average linkage clustering (UPGMA) with the Euclidean distance metric.

and target language are included in pre-training. Consequently, inclusion in pre-training, especially the target language, is highly important for achieving high cross-lingual performance. This is unfortunate for many low-resource languages that are not included in pre-training, as the benefit from cross-lingual transfer will be limited. Specific examples of underperforming languages that were not included in pre-trainig are discussed in Section 5.1.

4.2 LDND distance

The ASJP-based LDND measure has the strongest effect on predicted accuracy after target language inclusion in pre-training with a coefficient of -12.70. Figure 3 shows that low LDND distances between source and target language (i.e. when two languages share cognates) are indeed associated with high accuracy, whereas high LDND distances (very dissimilar languages) seem less informative.

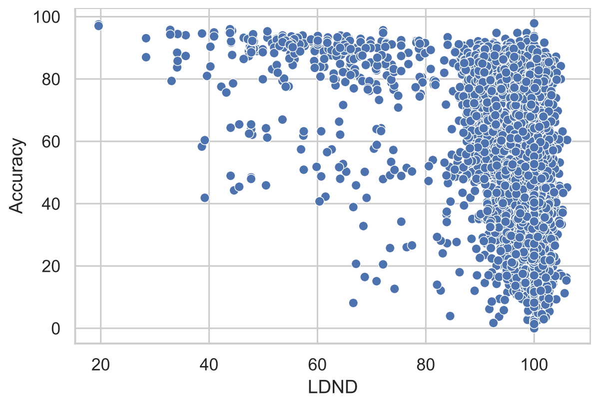

Figure 3: Relation between LDND lexical-phonetic distances and accuracy.

This significant effect might be surprising as the measure is based on (broad) phonetic transcriptions of single words, but measures of linguistic distance at different levels are correlated (Spruit et al., 2009).

4.3 Language family

Whether source and target languages are part of the same language family has a considerable effect on accuracy (see Table 1)[2]. Therefore, when choosing a source language, the best option would be a language from the same family. Figure 4 shows the average accuracies per language family combination. This figure is solely based on target languages

[2]Preliminary experiments have shown that splitting the large Indo-European language family into the major branches does not contribute to the explainability of the model.

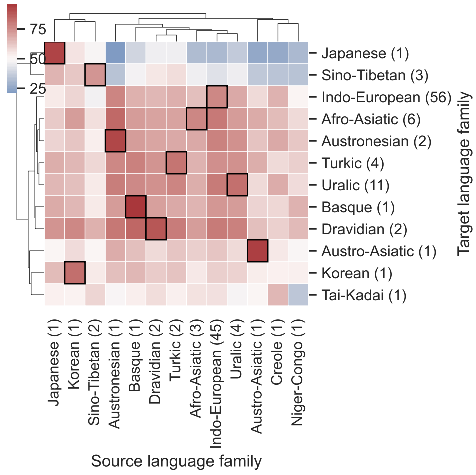

Figure 4: Average accuracies per source and target language family combination based on target languages that were included in pre-training. Numbers between parentheses indicate the number of languages in each family. High performance can be observed within language families (black outlines). Dendrograms are based on hierarchical clusters using unweighted average linkage clustering (UPGMA) with the Euclidean distance metric.

that were included in pre-training, since absence from pre-training has a large negative effect on performance as previously discussed (see Section 4.1).

The Japanese and Sino-Tibetan (Chinese, Classical Chinese and Cantonese) target languages only reach reasonable accuracies with Japanese, Sino-Tibetan or Korean source languages. These target languages reach a lower than 50% macro-averaged accuracy across language families. This could be a reflection of the type of writing system in those languages (see Section 4.4 for a dedicated discussion on this), but this is not certain. Tai-Kadai (Thai), Korean, and Austro-Asiatic (Vietnamese) languages also reach relatively low cross-family macro-average accuracies (up to 60%), whereas the remaining target language families generally reach a higher performance.

In Section 3, we found that accuracy is higher if the source and target language are the same, but transfer can work between different families. Figure 4 shows that some family combinations might not be suitable for transfer, but since the lower-performing families contain small numbers of languages, it is difficult to reach definitive conclusions.

4.4 Writing systems

Regarding writing systems, we distinguish writing system types (i.e. alphabetic, logosyllabic, abjad,

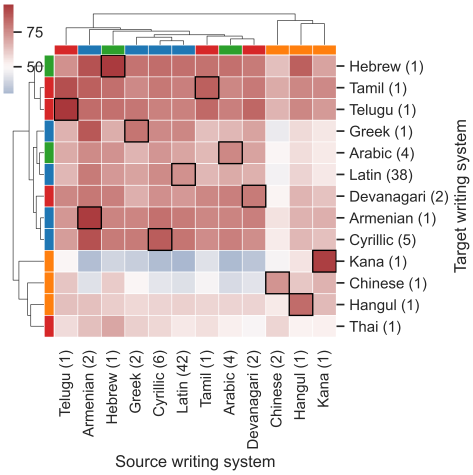

Figure 5: Average accuracies per source and target writing system combination based on target languages that were included in pre-training. Numbers between parentheses indicate the number of languages that use each writing system. Dendrograms are based on hierarchical clusters using unweighted average linkage clustering (UPGMA) with the Euclidean distance metric. Dendrogram colors represent writing system types (blue: alphabetic, orange: logosyllabic, red: abiguda, green: abjad.)

and abiguda[3]) from the more fine-grained writing systems (e.g., Armenian, Greek, Cyrillic, and Latin are all alphabetic). Cross-lingual POS-tagging accuracy is higher if the source and target writing system types are similar. If the two languages share the same writing system, performance is even better (see Table 1).

Languages that share a writing system, such as Latin, can benefit from a shared vocabulary if those languages have some lexical overlap (Pires et al., 2019). However, a shared vocabulary also introduces cross-lingual homography problems, where the same token has different meanings, and thus possibly different grammatical functions, in different languages. Both aspects are not present for languages that use different writing systems, even if the vocabulary is technically shared within a multilingual model.

Figure 5 shows average cross-writing-system accuracies. Some singleton writing systems reach very low accuracies. These are the logosyllabic Chinese characters, Kana (Japanese) and Hangul (Korean) writing systems and Thai, which is an abugida writing system. There are several other writing systems that are used by a single target language and achieve high performance regardless of source writing system, i.e. Hebrew, Tamil and Telugu. This might indicate that the data or the language itself is easier than other target languages.

Cross-script transfer seems to work well for a subset of writing systems. Languages with logosyllabic or the Thai writing system, tend to perform poorly with source languages that use different writing systems. However, these writing systems are not used across language families, so it is difficult to attribute these findings specifically to the writing systems themselves.

5 Qualitative discussion

Having discussed significant predictors in detail, we now take a closer look at "bad" source languages, thereby providing a better understanding of how to choose a "good" source language (Section 5.1). We also identify some optimal source-target language pairs (Section 5.2), and "optimal" source languages for our task (Section 5.3).

5.1 Underperforming source languages

Figure 2 shows that many source languages (columns) achieve high performance for at least a subset of the target languages, and also that some source languages never achieve high cross-lingual accuracies. While overall contributing factors have been discussed in Section 4, here we unpack why some source languages yield low accuracy.

Source languages should achieve highest performance on themselves as target languages. This is not the case for Arabic (higher accuracy on Ukrainian), Korean (higher accuracy on Hebrew) and Spanish (higher accuracy on Catalan). Excluding those languages, the lowest within-language accuracy is Sanskrit (84.2%). We identify poorly performing source languages as those where the best cross-lingual accuracy is below that 84.2% threshold. Based on this threshold, we identify 19 source languages that perform sub-optimally on every target language except themselves.

The full set of source languages contains 12 languages that were not included in XLM-RoBERTa pre-training (see red column labels in Figure 2). Out of these 12 languages, nine are in the bottom 25% of source languages ranked by overall accu-

[3] Characters in logosyllabic writing systems represent full words (logograms) or syllables. In abugida writing systems, consonants and vowels are combined as single units. This can make abugida writing systems similar to syllabic writing systems for character-based NLP systems. Abjad writing systems only use characters for consonants, whereas vowels are implied.

racy: Ancient Greek, Classical Chinese, Gothic, Maltese, Naija, North Sami, Old Church Slavonic, Old French and Wolof. The remaining three source languages that were not included in pre-training are Faroese, Old East Slavic and Western Armenian. The written forms of these three languages are considered mutually intelligible with at least one language that was included in pre-training.[4] Specifically, mutually intelligible are written Faroese with Icelandic (Barbour and Carmichael, 2000), Old East Slavic with Russian, Belarusian and Ukrainian (Andersen, 2003) and West Armenian with (Eastern) Armenian (Adalian, 2010). No similar mutual intelligibility pairs were found for the nine poorly performing non-pre-trained source languages. This indicates that while inclusion in pre-training is optimal for both the source and target language, inclusion of a mutually intelligible language variant can be sufficient for source languages.

Other source languages that never achieve high transfer performance but that were present in pre-training are Sanskrit, Arabic, Chinese, Japanese, Vietnamese, Uyghur, Irish, Marathi, Hebrew, Tamil. For Uyghur and Irish, no clear cause could be found for their low performance. This is not the case for the other languages, however.

Sanskrit is effectively not present in pre-training, since the Universal Dependencies data mainly contains romanized Sanskrit, whereas the data in the XLM-RoBERTa pre-training uses the Devanagari writing system. Serbian is the only other evaluated source language where the writing system in Universal Dependencies is not used in pre-training. However, the Latin script that is used in Universal Dependencies is used with the Croatian pre-training data, and Croatian is structurally and in written form effectively the same language as Serbian (Kordić, 2010).

For Arabic, the problem seems a poor model fit in general, since the trained model for Arabic also achieves only 75.9% accuracy on Arabic test data. We did not identify a clear external factor for why Arabic performance is so low, since other genetically related languages and other languages that use the Arabic writing system perform better.

Problems with Chinese, Japanese and Vietnamese might originate from issues with logosyllabic writing systems (see Section 4.4). Japanese uses its own unique syllabic writing system, and the Vietnamese language uses a romanized version of (logographic) Chinese characters. Logosyllabic writing systems therefore seem to transfer poorly to other languages. The languages in our set of source languages with logosyllabic writing systems are Japanese, Chinese, Classical Chinese and Cantonese. These four languages are in the bottom 20% lowest performing source languages for average cross-lingual accuracy. While the source writing system type was not identified as a significant predictor in the mixed-effects regression model, this could be because logosyllabic writing systems are not used across multiple language families.

The three remaining poorly performing languages are Marathi, Hebrew and Tamil. Those three languages are the only evaluated source languages with fewer than 200 training sentences. Therefore, the reason for the low performance of these source languages could be the lack of sufficient training data.

Overall, these findings suggest that a good source language should:

- Be included in pre-training data with the same writing system as the task-specific training data. Alternatively, a mutually intelligible related language must be included;
- Achieve good within-language performance. One cannot expect high cross-lingual performance, if a model performs poorly on the source language itself;
- Use the same type of writing system as the target language. Transfer between different alphabetic writing systems (i.e. Latin and Cyrillic) can work well, but lower performance is observed for logosyllabic writing systems (see Section 4.4);
- Have sufficient training data available. Using only 200 training sentences seems too little.

5.2 Optimal language pairs

For every target language, the best source language can be determined by taking the source language with the highest accuracy. Some highly similar languages are each other's best source language. In our set of languages, we found 11 of such pairs:

- Estonian and Finnish
- Icelandic and Faroese
- French and Italian
- Chinese and Japanese
- Irish and Scottish Gaelic

[4] If we consider these languages as pre-trained in the mixed effects model of Section 3, the marginal R^2 would increase from 47.1% to 54.6%.

- Croatian and Serbian
- Catalan and Spanish
- Belarusian and Ukrainian
- Hindi and Urdu
- Armenian and Western Armenian
- English and Swedish

All of these pairs, except *English and Swedish*, originate from countries that are geographic neighbours, or in the same country. Moreover, most of these pairs are closest siblings according to the Ethnologue genetic classification scheme (Eberhard et al., 2021), compared to alternative languages in our language set. The exceptions are *English and Swedish* (both are Germanic languages, but for instance Dutch is closer to English, and Norwegian is closer to Swedish), *Chinese and Japanese* (separate families, but Japanese has many Chinese loan words) and *Catalan and Spanish* (Portuguese is genetically closer to Spanish than Catalan).

Some of these pairs are known to have mutual intelligibility (see Section 5.1) and share common ancestor languages. This shows that optimal cross-lingual performance can be achieved by pairing highly similar languages. However, since all of these pairs are languages that were included in pre-training, it is unclear whether this also holds for low-resource languages that were not included.

5.3 The best source language

Romanian and Swedish are the most common best source language for any target language, with 10 and 7 target languages, respectively. Alternatively, optimal cross-lingual performance can be determined by taking the average cross-lingual accuracy per source language. According to this measure the best source languages are still Romanian (67.2%) and Swedish (65.9%). This criterion ranks English as 19th out of 65 source languages, with an average accuracy of 62.4%. All languages that perform better than English are Indo-European except Estonian (Uralic), and English is the fifth-best source language from the Germanic Indo-European branch. Romanian is also, on average, the best source language for both the set of target languages that were included in pre-training (81.5%) as well as the set of non-pre-trained languages (49.8%). This shows that even though cross-lingual tansfer commonly takes English as a source language, English might not be the best source language overall.

However, overall average performance might not be a good measure of source language quality because that introduces a strong Indo-European bias, due to the large amount of Indo-European languages in our target language set. If we determine the best source language per target language family (or Indo-European branch), we find that the best source language is from a different language family for 23 out of 30 families. Again, Romanian is the best general source language since it is the best source language for seven different families. All other best source languages occur twice (Chinese, Uyghur and Wolof) or once (17 languages).

In short, for this particular task, with this particular dataset, Romanian as source language achieves the best cross-lingual performance.

6 Conclusion

We show that simply fine-tuning a large multilingual pre-trained language model on English data does not necessarily make full use of the cross-lingual potential of the model. Especially when one applies cross-lingual training for a low-resource language with little or no evaluation data, the different factors that influence performance should be kept in mind. Unfortunately, one of the most important factors highlighted by our experiments is that the target language, or a highly similar language variant, should be included in pre-training for cross-lingual training to be successful. For current language models, this excludes many languages and a large number of language families. For those languages, the most important step is to collect unlabeled data for pre-training.

Languages that are included in pre-training can achieve high cross-lingual performance across language families and writing systems, at least for languages that use alphabetic writing systems. The English language, which is the *de facto* default source language for cross-lingual training, is not necessarily the best source language.

Due to data availability, our experiments focused on POS tagging, but we hypothesize that the factors we identified may be predictive for other tasks too. The significant influence of lexical-phonetic distances and word order differences on accuracies indicate that similar languages are encoded similarly in XLM-RoBERTa, even if there is no lexical overlap due to differing writing systems. Thus, these factors potentially also influence more syntax-dependent tasks, such as parsing, and semantically rich tasks, such as natural-language-inference.

Ethics statement

We used freely available data and a freely available pre-trained model for our experiments. Our experimental setup required fine-tuning many large language models, but we ran preliminary experiments on a few languages to determine whether we could achieve sufficient performance with a small model size. As this indeed was the case, environmental impact was limited compared to the larger model size. Moreover, to limit the need for future fine-tuning efforts for this task, we release all of the fine-tuned models.

References

Rouben Paul Adalian. 2010. *Historical dictionary of Armenia*. Scarecrow Press.

Henning Andersen. 2003. Slavic and the indo-european migrations. *Amsterdam studies in the theory and history of linguistic science. Series 4*, pages 45–76.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Stephen Barbour and Cathie Carmichael. 2000. *Language and nationalism in Europe*. OUP Oxford.

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 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP 2018*, pages 2475–2485.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2021. Ethnologue: Languages of the World. Twenty-fourth edition. SIL International.

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 generalization. *CoRR*, abs/2003.11080.

Snježana Kordić. 2010. Jezik i nacionalizam (language and nationalism). *Zagreb: Durieux (Rotulus Universitas)*.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Marco René Spruit, Wilbert Heeringa, and John Nerbonne. 2009. Associations among linguistic levels. *Lingua*, 119(11):1624 – 1642. The Forests behind the Trees.

Wietse de Vries, Martijn Bartelds, Malvina Nissim, and Martijn Wieling. 2021. Adapting monolingual models: Data can be scarce when language similarity is high. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4901–4907, Online. Association for Computational Linguistics.

Søren Wichmann, Eric W. Holman, Dik Bakker, and Cecil H. Brown. 2010. Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and its Applications*, 389(17):3632 – 3639.

Daniel Zeman, Joakim Nivre, et al. 2021. Universal dependencies 2.8.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.