Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Cross-lingual Strategies for Low-resource Language Modeling: A Study on Five Indic Dialects

Niyati Bafna[1] Cristina España-Bonet[3] Josef van Genabith[2,3]
Benoît Sagot[1] Rachel Bawden[1]
(1) Inria, Paris, France
(2) Saarland Informatics Campus, Saarland University, Germany
(3) DFKI GmbH, Germany
niyatibafna13@gmail.com,{josef.van_genabith,cristinae}@dfki.de,
{benoit.sagot,rachel.bawden}@inria.fr

ABSTRACT

Neural language models play an increasingly central role for language processing, given their success for a range of NLP tasks. In this study, we compare some canonical strategies in language modeling for low-resource scenarios, evaluating all models by their (finetuned) performance on a POS-tagging downstream task. We work with five (extremely) low-resource dialects from the Indic dialect continuum (Braj, Awadhi, Bhojpuri, Magahi, Maithili), which are closely related to each other and the standard mid-resource dialect, Hindi. The strategies we evaluate broadly include from-scratch pretraining, and cross-lingual transfer between the dialects as well as from different kinds of off-the-shelf multilingual models; we find that a model pretrained on other mid-resource Indic dialects and languages, with extended pretraining on target dialect data, consistently outperforms other models. We interpret our results in terms of dataset sizes, phylogenetic relationships, and corpus statistics, as well as particularities of this linguistic system.

RÉSUMÉ

Stratégies inter-langues pour la modélisation des langues à faibles ressources : étude sur cinq dialectes indo-aryens

Les modèles de langue neuronaux jouent désormais un rôle central en traitement automatique des langues, grâce à leurs performances sur de nombreuses tâches du domaine. Dans cet article, nous étudions le développement de tels modèles pour des langues (très) peu dotées, à savoir cinq langues du continuum dialectal indo-aryen (braj, awadhi, bhojpuri, magahi, maithili), toutes très proches du hindi, une langue moyennement dotée. Nous comparons plusieurs stratégies classiques par l'adaptation (*finetuning*) et l'évaluation sur la tâche d'étiquetage en parties du discours. Ces stratégies incluent le pré-entraînement à partir de zéro ainsi que le transfert entre dialectes et depuis des modèles multilingues existants. Nous constatons qu'un modèle préentraîné sur d'autres dialectes et langues indiennes moyennement dotées avec poursuite du préentraînement sur les données du dialecte cible surpasse systématiquement les autres modèles. Nous interprétons nos résultats à la lumière de la taille des jeux de données et de leurs propriétés statistiques, des relations phylogénétiques entre dialectes, ainsi que des particularités de ce système linguistique.

KEYWORDS: Language modeling, low-resource, Indic languages, cross-lingual transfer, POS tagging.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs, pages 28–42.

MOTS-CLÉS : Modèles de langue, peu dotée, langues indo-aryennes, transfert cross-lingue, étiquetage morpho-syntaxique.

1 Introduction

In the last decade, natural language models have made tremendous progress on multiple tasks (Kalyan *et al.*, 2021). Many recent advances in natural language processing (NLP) owe credit to neural models that are pretrained over large quantities of unlabeled text, such as BERT (Devlin *et al.*, 2019). Data inequity over the vast range of the world's languages has led efforts to "transfer" these data-hungry neural models from resource-rich languages such as English and Spanish, to lower-resource languages (Wu & Dredze, 2019), with many studies focusing on phylogenetic closeness between the source and target languages as one of the important factors determining the results (Lin *et al.*, 2019; Dhamecha *et al.*, 2021; Patil *et al.*, 2022).

Indic NLP, or NLP for Indian languages,[1] has also made corresponding advances, with the release of large corpora, language models, and benchmarks for 18 "major" Indian languages (Kakwani *et al.*, 2020). However, there are hundreds of other languages and dialects in India, many of them spoken by millions of people, such as Rajasthani, Kannauji, Garhwali, and others, that have non-existent or nascent NLP research (Bafna *et al.*, 2022).

We work with a typical real-world situation, with five (extremely) low-resource North Indian dialects belonging to the Indic language family—namely, Braj, Awadhi, Bhojpuri, Magahi, Maithili. These languages bear close relationships with Hindi, a mid-resource dialect[2] although with a number of morphosyntactic and lexical divergences. We compare strategies for language modeling for low-resource languages, using a part-of-speech (POS) tagging downstream task for evaluation. We report that relatedness at the level of the language family between the pretraining languages and the target language benefits downstream performance. However, while comparing low-resource dialects as sources for a particular target dialect in a transfer setup, the results are less explained by phylogeny than by corpus domain match and lexical overlap. Finally, we find that extended pretraining shows consistent benefits. We hope that these experiments will raise an interest in NLP for these dialects, and constitute a starting point for other work in this context.

2 Related Work

Recently, there have been some attempts to develop basic NLP tools and resources for some of the prominent languages of the Indic continuum. Mundotiya *et al.* (2021) collect monolingual data and POS-tagged corpora for Bhojpuri, Maithili, and Magahi, also providing CRF baselines for POS tagging. Priyadarshi & Saha (2020) collect a monolingual corpus for Maithili as well as some POS-annotated data,[3] Ojha (2019) contribute a similar effort for Bhojpuri, and Ojha *et al.* (2020)

[1] The word "Indic", depending on context, is used to refer to both the subfamily of the Indo-European family spoken in India (e.g. Hindi, Bengali, Marathi), and the Indian languages in general (including non-Indo-European languages such as the Dravidian family of languages). In this paper, unless otherwise mentioned, we use the first, phylogenetic, sense of the term.

[2] In this work, we will refer to Braj (bra), Awadhi (awa), Bhojpuri (bho), Magahi (mag), Maithili (mai), and (standard) Hindi as "dialects" belonging to the "macrolanguage" of the dialect continuum that forms the "Hindi" heartland of India. This terminology is not intended to have political connotations.

[3] Not publicly available.

Hindi	Awadi	Bhojpuri	Magahi	Maithili	Meaning
dʑaː rəheː hoː	dʑaːt̪ əhaːi	dʑaːt̪ baː	dʑaː həi	dʑaː rəhəl ətʃ^{h} i	(you) are going
ləɖkɑː	ləɖkɑː	ləikaː	ləiːkaː	ləɖkɑː	boy (nom.)
bət̪ːaːjaː/ kəːh lijaː	bət̪ːaːvət̪	kəhəl	kəhəlieː	kəhəlhu^{n}	told (completive)
aːpkiː	aːpən	aːpən	əpən	əhaːnk	your (hon., fem. sing. obj)
bəhən	bəhin	bəhin	bəhin	bəhin	sister

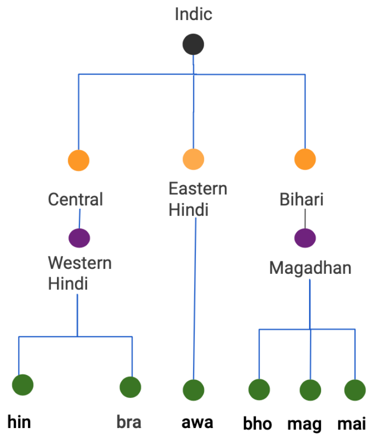
Table 1: Examples of cognates. Braj is not included due to lack of data. Since the Devanagari script is phonetically transparent, phonetic similarity is visible both in IPA and in Devanagari (not shown).

provide monolingual data and some parallel data for Bhojpuri and Magahi. As part of the NSURL 2019 shared task in POS-tagging for Bhojpuri and Magahi (Freihat & Abbas, 2019), Kumar M (2019) present an SVM-based system as well as a BERT-based classifier. Proisl *et al.* (2019) experiment with available taggers, including a BiLSTM+CRF architecture and the Stanford tagger.

There has also been work in language modeling for "dialects" of a standard variant, accompanying, of course, a rich literature in cross-lingual transfer to low-resource languages. Transformer-based pre-trained multilingual models such as mBERT (Devlin *et al.*, 2019; Conneau *et al.*, 2020) are often claimed to show multilingual generalization (Pires *et al.*, 2019). There are multiple aspects to the phenomenon of multilingual generalization, and many of them have received attention in the NLP community. One of the primary ways in which cross-lingual ability is demonstrated is through zero-shot transfer, i.e. a setting in which a multilingual pretrained model is trained on labeled or supervised data in one language and performs well in another language. Early papers found that mBERT performed remarkably well in the zero-shot setting (Pires *et al.*, 2019; Wu & Dredze, 2019) under certain conditions, such as similar typologies of source and target languages, but regardless of others, such as script and common vocabulary. Since then, many studies, such as (Chai *et al.*, 2022; Ri & Tsuruoka, 2022), have attempted to explore these conditions; notably, Muller *et al.* (2021) show that a common script indeed facilitates transfer, along with shared typological features, and Khemchandani *et al.* (2021) show the same for Indic languages. Studies are split on results regarding the relationship between subword overlap and ease of transfer. For example, K *et al.* (2020) show that shared subwords play a small role in positive transfer, while Deshpande *et al.* (2022) argue that this is only the case for languages with shared word order.

Research has also looked at the question of which languages may benefit from large multilingual models. For high-resource languages, it was quickly clear that monolingual models outperform or at least match multilingual models on most tasks (de Vries *et al.*, 2019; Martin *et al.*, 2020). However, ensuing studies have also found that monolingual models or language-family models can outperform multilingual counterparts for low-resource languages (Ulčar & Robnik-Šikonja, 2020; Ortiz Suárez *et al.*, 2020; Armengol-Estapé *et al.*, 2021; Micallef *et al.*, 2022; Barry *et al.*, 2022). In effect, there is no clear consensus in the community on the best strategies for solving a downstream task in typical conditions for a low-resource language, specifically, limited monolingual data, a related high-resource language, and possibly some annotated task data. This motivates our work in investigating different cross-lingual transfer strategies in the given context of dialects from the Indic continuum.

(a) Geography of the Indic dialect contiuum



(b) Theoretical phylogenetic tree.

Figure 1: Characterization of the Indic dialects used in this work

3 Languages

The Indic dialect continuum is a group of more than 40 dialects spoken across most of North India and surrounding regions, with hundreds of millions of speakers. The dialects of this continuum are classified under the Apabhramsic (e.g. Rajasthani), Western Hindi (e.g. Haryanvi), Eastern Hindi (e.g. Awadhi), and Bihari (e.g. Bhojpuri) branches of the Indic language family. We are working with six of these dialects, all of which are written in the Devanagari script, namely Awadhi, Bhojpuri, Braj, Magahi, Maithili, and the high-resource standardized dialect, i.e. Hindi. See Figure 1b for a phylogenetic tree of these dialects.[4] Geographically, these dialects are spread across the continuum (see Figure 1a[5]), with standard Hindi co-existing in many of these regions, although it is genetically part of the western sub-family of the continuum. This means that many of these dialects borrow from Hindi and share similarities with it due to contact, rather than via genetic transmission.

The dialects of the Indic continuum share cognates as well as morphosyntactic properties, such as a roughly common (free) word order, noun inflection for case, and verbal inflection for number and gender to a varying extent. However, they differ in specifics, for example, in the number of cases, the levels of honorifics, and the degree of inflection for gender. See Table 1 for examples of cognates in these dialects.[6]

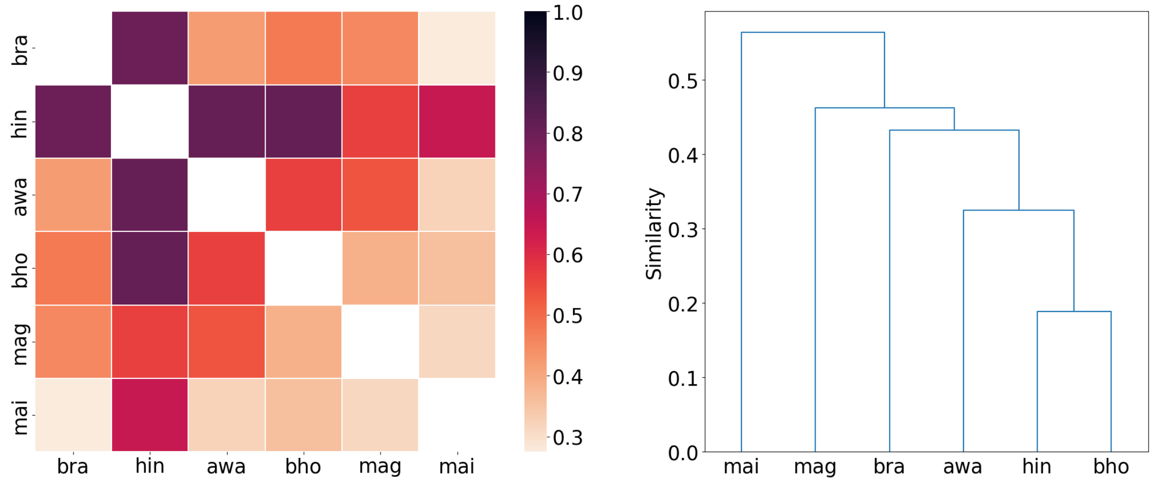
4 Data and Description

Data We use monolingual data from different (non-overlapping) sources for these dialects. These sources include the VarDial 2018 shared task (Zampieri *et al.*, 2018) for Bhojpuri, Magahi, Awadhi, and Braj, the BHLTR project for Bhojpuri (Ojha, 2019), LoResMT (Ojha *et al.*, 2020) for Bhojpuri and Magahi, and the BMM corpus (Mundotiya *et al.*, 2021) and the Wordschatz Leipzig corpus (Goldhahn *et al.*, 2012) for Maithili. For Hindi, we use the IndicCorp corpus (Kakwani *et al.*, 2020). This is the largest available consolidated Hindi corpus, as of the date of writing, and was used to

[4]Taken from Glottolog: https://glottolog.org/resource/languoid/id/midd1375.
[5]Taken from https://titus.fkidg1.uni-frankfurt.de/indexe.htm
[6]Translations are taken from Glosbe: https://glosbe.com.

	Monolingual #toks	POS #toks	POS labels #tags
awa	0.16M	21K	37
bho	2.99M	94K	34
bra	0.32M	62K	31
hin	1800.00M	351K	31
mag	3.04M	61K	19
mai	0.46M	211K	25

Table 2: Monolingual/POS dataset sizes (in #tokens) and POS tagset sizes, for all dialects.



(a) Pairwise lexical similarity for all dialects.

(b) Dendrogram, from pairwise similarities

Figure 2: Visualizations of lexical similarity among dialects.

train the publicly available pretrained Hindi model that we use for our experiments (described in Section 5).

For POS-annotated datasets, we use a single data source in a given dialect to maintain consistency in annotation style, the tagset, and its associated granularity, although these differ across dialect datasets. Specifically, we use the NSURL 2019 shared task datasets (Freihat & Abbas, 2019) for Bhojpuri and Magahi, the KMI-Linguistics datasets[7] for Awadhi and Braj, the BMM corpus for Maithili, and the Universal Dependencies Treebank HDTB project for Hindi (Palmer *et al.*, 2009; Bhat *et al.*, 2017). Aggregate token counts for each dialect are listed in Table 2.

Crosslingual interaction We would like to understand the crosslingual interactions between these dialects, in order to contextualize the results of our experiments, described in Section 5. We calculate the normalized lexical overlap[8] between monolingual corpora for all pairs of dialects (Figure 2a). The resulting similarity matrix can be used to cluster the dialects by similarity (as shown in Figure 2b); we see that the resulting tree does not resemble the gold phylogenetic tree in Figure 1b. This can have a few explanations: it is possible that lexical similarity is not a good measure of closeness—it does

[7] https://github.com/kmi-linguistics/

[8] This is calculated as the number of unique words that are common to both corpora, divided by the minimum of the number of unique words in the two corpora; thus, the measure lies between 0 and 1.

not, for example, take into account morphosyntactic similarity, or shared cognates that do not exactly match. It is also possible that the corpora we use are not representative of the true distributions of the dialects. Note that the outermost leaf in the tree, Maithili, is the only dialect whose monolingual data does not include VarDial shared task data. Finally, we can also credit widespread borrowing of words from Hindi with genetically more distant dialects having higher-than-expected lexical similarity.

The lexical similarity between monolingual pretraining corpora and the annotated datasets[9] may also be relevant in explaining our transfer results (see Figure 3a). Intuitively, low lexical similarity would indicate fewer benefits of pretraining. We note that the Hindi corpus has an almost perfect coverage of the annotated datasets of all dialects including itself. This is unsurprising given its size as well as the common subsumption of dialects on this continuum under "Hindi."

5 Experiments

We compare different multilingual or cross-lingual transfer strategies in the context of Indic dialects, evaluated on the downstream task of POS-tagging. We will refer to the two following operations: "pretraining" refers to training a model on mono/multilingual raw text, "EP" or "extended pretraining" refers to further pretraining an already pretrained model on different mono/multilingual raw text.[10] All models are finetuned and evaluated on the annotated dataset of the target dialect. Due to tagset differences, we cannot attempt zero-shot transfer without performing label alignment. Finally, we do not pretrain from scratch on Hindi data and use a publicly available Hindi pretrained model instead (Joshi, 2022).

We use the HuggingFace transformers library (Wolf *et al.*, 2020) for training and accessing publicly available pretrained models. All models (including the publicly available models that we use) have a BERT-base architecture, consisting of 12 attention heads, 12 layers, and with hidden layer size 768. Models trained from scratch on dialect data are trained on the Masked Language Modeling (MLM) objective for ~40 epochs, EP over pretrained models is performed for 15 epochs, and finetuning is performed over the best performing pretraining checkpoint for 5 epochs.[11]

Baseline As the baseline for each dialect, we use the BERT architecture described above, pretrained from scratch on monolingual dialect data, and finetuned on task data for the dialect.

Using pretrained models We report the results obtained by finetuning three publicly available large pretrained models on the POS-tagged dataset for each dialect:

- Hindi (we hereafter refer to this model as Hin-BERT) (Joshi, 2022): We use a pretrained Hindi model[12] trained on the MLM objective; we want to see how well a pretrained model in a *related mid-resource dialect* transfers to a low-resource dialect.

[9] Calculated in the same way as for corpus lexical similarity.
[10] Other works may use different terms for what we call extended pretraining, or may carry out extended pretraining in a different manner.
[11] Further finetuning did not improve performance.
[12] https://huggingface.co/l3cube-pune/hindi-bert-scratch/tree/main

- MuRIL – Multilingual Representations for Indian Languages (Khanuja *et al.*, 2021): The publicly available MuRIL model represents a *mid/high-resource related language family* model. MuRIL is trained on the MLM and Translation Language Modeling (TLM) objectives on 17 languages in total, including other Indic languages such as Marathi and Bengali, as well as genealogically unrelated languages from the Indian subcontinent, such as Tamil and Kannada.[13]

- mBERT (Devlin *et al.*, 2019):[14] Finally, we also finetune mBERT for each dialect; mBERT is trained on the MLM and Next Sentence Prediction (NSP) objectives, on 104 languages, including Indic languages, but also several other languages and language families.

We also perform extended pretraining for the MuRIL and Hin-BERT models to observe potential benefits. Specifically, these models are pretrained further with an MLM objective on the monolingual data of the target dialect, and then (like all other models) finetuned and evaluated on the target dialect. The MuRIL model was chosen over mBERT due to its better initial performance.[15]

Using related dialects One can use data in related low-resource dialects to boost the learning of shared properties and similar words. We conduct the following experiments to investigate this idea:

- Pairwise transfer (D+ft): We pretrain a BERT model from scratch on monolingual data from a low-resource source dialect, therefore excluding Hindi, and finetune and evaluate it on task-specific data of the target dialect. We do this for all possible pairs, and report the best F1 performance over all source languages for every target dialect. We would also like to draw inference from the above setup as to which dialects perform best as sources for a given dialect, in relation to their genealogical or other type of closeness to the target. In the D+ft setup, however, the results are confounded by varying amounts of monolingual data available for different dialects. Therefore, we fix the monolingual as well as the evaluation data size in tokens for all (source and target) dialects, using the minimum available dataset size (Awadhi), and repeat pairwise transfer experiments.

- All dialects together (ABBMM+ft): In the setup, we pretrain a BERT model from scratch on all available low-resource dialect data,[16] and finetune and evaluate separately on each dialect. The aim of this experiment is to investigate how far joint training on related dialects (without a high-resource dialect) can benefit the target.

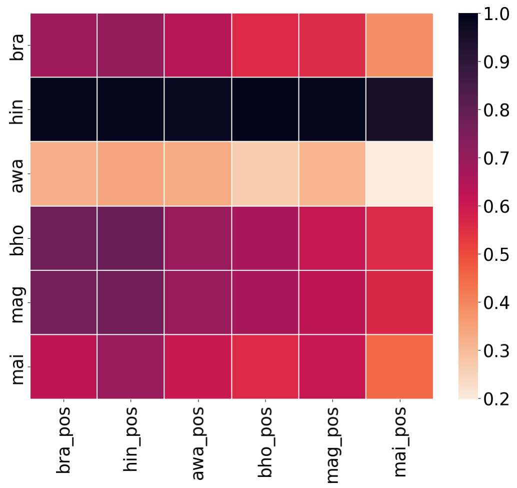
- MuRIL with EP on all dialects (MuRIL+EP$_{ABBMM}$+ft): Finally, we choose the best performing large pretrained model from the previous set of experiments (namely, MuRIL), and extended-pretrain it with all low-resource dialects, followed by separate finetuning and evaluation in each dialect.

[13] See more details here: https://huggingface.co/google/muril-base-cased/tree/main.

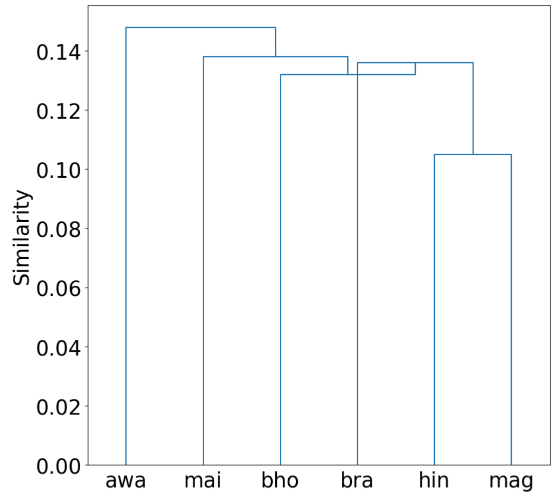
[14] https://huggingface.co/bert-base-multilingual-cased/tree/main

[15] We do not perform this experiment for Hindi, i.e. we do not do extended pretraining on Hindi data, for two reasons: firstly, both MuRIL and mBERT have already seen Hindi data, therefore rendering this a different experiment to that with the dialects, and secondly, in our work, our focus is on the low-resource dialects. We leave the exploration of best-performing transfer setups for mid-to-high range resource languages to other works.

[16] Hindi is not included in these experiments since it would easily dominate the low-resource data, and the resulting experiment would not be very different from Hin-BERT+ft, which we conduct separately.



(a) Lexical similarity of monolingual corpus with POS dataset (mono-POS overlap)



(b) Dendrogram from pairwise transfer results with fixed dataset sizes

Figure 3: Pairwise transfer results

	bho	bra	awa	hin	mag	mai
Monolingual	92.97	87.98	80.46	97.10	90.53	85.00
Hin-BERT+ft	92.13	93.31	84.05	97.10	90.89	87.98
Hin-BERT+EP$_{mono}$+ft	92.42	93.74	84.10	-	91.12	87.46
mBERT+ft	93.05	93.5	83.53	97.08	90.47	87.90
MuRIL+ft	92.96	94.14	82.31	98.01	91.25	88.54
MuRIL+EP$_{mono}$+ft	**93.62**	**94.73**	84.22	-	91.81	88.30
D+ft	91.40	92.12	82.95	96.39	90.39	87.01
ABBMM+ft	92.86	93.14	83.12	96.35	90.96	87.48
MuRIL+EP$_{ABBMM}$+ft	93.59	94.68	**85.72**	-	**91.95**	**88.69**

Table 3: Evaluation on POS-tagging (F1). EP: extended pretraining, ABBMM: all low-resource dialect data, D+ft: best performing low-resource source-to-target model per target dialect.

Target Lang	Source Language bho	bra	awa	hin	mag	mai
bho	**92.97**	87.49	87.67	92.13	91.40	87.48
bra	92.12	87.98	87.79	**93.31**	91.31	89.31
awa	82.95	80.06	80.46	**84.05**	82.43	79.88
hin	96.39	94.69	94.55	**97.10**	96.11	94.40
mag	90.39	87.97	88.06	**90.89**	90.53	87.72
mai	87.01	85.53	85.25	**87.98**	86.79	85.01

(a) Original dataset sizes

Source Language bho	bra	awa	hin	mag	mai
84.17	83.77	83.50	83.63	82.51	83.00
90.57	**93.31**	92.83	90.44	92.79	92.09
79.25	77.92	77.65	77.37	79.53	**80.19**
84.28	83.69	**87.02**	86.24	83.45	86.39
85.24	86.23	85.93	83.31	**86.87**	86.08
81.84	80.22	80.89	80.56	79.50	81.39

(b) Equal dataset sizes

Table 4: Evaluation on POS-tagging (F1), for pairwise transfer experiments.

6 Results and Discussion

Monolingual Performance Within dialects, differences in performance can be explained by the size of the annotated dataset, the amount of monolingual data, and the complexity of the tagset (see Table 2), as well as the token coverage of the annotated dataset by the monolingual data (mono-POS overlap, see Figure 3a). We see in Table 3 that the Maithili and Awadhi monolingual models perform worse than the others; they also show the least mono-POS overlap. Similarly, Braj has less monolingual and annotated data than Maithili, but still performs better, possibly because of its higher mono-POS overlap.

Bhojpuri and Magahi, the highest resourced dialects, have baseline performances roughly on par with {mBERT | MuRIL}+ft. Although these dialects are still very low-resource by several orders of magnitude compared to Hindi, their datasets are already big "enough" to yield a good performance on a relatively shallow task such POS-tagging. The transfer methods are mainly beneficial for the lowest-resourced dialects, Braj, Awadhi, and Maithili.

Pretrained models Comparing {Hin-BERT | MuRIL | mBERT}+ft, we observe that MuRIL-based models do better than mBERT-based models on four out of six dialects. This extends the results shown by Khanuja *et al.* (2021) that demonstrate that MuRIL outperforms mBERT consistently on its 17 pretraining languages. We also see that Hin-BERT and mBERT seem to perform on par, with perhaps a slight edge to Hin-BERT, although mBERT is pretrained on much more data. This corroborates the intuition that the relatedness of the pretraining languages with target languages could positively affect transfer results, "compensating", in a way, for less data. However, the fact that these pretrained models differ in their pretraining objectives must be kept in mind while making observations about the effects of pretraining languages.[17]

The increase in performance over the monolingual baseline with pretrained models, especially for Hin-BERT+ft, can also be contextualized by crosslingual lexical similarity between monolingual corpora (see Figure 2a). We see that Braj, Awadhi, and Bhojpuri show the highest lexical similarity with Hindi. This explains the jump in performance for low-resource Braj and Awadhi (whereas for Bhojpuri, which already has a good amount of monolingual data, training on Hindi causes worsening).

Extended pretraining EP helps consistently; language-specific pretraining possibly serves to expose the model to non-cognate words or language specific constructions in the target language. The only performance drop is observed for Maithili; monolingual EP slightly worsens performance in both Hin-BERT+ft and MuRIL+ft. This accords with our earlier observation of the low Maithili mono-POS overlap, and its possible effects.

Using other dialect data The best performing single-language transfer from another low-resource dialect, i.e. the D+ft model, does better than the monolingual model for dialects with little monolingual data, namely, Braj, Maithili and Awadhi, and worse for the higher resourced dialects i.e. Bhojpuri and Magahi. ABBMM+ft always does better than D+ft, presumably because the model sees monolingual data in the target dialect as well as more related dialect data in general. We also observe that

[17] Note that the training data of mBERT does include Hindi and other Indic languages; however, these are naturally accorded less "space" or percentage of training data as compared to with MuRIL or simply a Hindi pretrained model.

ABBMM+ft is on par with {mBERT | MuRIL}+ft even for dialects without much monolingual data; again, this indicates that models pretrained on roughly of a few million tokens, even from closely related dialects, perform comparably with much larger (language family or other) multilingual models (pretrained on three orders of magnitude more data) for a downstream POS-tagging task. It is possible that this is not the case for tasks requiring more language understanding.

Pairwise transfer with equal dataset size By fixing dataset sizes for all dialects, we aim to directly compare different dialects as (pretraining) sources for a given target dialect.[18] The resulting F1 scores are presented in Table 4a. We see that the differences in performance for a given dialect with different pretraining dialects are much lower than before. Interestingly, we observe that for Awadhi, Hindi, and Maithili, it is better to pretrain on a different dialect than itself. This can be partially understood in view of similarities between different dialect corpora and annotated datasets (Figure 3a), although these similarities are calculated over the full datasets rather than same-sized subsets. For example, we see that the Maithili POS-dataset has lower lexical overlap with the Maithili corpus than with the Bhojpuri corpus. A similar argument holds for Awadhi.

We also use these scores (Table 4b) to extract a dendrogram of language relatedness, with the hypothesis that this may recover a phylogenetic tree, which would mean that genetically close dialects behave similarly as sources and targets. We use the 0-1 normalized mean source-target performance of each pair of dialects as their "similarity" score[19] (Figure 3b). The resulting tree is not in fact a good representation of the phylogenetic tree of these dialects; the effect of genealogy seems to be outweighed by other factors, possibly including lexical overlap due to borrowings, and domain match.

Takeaways The takeaways from the results can be summarized as follows:

- Multilingual models pretrained on (a) data from the same language family, (b) a closely related high-resource dialect, (c) "general" multilingual data, as well as (d) low-resource closely related dialects are all good candidates for base pretrained models (to be finetuned on task data), with (a) consistently outperforming the others. Their relative performance can be interpreted as a function of the relatedness of the pretraining corpus languages, and the amount of such data.

- Among closely related dialects, the best performing source dialect pretrained model may be determined by lexical overlap or domain match with the target dialect annotated data rather than phylogenetic closeness between the source and target dialects; in general, especially for lower-resource dialects, closely-related dialect data helps performance.

- Extended pretraining, even on very little data, consistently helps. The best performing models are obtained by extended pretraining MuRIL with either monolingual data (for Bhojpuri and Braj) or all dialect data together (for Awadhi, Maithili, and Magahi).

[18] Although we do also fix the annotated dataset size, this does not mean that the downstream task is of the same difficulty for all dialects. Different datasets have different inherent difficulty due to the tagset size, length of sentences, rare words, tag distributions, etc. Therefore, it is still not advisable to make comparisons across target dialects.

[19] This clustering algorithm makes the assumption that the similarity of a dialect with itself is 1, or at least higher than that with any other dialect; we therefore ignore self-source-target scores.

7 Conclusion

In this paper, we looked at different strategies for developing language models for low-resource languages, using five extremely low-resource dialects belonging to the Indic continuum as a testbed. We compared conventional pretraining and cross-lingual transfer methods, and concluded that large pretrained models trained on the same language family (in our case MuRIL, for the Indic language family) are particularly successful as base models, especially if followed by extended pretraining, either monolingually or on closely related dialect data. We hope that this work contributes to building a basic research base for the Indic dialect continuum, as well as other dialect systems.

Acknowledgments

This work was partly funded by R. Bawden's and B. Sagot's chairs in the PRAIRIE institute funded by the French national agency ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001 and by the Emergence project, DadaNMT, funded by Sorbonne Université. The work was also supported by the German Research Foundation (Deutsche Forschungsgemeinschaft) under grant SFB 1102: Information Density and Linguistic Encoding.

References

ARMENGOL-ESTAPÉ J., CARRINO C. P., RODRIGUEZ-PENAGOS C., DE GIBERT BONET O., ARMENTANO-OLLER C., GONZÁLEZ-AGIRRE A., MELERO M. & VILLEGAS M. (2021). Are Multilingual Models the Best Choice for Moderately Under-resourced Languages? A Comprehensive Assessment for Catalan. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, p. 4933–4946.

BAFNA N., VAN GENABITH J., ESPAÑA-BONET C. & ŽABOKRTSKÝ Z. (2022). Combining Noisy Semantic Signals with Orthographic Cues: Cognate Induction for the Indic Dialect Continuum. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, p. 110–131, Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics.

BARRY J., WAGNER J., CASSIDY L., COWAP A., LYNN T., WALSH A., Ó MEACHAIR M. J. & FOSTER J. (2022). gaBERT — an Irish Language Model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 4774–4788, Marseille, France: European Language Resources Association.

BHAT R. A., BHATT R., FARUDI A., KLASSEN P., NARASIMHAN B., PALMER M., RAMBOW O., SHARMA D. M., VAIDYA A., VISHNU S. R. *et al.* (2017). The Hindi/Urdu Treebank Project. In *Handbook of Linguistic Annotation*. Springer Press.

CHAI Y., LIANG Y. & DUAN N. (2022). Cross-Lingual Ability of Multilingual Masked Language Models: A Study of Language Structure. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 4702–4712, Dublin, Ireland: Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.322.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8440–8451, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.747.

DE VRIES W., VAN CRANENBURGH A., BISAZZA A., CASELLI T., VAN NOORD G. & NISSIM M. (2019). BERTje: A Dutch BERT Model. arXiv:1912.09582 [cs], DOI : 10.48550/arXiv.1912.09582.

DESHPANDE A., TALUKDAR P. & NARASIMHAN K. (2022). When is BERT Multilingual? Isolating Crucial Ingredients for Cross-lingual Transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 3610–3623, Seattle, United States: Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.264.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota: Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DHAMECHA T., MURTHY R., BHARADWAJ S., SANKARANARAYANAN K. & BHATTACHARYYA P. (2021). Role of Language Relatedness in Multilingual Fine-tuning of Language Models: A Case Study in Indo-Aryan Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 8584–8595, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.675.

FREIHAT A. A. & ABBAS M., Éds. (2019). *Proceedings of the First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019 - Short Papers*. Trento, Italy: Association for Computational Linguistics.

GOLDHAHN D., ECKART T. & QUASTHOFF U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, p. 759–765, Istanbul, Turkey: European Language Resources Association (ELRA).

JOSHI R. (2022). L3Cube-HindBERT and DevBERT: Pre-Trained BERT Transformer models for Devanagari based Hindi and Marathi Languages. *arXiv preprint arXiv:2211.11418*.

K K., WANG Z., MAYHEW S. & ROTH D. (2020). Cross-Lingual Ability of Multilingual BERT: An Empirical Study. In *International Conference on Learning Representations*. DOI : 10.48550/arXiv.1912.07840.

KAKWANI D., KUNCHUKUTTAN A., GOLLA S., N.C. G., BHATTACHARYYA A., KHAPRA M. M. & KUMAR P. (2020). IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, p. 4948–4961, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.445.

KALYAN K. S., RAJASEKHARAN A. & SANGEETHA S. (2021). AMMUS: A Survey of Transformer-based Pretrained Models in Natural Language Processing. arXiv:2108.05542 [cs], DOI : 10.48550/arXiv.2108.05542.

KHANUJA S., BANSAL D., MEHTANI S., KHOSLA S., DEY A., GOPALAN B., MARGAM D. K., AGGARWAL P., NAGIPOGU R. T., DAVE S. *et al.* (2021). Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

KHEMCHANDANI Y., MEHTANI S., PATIL V., AWASTHI A., TALUKDAR P. & SARAWAGI S. (2021). Exploiting Language Relatedness for Low Web-Resource Language Model Adaptation: An Indic Languages Study. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, p. 1312–1323, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.105.

KUMAR M A. (2019). NITK-IT_NLP@NSURL2019: Transfer Learning based POS Tagger for Under Resourced Bhojpuri and Magahi Language. In *Proceedings of the First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019 - Short Papers*, p. 68–72, Trento, Italy: Association for Computational Linguistics.

LIN Y.-H., CHEN C.-Y., LEE J., LI Z., ZHANG Y., XIA M., RIJHWANI S., HE J., ZHANG Z., MA X., ANASTASOPOULOS A., LITTELL P. & NEUBIG G. (2019). Choosing Transfer Languages for Cross-Lingual Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 3125–3135, Florence, Italy: Association for Computational Linguistics. DOI : 10.18653/v1/P19-1301.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE E., SEDDAH D. & SAGOT B. (2020). CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

MICALLEF K., GATT A., TANTI M., VAN DER PLAS L. & BORG C. (2022). Pre-training Data Quality and Quantity for a Low-Resource Language: New Corpus and BERT Models for Maltese. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, p. 90–101. arXiv:2205.10517 [cs], DOI : 10.18653/v1/2022.deeplo-1.10.

MULLER B., ANASTASOPOULOS A., SAGOT B. & SEDDAH D. (2021). When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 448–462, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.38.

MUNDOTIYA R. K., SINGH M. K., KAPUR R., MISHRA S. & SINGH A. K. (2021). Linguistic Resources for Bhojpuri, Magahi, and Maithili: Statistics about Them, Their Similarity Estimates, and Baselines for Three Applications. *ACM Transactions on Asian and Low-Resource Language Information Processing*, **20**(6), 95:1–95:37. DOI : 10.1145/3458250.

OJHA A. K. (2019). English-Bhojpuri SMT System: Insights from the Karaka Model. arXiv:1905.02239 [cs], DOI : 10.48550/arXiv.1905.02239.

OJHA A. K., MALYKH V., KARAKANTA A. & LIU C.-H. (2020). Findings of the LoResMT 2020 Shared Task on Zero-Shot for Low-Resource languages. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, p. 33–37, Suzhou, China: Association for Computational Linguistics.

ORTIZ SUÁREZ P. J., ROMARY L. & SAGOT B. (2020). A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 1703–1714, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.156.

PALMER M., BHATT R., NARASIMHAN B., RAMBOW O., SHARMA D. M. & XIA F. (2009). Hindi syntax: Annotating dependency, lexical predicate-argument structure, and phrase structure. In *The 7th International Conference on Natural Language Processing*, p. 14–17.

PATIL V., TALUKDAR P. & SARAWAGI S. (2022). Overlap-based Vocabulary Generation Improves Cross-lingual Transfer Among Related Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 219–233, Dublin, Ireland: Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.18.

PIRES T., SCHLINGER E. & GARRETTE D. (2019). How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 4996–5001, Florence, Italy: Association for Computational Linguistics. DOI : 10.18653/v1/P19-1493.

PRIYADARSHI A. & SAHA S. K. (2020). Towards the first Maithili part of speech tagger: Resource creation and system development. *Computer Speech & Language*, **62**, 101054. DOI : 10.1016/j.csl.2019.101054.

PROISL T., UHRIG P., BLOMBACH A., DYKES N., HEINRICH P., KABASHI B. & MAMMARELLA S. (2019). The_Illiterati: Part-of-Speech Tagging for Magahi and Bhojpuri without Even Knowing the Alphabet. In *Proceedings of The First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019-Short Papers*, p. 73–79.

RI R. & TSURUOKA Y. (2022). Pretraining with Artificial Language: Studying Transferable Knowledge in Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 7302–7315, Dublin, Ireland: Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.504.

ULČAR M. & ROBNIK-ŠIKONJA M. (2020). FinEst BERT and CroSloEngual BERT. In P. SOJKA, I. KOPEČEK, K. PALA & A. HORÁK, Éds., *Text, Speech, and Dialogue*, p. 104–111, Cham: Springer International Publishing.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., LE SCAO T., GUGGER S., DRAME M., LHOEST Q. & RUSH A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, p. 38–45, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-demos.6.

WU S. & DREDZE M. (2019). Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 833–844, Hong Kong, China: Association for Computational Linguistics. DOI : 10.18653/v1/D19-1077.

ZAMPIERI M., NAKOV P., LJUBEŠIĆ N., TIEDEMANN J., MALMASI S. & ALI A., Éds. (2018). *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*. Santa Fe, New Mexico, USA: Association for Computational Linguistics.