Investigating Lexical Sharing in Multilingual Machine Translation for Indian Languages

Sonal Sannigrahi
Saarland University, Saarland Informatics Campus
Saarbrücken, Germany
sosa00001@stud.uni-saarland.de

Rachel Bawden
Inria, Paris, France
rachel.bawden@inria.fr

Abstract

Multilingual language models have shown impressive cross-lingual transfer ability across a diverse set of languages and tasks. To improve the cross-lingual ability of these models, some strategies include transliteration and finer-grained segmentation into characters as opposed to subwords. In this work, we investigate lexical sharing in multilingual machine translation (MT) from Hindi, Gujarati, Nepali into English. We explore the trade-offs that exist in translation performance between data sampling and vocabulary size, and we explore whether transliteration is useful in encouraging *cross-script* generalisation. We also verify how the different settings generalise to unseen languages (Marathi and Bengali). We find that transliteration does not give pronounced improvements and our analysis suggests that our multilingual MT models trained on original scripts seem to already be robust to cross-script differences even for relatively low-resource languages. Our code will be made publicly available.[1]

1 Introduction

As research in natural language processing (NLP) moves towards handling an increasing number of languages (Aharoni et al., 2019; Fan et al., 2021), one of the key challenges is targeting low-resource and morphologically rich languages (Johnson et al., 2017; Magueresse et al., 2020). Multilingual language models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) have shown surprising cross-lingual ability in zero and few-shot scenarios for a diverse set of languages (Wu and Dredze, 2020).

In order for low-resource languages to optimally benefit from data available for related and higher-resource languages, one research direction has been to explore what encourages better cross-lingual sharing in multilingual models, particularly in models that have joint vocabularies (Ha et al., 2016; Johnson et al., 2017; Aharoni et al., 2019).

One strategy for doing this is to preprocess the texts to reduce variation linked to differences in script and orthographic conventions, for example phonetisation, transliteration and transcription, in order to increase lexical overlap across languages. These pre-processing steps have been used in the literature across several multilingual NLP tasks (Nakov and Tiedemann, 2012; Nguyen and Chiang, 2017; Chakravarthi et al., 2019; Goyal et al., 2020; Sun et al., 2022; Muller et al., 2021; Alabi et al., 2022). However, there is still some debate over how much transliteration helps in multilingual setups, despite it theoretically encouraging better lexical overlap, particularly for low-resource languages. For example, Pires et al. (2019) found that transfer may be helped by increased lexical overlap (although it also works without it) and K et al. (2020) argue that lexical overlap has a negligible impact on transfer. Chakravarthi et al. (2019) and Muller et al. (2021) found gains when transliterating, whereas for Alabi et al. (2022), results were less clear.

In this study, we build on this previous work to further investigate how lexical overlap can help multilingual machine translation (MT) by taking as a case study several Indian languages. Figure 1 il-

[1] https://github.com/sonalsannigrahi/Multilingual_Strategy

HI: वह लंबे समय से राजनीति के बारे में बात कर रहे हैं।
GU: તેઓ ઘણા સમયથી રાજનીતિ વિશે વાત કરી રહ્યા છે.
NE: उनी लामो समयदेखि राजनीतिको कुरा गर्दै आएका छन्।
BE: উনি রাজনীতির সম্পর্কে অনেক ক্ষণ ধরে কথা বলছেন।
MR: ते बरेच दिवस राजकारणावर बोलत आहेत.

Gloss: He long time politics about talk doing.
EN: He has been talking about politics for a long time

Figure 1: Illustration of partial lexical overlap in different scripts and languages (Hindi, Gujarati, Nepali, Bengali, Marathi). Highlighted text is an exact phonetic match at word or partial word coverage level.

lustrates the degree of lexical overlap in the chosen languages of study: Hindi, Gujarati, Nepali, Bengali, and Marathi. Despite script differences, this example shows a sizeable amount of shared token overlap in terms of both characters and words.

Focusing on the translation of these languages (Hindi, Gujarati, Nepali) into English, we explore the ideal parameter settings for multilingual MT (sampling vs. segmentation size) and look at how transliterating into a single script (i.e. Gujarati into Devanagari) may help performance. In addition, we look at how the trained models can transfer to other related languages (Bengali and Marathi) in zero- and few-shot settings. We find that transliteration does not significantly help performance in our multilingual MT setup, even for the lowest-resourced language directions. Our analysis suggests that even with relatively little data, the multilingual model trained on the original scripts seems to learn a sufficient mapping between original and transliterated tokens, possibly making transliteration redundant. Even in zero- and few-shot transfer settings, we find only marginal improvements in the languages considered by using the multilingual model that uses transliteration as opposed to the multilingual model with the original scripts.

2 Related Work

Multilingual models have been proposed for MT as well as other NLP tasks (Doddapaneni et al., 2021). Within multilingual models, the promotion of lexical sharing has been the primary motivation to train multilingual models, which can especially aid low-resource languages (Conneau et al., 2020).

The choice of input unit has received a lot of attention, from the use of joint multilingual vocabularies (Sennrich et al., 2016a; Ha et al., 2016; Johnson et al., 2017; Aharoni et al., 2019) and subword segmentation strategies (Sennrich et al., 2016b; Kudo and Richardson, 2018) to character-based (Kreutzer and Sokolov, 2018) and byte-based (Xue et al., 2022) models. Other works have explored phonetisation (Liu et al., 2019; Rosales Núñez et al., 2019) and transliteration/transcription in order to create a higher degree of lexical overlap in related languages that do not shared scripts (Nakov and Tiedemann, 2012; Nguyen and Chiang, 2017; Chakravarthi et al., 2019; Goyal et al., 2020; Muller et al., 2021; Alabi et al., 2022).

Cross-lingual word embedding spaces have been of interest as well. Chronopoulou et al. (2021) map separately learnt embeddings to the same space, and other related works attempt to jointly learn a shared embedding space for multiple languages. Cross-lingual transfer studies on multilingual models such as mBERT (Devlin et al., 2019) have also shown the utility of multilingual pre-training especially for zero-shot transfer (Pires et al., 2019). They show that overlap can lead to better zero-shot transfer, although there can still be transfer with no overlap, as also seen by K et al. (2021). Wu and Dredze (2020) also see a positive correlation between lexical overlap and the zero-shot transfer performance. Additionally, (Oladipo et al., 2022) experiment with effect of shared vocabulary spaces in multilingual setups for several low-resource African languages (Amharic, Hausa, and Swahili) and find that the number of languages used during pre-training has a positive effect on cross-lingual transfer only up to a certain point- which is improved by simply using a monolingual model with a multilingual tokeniser.

Variation in data availability, scripts, and morpho-syntactic properties make adapting multilingual models to unseen languages challenging. Transliteration, which directly encourages lexical overlap, has shown positive results for texts in different scripts (Muller et al., 2021; Chakravarthi et al., 2019). Muller et al. (2021) show that script plays a crucial role in improving transferability of multilingual models for languages that otherwise lag behind in performance. However, Alabi et al. (2022) find that transcription (for Slavic languages) degraded rather than aided performance, with the hypothesis that the high-resource setup made transcription unnecessary, especially given the noise introduced by transcription. In our work, we study the role of transliteration in the case of multilingual MT for a set of lower-resource language directions, using related Indian languages with script differences.

3 Background on the Languages of Study

Hindi, Nepali, Gujarati, Bengali, and Marathi are all Indo-Aryan languages, a sub-branch of the Indo-European language family, with speakers primarily concentrated in the Indian subcontinent. Hindi (excluding Urdu)[2] is spoken by approximately 340M L1 speakers (and 600M L1 or L2 speakers) and is considered to be the largest in terms of L1 speakers, whereas Nepali, Gujarati, Bengali, and Marathi have 16M, 57M, 272M, and 99M L1 speakers respectively.[3] Hindi, Nepali, and Marathi share the same script (Devanagari) and also certain morpho-syntactic properties such as split ergativity and Subject-Object-Verb word order with constraint-based reordering allowed. Gujarati and Bengali each use their own scripts, although they are still considered closely related to the other Indo-Aryan languages, with both lexical and grammatical similarities. In particular, in both languages there exist many words that are an exact phonetic match with Hindi due to direct borrowing from Sanskrit. Due to these properties and the fact that the writing systems correspond well to the phonetic systems, transliteration from either the Gujarati and Bengali script into Devanagari is mostly straightforward (see Figure 3 for an example).

4 Experiments

We study the effect of transliteration for multilingual MT to test the hypothesis that increased lexical overlap between the training languages could boost performance, particularly for lower-resourced language pairs. We study two different scenarios: (i) an *in-language* scenario, whereby we train and evaluate on the same set of language pairs, namely Hindi (hi), Nepali (ne), and Gujarati (gu) into English, and (ii) zero- and few-shot transfer (via fine-tuning) of these models to two unseen related language pairs, namely Marathi (mr) and Bengali (bn) into English. We compare models trained on the original scripts and after transliteration (i.e. Gujarati is transliterated into Devanagari).

Since the aim of transliteration is to increase lexical overlap between the languages, we make sure to monitor for the degree of tokenisation, as well as data sampling, both crucial parameters in multilingual MT performance that directly affect token overlap, to ensure a fair comparison.

4.1 Data

The chosen languages cover a variety of scripts (Devanagari, Gujarati, and Bengali) as illustrated in Figure 1. Table 1 lists the data sources and sizes used (ranging from 65k sentences for gu–en to 1M sentences for hi–en after post-processing).

We clean the data by normalising punctuation, and removing duplicate sentence pairs from the training data. For experiments involving transliteration, we use the IndicNLP toolkit[6] (Kunchukuttan, 2020) to transliterate Gujarati and Bengali scripts into the Devanagari script. For subword segmentation, we use the Sentencepiece toolkit (Kudo and Richardson, 2018) and the BPE strategy (Sennrich et al., 2016b) to train joint models covering the specific training languages for each model, i.e. the source and target language for bilingual models and Hindi, Gujarati, Nepali and English for the multilingual ones. We test a range of vocabulary sizes: 4k, 8k, 16k and 32k for the multilingual models and 4k, 8k, 10k for the bilingual models.[7]

Due to differences in the amount of data available, we use temperature sampling to address imbalances (Fan et al., 2021). We sample data with probability p_l from each language pair, l with D_l size parallel corpora, included in the data during training of the SentencePiece models and the training of the multilingual MT model as follows:

$$p_l \propto (\frac{D_l}{\sum_k D_k})^{\frac{1}{T}},$$

where T corresponds to the temperature, which adjusts how much the original distribution is favoured (T=1) versus a more uniform distribution of the data (higher T value) as illustrated in Figure 2.

We test the temperature values 1.2, 1.5 and 1.8.[8]

4.2 Models

We train multilingual models for Hindi, Gujarati, and Nepali into English for the vocabulary sizes and

[2] We exclude Urdu in the speaker counts, since Hindi and Urdu, although nearly identical phonetically, are written in different scripts (Devanagari and Arabic script respectively). This is an important distinction given that we focus on transliteration.

[3] Figures from Ethnologue, `https://www.ethnologue.com/insights/ethnologue200/`.

[4] (Kunchukuttan et al., 2018)

[5] (Christodouloupoulos and Steedman, 2015)

[5] (Reimers and Gurevych, 2020)

[6] `https://github.com/anoopkunchukuttan/indic_nlp_library`

[7] Preliminary experiments showed that larger vocabulary sizes degraded the performance.

[8] Preliminary experiments showed that more extreme (higher) values worked less well, despite these being used previously in the literature (Aharoni et al., 2019).

	Data sources			#sentences		
	Train	Dev	Test	Train	Dev	Test
hi–en	Wikititles, HindEnCorp, IITB[4]	WMT-dev14	WMT-test14	1.3M	520	2,507
ne–en	Bible,[5] Ted2020,[6] QED, GlobalVoices, GNOME, KDE	Flores-dev	Flores-devtest	115k	997	1,012
gu–en	Bible, Wiki, Wikititles, Govin-clean, localisation	WMT-dev19	WMT-test19	70k	997	1,012
mr–en	Bible-UEDIN, cvit-pib, jw, PMI, Ted2020, Wikimatrix	Flores-dev	Flores-devtest	330k	997	1,012
be–en	alt, cvit-pib, jw, OpenSubtitles, PMI, Tanzil, Ted2020, Wikimatrix	Flores-dev	Flores-devtest	86k	997	1,012

Table 1: Data sources and dataset sizes for each language pair.

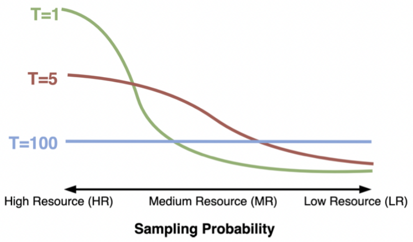


Figure 2: Illustration of data distribution with temperature sampling, taken from (Arivazhagan et al., 2019).

temperatures specified in Section 4.1, comparing models using (i) the original scripts and (ii) when Gujarati is transliterated into Devanagari (i.e. all sources languages use Devanagari). We compare these models to bilingual baselines for each of the three main language pairs, trained in the same way but only with the source and target languages concerned.

All models are transformers as implemented in Fairseq (Ott et al., 2019). We use the following default parameters unless stated otherwise:[9] 6 encoder and decoder layers with 512 embedding dimension, 2048 FFN embedding dimension, and 8 heads for both the encoder and decoder. For the multilingual models, we use a shared encoder to promote language sharing. All models are trained using the Adam optimiser with a learning rate of $3e{-}5$. All the models, multilingual and bilingual, use the same hyperparameters. Models are trained until convergence and the best model is selected according to the BLEU score on the development set. We evaluate using BLEU (Papineni et al., 2002) using the SacreBLEU toolkit (Post, 2018).[10]

5 Results

The main results are shown in Table 2a for bilingual models and Table 2b for multilingual models.

5.1 Does multilinguality help?

We start by evaluating whether multilinguality helps by comparing the models trained on original scripts. Tables 2a and 2b summarise these results for each of the language directions considered (hi→en, gu→en, ne→en). For the lower-resourced pairs, the bilingual MT models perform poorly (less than 5 BLEU points). However, these scores are greatly improved in the multilingual MT model (ne→en and gu→en achieve 12.52 and 11.82 BLEU respectively as the highest scores across all configurations tested). This performance jump demonstrates the large gains that can be observed via knowledge transfer in multilingual models, confirming previous work (Dabre et al., 2020).

In terms of temperature and vocabulary size, our multilingual results are coherent with the existing literature (Cherry et al., 2018; Kreutzer and Sokolov, 2018), which suggests that using smaller sub-word tokens perform better in low-resource settings due to their improved ability to generalise; for the lower-resource language pairs ({ne,gu}→en) a higher temperature and smaller vocabulary size combination was preferred,[11] while for the higher-resource language pair (hi→en) a lower temperature and larger vocabulary size combination was better.[12]

5.2 Is Transliteration Useful?

Our hypothesis was that by transliterating Gujarati into the Devanagari script, we might be able to see gains through increased lexical sharing amongst the three source languages in a multilingual setup.

As a control experiment to test the impact of transliteration outside of the multilingual setup, we compare results for the bilingual model using the original Gujarati script and when transliterated into Devanagari script (Table 2a *Transliterated*). The transliterated model performs slightly worse than

[9] `https://github.com/facebookresearch/fairseq`

[10] Signature =`nrefs:1|case:mixed|eff:no|tok:13a smooth:exp|version:2.0.0`

[11] 4k vocabulary size, T=1.8.

[12] 32k vocabulary size, T=1.2.

Vcb.	gu→en	hi→en	ne→en
	Original		
4k	3.87	10.12	2.06
8k	3.95	10.44	2.33
10k	4.12	12.32	2.37
	Transliterated		
4k	3.48	–	–
8k	3.68	–	–
10k	4.11	–	–

(a) Bilingual models.

	gu→en			hi→en			ne→en		
Temp.	1.2	1.5	1.8	1.2	1.5	1.8	1.2	1.5	1.8
Vcb. ↓	*Original*								
char	11.30	11.45	11.63	14.78	15.12	15.64	11.02	10.46	10.89
4K	11.10	11.40	***11.82***	15.03	14.14	14.34	11.12	**12.10**	***12.52***
8K	**11.46**	**11.69**	11.58	15.01	14.60	14.66	**11.85**	11.80	11.79
16K	11.42	9.99	11.59	15.11	14.70	**14.78**	11.73	10.44	11.56
32K	11.37	11.11	11.01	***15.32***	**14.76**	14.57	11.60	11.20	11.31
	Transliterated								
char	**11.67**	**11.82**	***11.96***	12.78	13.35	13.41	10.87	11.21	11.30
4K	11.42	11.65	11.78	13.32	13.28	13.61	**12.23**	**12.52**	***12.56***
8K	11.21	11.34	11.68	13.28	13.56	13.55	11.32	11.50	11.87
16K	11.12	11.46	11.54	**13.10**	***14.38***	**14.33**	11.11	11.24	11.73
32K	11.00	11.08	11.56	13.14	13.44	13.75	11.10	11.20	11.65

(b) Multilingual models.

Table 2: BLEU scores for bilingual baseline and multilingual models (original and transliterated) for different vocab sizes (Vcb.) and temperature values (for multilingual models only) averaged over three runs with different starting seeds. Bold represent the best score for each temperature, italics represents best score overall.

the original bilingual model (0.24% decrease between the highest scores) suggesting that transliteration may be introducing ambiguity or noise, as also suggested by Alabi et al. (2022). For the multilingual models (Table 2b), in the case of hi→en (the highest-resourced language) transliteration leads to a 8.6% decrease in the BLEU score. This decrease does not appear for gu→en and ne→en, where instead marginal improvements of 0.08 and 0.04 BLEU between the highest scores respectively are observed. However this improvement is not as large as suggested by some previous work (Muller et al., 2021). The results here could suggest that the original model might be sufficiently capturing the same level of information regarding token overlap as transliteration.

Overall compared to the original model in both the bilingual and multilingual setup, we find the improvements from transliteration (when applicable) to be not as pronounced.

5.3 Mapping Tokens in the Multilingual Embedding Space

The lack of significant improvement in in-language performance for the transliterated model is in line with results seen by Alabi et al. (2022), but is more surprising given that we test on two lower-resourced language pairs. So does this mean that the original model is already able to map between tokens written in different scripts?

To test this, we look at the similarity of tokens that are phonetically equivalent aside from being written in different scripts. Figure 3 shows some examples of Gujarati and Devanagari characters and (for illustration purposes) their romanised phonetic equivalents. Figure 4 illustrates the embedding projection of the original multilingual model (16k vocab size, T=1). We use PCA to perform dimension-reduction, and we use 10000 tokens from the vocabulary to learn the embedding space. We observe that phonetically equivalent tokens in the Devanagari and Gujarati scripts are mapped reasonably close together in this embedding space suggesting that despite script differences, the model seems to have learnt similar representations.

ક-क-ka ગ-ग-ga
ચ-च-cha જ-ज-ja
છ-छ-chha ઘ-घ-gha

Figure 3: Examples of six consonants and their realisation in Gujarati, Devanagari and Latin scripts.

	Hindi					
Gujarati ↓	Pa	Ma	Da	Ka	Fa	Avg.
Pa	**0.73**	0.12	0.02	0.14	0.02	0.01
Ma	0.18	**0.75**	0.05	0.20	0.13	0.04
Da	0.02	0.25	**0.35**	0.26	0.03	0.02
Ka	0.15	0.26	0.02	**0.66**	0.01	0.03
Fa	0.02	0.25	0.12	0.20	**0.45**	0.01
Avg.	0.01	0.02	0.02	0.01	0.03	-

Table 3: Cosine similarity scores between phonetically identical units in Devanagari (horizontal) and Gujarati (vertical) scripts with an average score (Avg.) between all other tokens.

5.4 Cross-script Robustness

We additionally experiment with cross-script switching to test how robust the original multilingual model is to changes in the script being used,

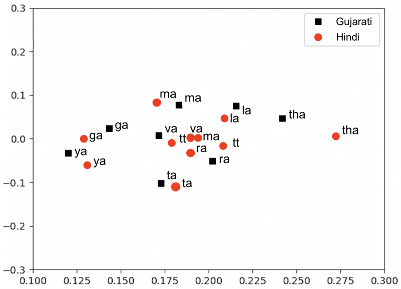


Figure 4: PCA projection of the multilingual embedding space (Original, 16k vocabulary size, T=1.5), where labelled points are a selection of phonetically equivalent tokens in Devanagari script (red dots) and Gujarati script (black squares).

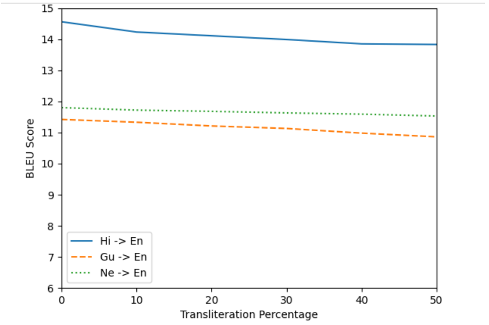
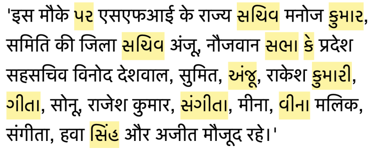


Figure 6: BLEU scores of the multilingual model (8k vocab, T=1.5) with an increasing percentage of cross-script switching.

as it appears to provide reasonably similar mappings between the same tokens written in different scripts. We artificially create texts with increasing percentages of transliteration into a different script seen by the model and evaluate the model at inference on these texts in a zero-shot fashion. For Devanagari text (in Hindi and Nepali), we transliterate parts of the text into Gujarati and vice versa. We randomly select a certain percentage of words to transliterate in each sentence. Figure 5 shows an example of cross-script switching for Hindi with 30% of words transliterated into Gujarati. We plot the BLEU scores of the different model configurations against the percentage of word-level transliteration in the test set in Figure 6. For brevity, we only plot results with $T = 1.5$ and subword vocabulary size of 16k tokens in the original multilingual model that keeps the scripts as they are.[13]



'इस मौके પર एसएफआई के राज्य સચિવ मनोज કુમાર,
समिति की जिला સચિવ अंजू, नौजवान સભા કે प्रदेश
सहसचिव विनोद देशवाल, सुमित, અંજૂ, राकेश કુમારી,
ગીતા, सोनू, राजेश कुमार, સંગીતા, मीना, વીના मलिक,
संगीता, हवा સિંહ और अजीत मौजूद रहे।'

Figure 5: Example of Hindi text in Devanagari script with 30% of words transliterated into Gujarati script (highlighted).

Although there is a downward trend in the BLEU scores, there is no significant degradation in performance with increasingly transliterated texts (only -0.2 BLEU with 50% transliteration for gu→en). The degradation of performance in the case of Hindi is more pronounced (-0.7 BLEU with 50% transliteration for hi→en). It is to be noted that in the earlier experiments (Table 2b) we found similar performance drops in Hindi between the original multilingual model and the transliterated multilingual model. This suggests that transliteration may not be a particularly useful strategy to promote lexical sharing as the models appear to already be reasonably robust to script differences.

5.5 How Well do Models Generalise to Unseen Languages?

Lastly, we study the models' ability to generalise to previously unseen but related languages. Adelani et al. (2022) find that the most effective strategy for transferring to additional languages is to use a small quantities of high-quality data. In our case, we do not fine-tune a large pre-trained language model but rather a multilingual translation model trained on Hindi, Nepali, Gujarati, and English. We therefore expect gains to be more limited than those demonstrated in (Adelani et al., 2022).

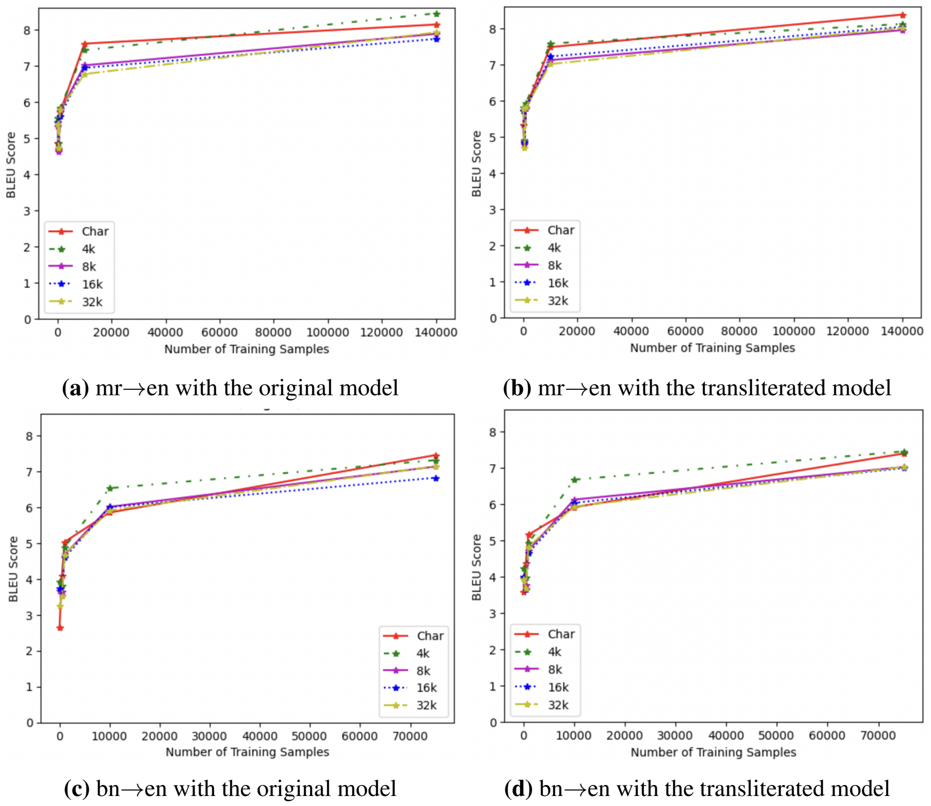
We evaluate zero-shot and few-shot transfer from the multilingual models with and without transliteration into two languages that share morphological similarities with the previous languages: Marathi (written with the Devanagari Script) and Bengali (written with the Bengali Script).[14] In this setup we incrementally increase the amount of data used to fine-tune different models (zero-shot and 500, 1k, and 10k samples for the few-shot settings). We also include a topline in which we finetune the same models on all the available data (140k sentence pairs for mr→en and 75k sentence pairs for bn→en). Figure 7 summarises our results. The raw results are in Appendix A.

The results of the zero-shot performance of the configurations illustrated[15] show that there is mini-

[13] We observe similar results across the other temperature-vocabulary size configurations.

[14] Across all models (original and transliterated) we first transliterate Bengali into Devanagari script in order to use the learned representations of the model. We leave Marathi in its original script (Devanagari)

[15] We plot the best result for each vocabulary size in char, 4k, 8k, 16k, 32k



(a) mr→en with the original model

(b) mr→en with the transliterated model

(c) bn→en with the original model

(d) bn→en with the transliterated model

Figure 7: BLEU scores after fine-tuning on different amounts of supervised training data (starting with zero-shot performance, i.e. no language-pair-specific data) for both the original multilingual model and after transliteration with varous across mvocabulary sizes: char, 4k, 8k, 16k and 32k. Only the best performing temperature value is plotted for clarity and space reasons.

mal generalisation of our multilingual model (original and transliterated) to new languages, despite their linguistic relatedness, with BLEU scores under 6 for both language directions. Using transliteration, the zero-shot transfer results are marginally improved (an increase from 5.56 to 5.81 BLEU for mr→en and from 3.93 to 4.23 BLEU for bn→en when using the transliterated rather than original model).

In the few-shot setup, similar to the results in Section 5.1 for the lower-resourced language pairs, smaller vocabulary sizes and higher temperature values are preferred (T=1.8 and either 5k or character-based segmentation). As with the zero-shot setup, marginal improvements with transliteration are observed in the few-shot setup. This result agrees with our earlier results (Section 5.2), which show that transliteration does not provide significant gains, possibly as the original multilingual model is already robust to cross-script differences.

6 Conclusions

In this work, we studied language sharing in multilingual MT of several languages in the Indo-Aryan language family (Gujarati, Nepali, and Hindi into English). Experimenting with sampling temperature and vocabulary size, we compare multilingual models using the original scripts and when transliterating Gujarati into the same script as Nepali and Hindi (Devanagari). Surprisingly, even for the low-resource language directions (gu→en and ne→en), we find that transliteration is not particularly helpful. It seems that our multilingual models using the original scripts are able to correctly map phonetically equivalent tokens together, as suggested by (i) our analysis of the embeddings of identical characters across scripts and (ii) testing the robustness of the model to cross-script switching. Finally, we test how well the models transfer to unseen related languages (Marathi and Bengali into English). We find that the model with transliteration does not perform significantly better with respect to generalisation to unseen languages, further supporting our previous findings.

7 Acknowledgments

R. Bawden's participation was partly funded by her chair position 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. The work was also funded by R. Bawden's Emergence project, DadaNMT, funded by Sorbonne Université.

References

Adelani, David, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States, July. Association for Computational Linguistics.

Aharoni, Roee, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. 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)*, pages 3874–3884, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Alabi, Jesujoba O, Lydia Nishimwe, Benjamin Muller, Camille Rey, Benoît Sagot, and Rachel Bawden. 2022. Inria-ALMAnaCH at the WMT 2022 shared task: Does Transcription Help Cross-Script Machine Translation? In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Arivazhagan, Naveen, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Chakravarthi, Bharathi Raja, Mihael Arcan, and John P McCrae. 2019. Comparison of different orthographies for machine translation of under-resourced dravidian languages. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Cherry, Colin, George Foster, Ankur Bapna, Orhan Firat, and Wolfgang Macherey. 2018. Revisiting character-based neural machine translation with capacity and compression. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4295–4305, Brussels, Belgium, October-November. Association for Computational Linguistics.

Christodouloupoulos, Christos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.

Chronopoulou, Alexandra, Dario Stojanovski, and Alexander Fraser. 2021. Improving the lexical ability of pretrained language models for unsupervised neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 173–180, Online, June. Association for Computational Linguistics.

Conneau, Alexis, 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, July. Association for Computational Linguistics.

Dabre, Raj, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)*, 53(5):1–38.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Doddapaneni, Sumanth, Gowtham Ramesh, Mitesh M Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2021. A primer on pretrained multilingual language models. *arXiv preprint arXiv:2107.00676*.

Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond English-Centric Multilingual Machine Translation. *Journal of Machine Learning Research*, 22(107):1–48.

Goyal, Vikrant, Sourav Kumar, and Dipti Misra Sharma. 2020. Efficient neural machine translation for low-resource languages via exploiting related languages.

In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 162–168, Online, July. Association for Computational Linguistics.

Ha, Thanh-Le, Jan Niehues, and Alex Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C, December 8-9. International Workshop on Spoken Language Translation.

Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

K, Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-Lingual ability of multilingual BERT: An empirical study. In *Proceedings of the 8th International Conference on Learning Representations*, Online.

K, Karthikeyan, Aalok Sathe, Somak Aditya, and Monojit Choudhury. 2021. Analyzing the effects of reasoning types on cross-lingual transfer performance. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 86–95, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Kreutzer, Julia and Artem Sokolov. 2018. Learning to segment inputs for NMT favors character-level processing. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 166–172, Brussels, October 29-30. International Conference on Spoken Language Translation.

Kudo, Taku and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November. Association for Computational Linguistics.

Kunchukuttan, Anoop, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Kunchukuttan, Anoop. 2020. The IndicNLP Library. `https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf`.

Liu, Hairong, Mingbo Ma, Liang Huang, Hao Xiong, and Zhongjun He. 2019. Robust neural machine translation with joint textual and phonetic embedding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3044–3049, Florence, Italy, July. Association for Computational Linguistics.

Magueresse, Alexandre, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Muller, Benjamin, 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 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online, June. Association for Computational Linguistics.

Nakov, Preslav and Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 301–305, Jeju Island, Korea, July. Association for Computational Linguistics.

Nguyen, Toan Q. and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Oladipo, Akintunde, Odunayo Ogundepo, Kelechi Ogueji, and Jimmy Lin. 2022. An exploration of vocabulary size and transfer effects in multilingual language models for african languages. In *3rd Workshop on African Natural Language Processing*.

Ott, Myle, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Pires, Telmo, 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, July. Association for Computational Linguistics.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Reimers, Nils and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11.

Rosales Núñez, José Carlos, Djamé Seddah, and Guillaume Wisniewski. 2019. Phonetic normalization for machine translation of user generated content. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 407–416, Hong Kong, China, November. Association for Computational Linguistics.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany, August. Association for Computational Linguistics.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Sun, Simeng, Angela Fan, James Cross, Vishrav Chaudhary, Chau Tran, Philipp Koehn, and Francisco Guzmán. 2022. Alternative input signals ease transfer in multilingual machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5291–5305, Dublin, Ireland, May. Association for Computational Linguistics.

Wu, Shijie and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online, July. Association for Computational Linguistics.

Xue, Linting, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

A Generalisation of Models

Table 4 reports results for the zero-shot and few-shot set-up for Marathi-English and Bengali-English. We use samples of sizes 500, 1k, 10k, and further report a fine-tuning topline, which uses all available data for each of the language pairs. Similar to the earlier setups, we evaluate vocabulary sizes in { character, 4k,8k,16k,32k } and temperature values in { 1.2,1.5, 1.8 }.

	#fine-tuning examples														
	0			**0.5k**			**1k**			**10k**			**full set**		
	1.2	1.5	1.8	1.2	1.5	1.8	1.2	1.5	1.8	1.2	1.5	1.8	1.2	1.5	1.8
Original															
mr→en															
char	4.23	4.54	**4.86**	4.50	4.61	**4.68**	5.58	5.63	**5.72**	7.37	7.58	***7.61***	8.07	**8.14**	8.02
4K	4.89	5.12	***5.56***	4.12	4.72	***4.86***	5.23	5.65	***5.83***	6.98	7.12	**7.43**	8.02	8.34	***8.45***
8K	4.45	4.83	**5.32**	4.03	4.53	**4.62**	5.11	5.44	**5.73**	6.87	6.99	**7.01**	7.63	7.72	**7.88**
16K	4.36	4.49	**5.43**	4.41	4.46	**4.72**	5.08	5.39	**5.61**	6.76	6.87	**6.94**	7.58	7.63	**7.74**
32K	4.11	4.35	**5.40**	4.52	4.68	**4.71**	5.33	5.46	**5.79**	6.54	6.57	**6.77**	7.41	7.64	**7.92**
bn→en															
char	2.53	2.61	**2.64**	4.03	4.08	***4.10***	4.98	5.01	***5.03***	5.72	5.78	**5.85**	***7.45***	7.36	7.40
4K	3.31	3.41	***3.93***	3.02	3.43	**3.81**	4.40	4.51	**4.88**	6.12	6.49	***6.53***	6.98	7.08	**7.31**
8K	3.50	3.55	**3.67**	3.01	3.48	**3.65**	4.35	4.48	**4.67**	5.56	5.93	**6.01**	6.48	6.75	**7.13**
16K	3.65	3.70	**3.74**	3.00	3.49	**3.52**	4.28	4.37	**4.59**	5.71	5.83	**6.00**	6.16	6.37	**6.82**
32K	3.21	3.25	**3.26**	3.07	3.35	**3.52**	4.36	4.48	**4.67**	5.74	5.86	**5.91**	6.80	**7.14**	7.04
Transliterated															
mr→en															
char	5.02	5.12	**5.33**	4.66	4.76	**4.78**	5.73	**5.81**	5.80	7.32	7.46	**7.48**	8.20	8.34	***8.38***
4K	5.02	5.33	***5.81***	4.51	4.73	***4.91***	5.61	5.72	***5.92***	7.11	7.34	***7.58***	8.10	**8.12**	7.99
8K	5.24	5.41	**5.71**	4.34	4.65	**4.85**	5.50	5.61	**5.80**	6.95	7.02	**7.12**	7.71	7.86	**7.95**
16K	5.15	5.41	**5.71**	4.22	4.60	**4.83**	5.48	5.65	**5.78**	6.92	6.98	**7.22**	7.95	7.98	**8.04**
32K	5.17	5.76	**5.78**	4.40	4.70	**4.70**	5.45	5.58	**5.81**	6.87	**7.01**	6.97	7.58	7.67	**8.01**
bn→en															
char	3.39	3.42	**3.58**	4.03	4.12	***4.37***	4.98	5.04	***5.15***	5.76	5.85	**5.91**	7.27	7.38	**7.39**
4K	3.68	3.79	***4.23***	3.15	3.66	**3.98**	4.36	4.68	**4.92**	6.33	6.56	***6.67***	7.02	7.13	***7.45***
8K	3.75	3.86	**3.95**	3.10	3.54	**3.76**	4.48	4.55	**4.72**	5.95	6.02	**6.12**	6.64	6.83	**7.02**
16K	3.77	3.83	**3.99**	3.02	3.51	**3.65**	4.31	4.48	**4.65**	5.86	5.98	**6.03**	6.54	6.77	**6.99**
32K	3.76	3.91	**3.93**	3.14	3.42	**3.68**	4.43	4.56	**4.82**	5.81	5.90	**5.93**	6.83	**7.02**	6.95

Table 4: BLEU scores for few-shot performance on transliterated English-Bengali and English-Marathi pairs using character tokenisation and shared BPE with vocabulary size v in $\{4000, 8000, 16000, 32000\}$. Bold shows best score for each vocabulary size and bold italic represents best score overall.