

# Unsupervised Machine Translation On Dravidian Languages

Sai Koneru, Danni Liu and Jan Niehues

Department of Data Science and Knowledge Engineering, Maastricht University

s.koneru@student.maastrichtuniversity.nl

{danni.liu, jan.niehues}@maastrichtuniversity.nl

## Abstract

Unsupervised neural machine translation (UNMT) is beneficial especially for low resource languages such as those from the Dravidian family. However, UNMT systems tend to fail in realistic scenarios involving actual low resource languages. Recent works propose to utilize auxiliary parallel data and have achieved state-of-the-art results. In this work, we focus on unsupervised translation between English and Kannada, a low resource Dravidian language. We additionally utilize a limited amount of auxiliary data between English and other related Dravidian languages. We show that unifying the writing systems is essential in unsupervised translation between the Dravidian languages. We explore several model architectures that use the auxiliary data in order to maximize knowledge sharing and enable UNMT for distant language pairs. Our experiments demonstrate that it is crucial to include auxiliary languages that are similar to our focal language, Kannada. Furthermore, we propose a metric to measure language similarity and show that it serves as a good indicator for selecting the auxiliary languages.

## 1 Introduction

Around 200 million people in the world speak a Dravidian language which has almost 80 varieties (Kolipakam et al., 2018). Given the significant number of speakers, improving translation quality on these languages is an important task to facilitate intercultural communication. While neural machine translation (NMT) systems (Bahdanau et al., 2015; Vaswani et al., 2017) led to great improvements on high resource languages, a major challenge remains when there is little parallel data between the source and target languages. Compared to resource-rich languages like German and French, the amount of high-quality parallel data for Dravidian languages is minimal. As a result, the

current translation quality for Dravidian languages is largely lagging behind.

Unsupervised neural machine translation (UNMT; Artetxe et al., 2018, Lample et al., 2018) learns to translate between languages without relying on parallel data. Instead, only monolingual corpora are required. This relieves the need to acquire large parallel datasets, which is especially difficult for low resource languages. Despite this attractive property, existing literature (Marchisio et al., 2020; Kim et al., 2020) suggests that UNMT performs well under the following conditions: 1) Source and target languages are similar; 2) Monolingual data for both the languages belong to the same domain. In most realistic scenarios that call for unsupervised translation, these conditions are rarely fulfilled.

Considering these challenges, one promising research direction to address these conditions is to combine multilingual NMT models with unsupervised translation (Liu et al., 2020; Li et al., 2020b; Garcia et al., 2020b; Guzmán et al., 2019). Incorporating similar high resource languages could alleviate the data scarcity of low resource languages by exploiting the language similarity between them. Fraser (2020) describes this scenario as “unsupervised machine translation with multilingual transfer”. For example, Li et al. (2020a) investigate translating between German to Upper Sorbian with no parallel data between them but using parallel data between the high resource languages German and English. Among Dravidian languages, the comparatively better resourced ones are Tamil, Telugu, Malayalam and Kannada. In turn, among these four languages, Kannada is relatively low resource compared to the others (Reddy and Sharoff, 2011). Therefore, we consider a realistic scenario of improving translation quality on Kannada using auxiliary data in the other three languages. We refer to the additional language as *reference language*

following the terminology in Li et al. (2020b).

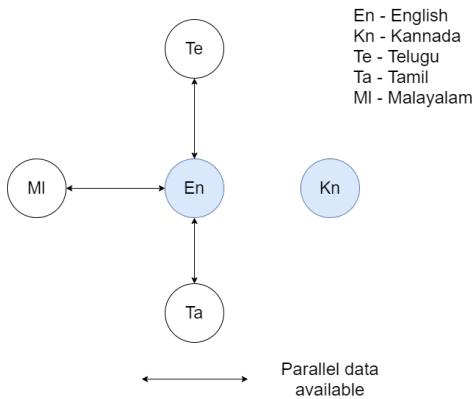


Figure 1: Illustration of the data conditions of the languages in our work. Connected nodes indicate availability of parallel data. We focus on translating between English and Kannada (blue nodes) without any parallel data.

In this work, with a focus on Dravidian languages, we analyze the influence of related languages as auxiliary data in various setups of multilingual UNMT (MUNMT). As shown in Figure 1, we focus on a scenario of translating between Kannada and English with the following conditions: 1) no parallel data for Kannada-English; 2) very limited in-domain parallel data ( $\sim 30k$  sentences) between English and three other Dravidian languages; 3) monolingual data for all five languages in a different domain from the test data.

Our main contributions in this paper are:

- We show that transliteration is beneficial in UNMT between related languages that do not share the same writing system.
- We investigated multiple architectures to include reference languages and enable MUNMT for diverse language pair such as English and Kannada.
- We analyze the influence of language similarity and show the importance of using reference languages similar to Kannada.

## 2 Related Work

Initial work on UNMT (Lample et al., 2018) demonstrates the possibility of building unsupervised bilingual systems between similar and high resource languages. However, as shown recently (Guzmán et al., 2019; Marchisio et al., 2020; Kim et al., 2020), purely unsupervised systems perform poorly in situations with low resource and distant language pairs. However, often it is in this very

case that unsupervised translation could be helpful. Guzmán et al. (2019) propose a hybrid approach that translates between Nepali and English without parallel data by utilizing parallel data between Hindi and English. Specifically, the model is trained jointly with supervised loss between English and Hindi and unsupervised loss between English and Nepali. In our work, we employ additional loss terms on top of this in order to more tightly connect the languages without parallel data. Moreover, we analyze ways to incorporate auxiliary parallel data and investigate the correlation to language similarity.

The current state-of-the-art UNMT systems build upon multilingual pretraining (Liu et al., 2020; Garcia et al., 2020b) and include a large number of languages from diverse languages families. The high degree of multilinguality is expected to lead to better model generalization. In this work, we zoom in on Dravidian languages and study how we can best utilize the related languages within this particular family. We also study data scenarios with minimal auxiliary parallel data and use a training procedure that relies more on the monolingual data.

The languages in our setup belong to the same family, but have unique scripts. Because of this, there is little vocabulary overlap. Chakravarthi et al. (2018, 2019a,b,c) compare the impact of transliteration in supervised multilingual NMT and shows that it was helpful. Different from Chakravarthi et al. (2019c), the focus of our work is unsupervised translation. We study the role of lexical overlap in a unsupervised training pipeline, including pretraining on monolingual data and supervised training on auxiliary language pairs. We also highlight the information loss caused by transliteration from code-mixing in data from certain domains.

## 3 Background

We define  $L_P$  as the set of language pairs with parallel data, and  $L_M$  as the set of languages with monolingual data. The sample space of sentences for language  $X \in L_M$  is  $\phi_X$  following the terminology in (Li et al., 2020c). Similarly, for language pair  $X - Y \in L_P$ , the sample space of parallel sentences is  $\phi_{X-Y}$ .

A standard bilingual supervised model translating from language  $X$  to  $Y$  is trained on parallel dataset  $\phi_{X-Y}$ . The model parameters  $\theta$  are optimized to maximize the likelihood of target sen-

tences in  $Y$  given the source sentences in  $X$ :

$$\mathcal{L}_{MT}(\theta, X, Y) = \mathbf{E}_{(x,y) \sim \phi_{X-Y}}[-\log p_{\theta}(y|x)]. \quad (1)$$

For unsupervised translation, in absence of parallel sentences, Lample et al. (2018) propose to combine a denoising auto-encoding task with an on-the-fly back-translation task, with the former learning language models from monolingual data, and the latter bridging the two languages.

In the auto-encoding task, the model aims to reconstruct sentences from corrupted inputs. Specifically, a noise model  $C(x)$  corrupts an input sentence  $x$  by dropping and shuffling the words according to a given probability. The loss function for auto-encoding using a monolingual dataset for language  $X$  from the collection  $L_M$  is defined as

$$\mathcal{L}_{AE}(\theta, X) = \mathbf{E}_{(x) \sim \phi_X}[-\log p_{\theta}(x|C(x))]. \quad (2)$$

In the on-the-fly back-translation task, the model only needs monolingual data and its loss function is defined as

$$\mathcal{L}_{BT}(\theta, X, Y) = \mathbf{E}_{(x) \sim \phi_X}[-\log p_{\theta}(x|\hat{y})], \quad (3)$$

where  $\hat{y} = \operatorname{argmax}_y \log p_{\theta}(y|x)$  is the intermediate translation generated in language  $Y$  during training.

The full loss function for unsupervised translation combines the auto-encoding loss in Equation (2) on  $X$  and  $Y$  respectively and the back-translation loss in Equation (3) between  $X$  and  $Y$  in both directions:

$$\begin{aligned} \mathcal{L}_{UMT}(\theta, X, Y) = & \mathcal{L}_{AE}(\theta, X) + \mathcal{L}_{AE}(\theta, Y) + \\ & \mathcal{L}_{BT}(\theta, X, Y) + \mathcal{L}_{BT}(\theta, Y, X). \end{aligned} \quad (4)$$

Cross translation (Garcia et al., 2020a; Li et al., 2020b) is a recently proposed technique that uses a reference language to improve the performance of unsupervised multilingual models. The procedure is similar to on-the-fly back-translation but additionally utilizes parallel dataset with *reference languages*. When training an unsupervised translation model from  $X$  to  $Y$ , we use an additional parallel dataset  $X - R$  from the collection  $L_P$ , where  $R$  is the reference language. Beside the UNMT loss, we additionally optimize the following cross translation loss function:

$$\mathcal{L}_{CT}(\theta, X, R, Y) = \mathbf{E}_{(x-r) \sim \phi_{X,R}}[-\log p_{\theta}(r|\hat{y})], \quad (5)$$

where  $\hat{y} = \operatorname{argmax}_y \log p_{\theta}(y|x)$  is the intermediate translation.

For both on-the-fly back-translation and cross translation, the quality of intermediate translations is essential for the final UNMT model (Li et al., 2020b). High-quality intermediate translations further needs similar encoder outputs for parallel sentences in the source and target language. While the denoising auto-encoding objective is expected to encourage this, achieving this is expected to be easier between similar languages. Therefore, in this work we focus on utilizing similar languages for UNMT.

## 4 Language Similarity in Unsupervised Machine Translation

As outlined in Section 2, distant language pairs pose a great challenge to UNMT systems. In order to apply UNMT between highly dissimilar languages, in our case English and Kannada, we explore several directions to leverage language similarity using other reference languages in the Dravidian family.

First, we investigate the role of transliteration in UNMT, as the languages we work with do not share a common writing system. Secondly, we evaluate multiple model architectures to enable unsupervised translation of dissimilar language pairs. We experiment with using auxiliary data from various related reference languages in each architecture to show the influence of language similarity. Last but not least, we propose a method to measure language similarity using  $n$ -gram overlap. Based on this metric, we show the benefit of including reference languages with high similarity score.

### 4.1 Transliteration

Unsupervised translation is more difficult when the writings systems differ (Marchisio et al., 2020). Although the languages in our work belong to the Dravidian language family, they do not share a common script. In fact, the ones we focus on (Kannada, Tamil, Telugu, Malayalam) all have unique scripts. To enable a unified writing system and potentially common cross-lingual representations, we transliterate the Dravidian text into the Latin script. As a result, when translating into Dravidian languages, the model outputs are romanized. To restore the original script, we transliterate once again to convert the outputs back into the respective Dravidian writing system.

While this transliteration procedure is conceptually straightforward, in practice an additional challenge arises due to code-mixed inputs. We illustrate this with the example in Figure 2. As shown in the original Telugu sentence in the first row, the website name is written in the Latin script. The second row shows a perfect translation in the Latin script. However when restoring the original Telugu script in the third row, we incorrectly convert the website name into the wrong writing system, as the transliteration model has no knowledge of the code-mixing in the original sentence. A similar issue occurs with numbers and punctuation marks, which could get converted to the Dravidian versions (e.g. “.” mapped to “|” in Figure 2).

Figure 2: Example sentence in Telugu showing information loss from transliteration and partial recovery through post-processing. Orange lines mark the words transliterated into the incorrect script as a result of code-mixing. Blue lines mark the correct forms.

*Original* : ఇలాంటి ఎన్నో విషయాలను నేను  
thebetterindia.com website లో చదివాను.

*Romanized* : iIAMTi ఎన్నో viSayalanu nenu  
thebetterindia.com website lo cadivAnu.

*Deromanized* : ఇలాంటి ఎన్నో విషయాలను నేను  
ధైర్యేరినిఅ/చమ్ wవిభిత్తే లో చదివాను!

*Post-processed* : ఇలాంటి ఎన్నో విషయాలను నేను  
ధైర్యేరినిఅ/చమ్ wవిభిత్తే లో చదివాను.

To partially cope with the problem described above, we devise a simple post-processing step. When we restore the original script of a Dravidian language, we keep the numbers and punctuation marks in the Latin script. Nevertheless, we still lose information because of code-mixing in the data, such as the website name in Figure 2.

#### 4.2 Exploiting Language Similarity

A main focus of this work is to investigate the impact of language similarity in MUNMT. After allowing the model to exploit language similarity by mapping the text to a common script, we analyze the impact of various model architectures on enabling knowledge sharing. We further study the influence of language similarity in each architecture by comparing different reference languages.

In our setup, the goal is to translate between English and Kannada but we do not have parallel data between them. Although, parallel data is available between English and reference languages such as Telugu, Tamil, Malayalam. We explain the

different architectures we investigate and training procedure of these models below.

We explore both bilingual and multilingual models to include these reference languages. We build these models by including language embedding to each word embedding as proposed in [Conneau and Lample \(2019\)](#). This is similar to the inclusion of position embedding in the Transformer architecture [Vaswani et al. \(2017\)](#). As a first step, we pretrain a Masked Language Model (MLM; [Devlin et al., 2019](#)) on the monolingual corpora for all languages in the model following the procedure proposed by [Conneau and Lample \(2019\)](#). The goal of this step is to get better initialization of word embeddings.

After pretraining, we consider several training setups to utilize auxiliary data, as illustrated in Figure 3 and described next:

**Cascaded** The cascaded system consists of two separate NMT models as shown in Figure 3.1, a supervised one translating English  $\leftrightarrow$  reference language, and an unsupervised one translating reference language  $\leftrightarrow$  Kannada. The assumption here is that it is easier to learn unsupervised between languages that are in the same family. The reference language will therefore be a Dravidian language. For example, when translating from English to Kannada, we first generate a translation from English to the reference language using the supervised model and pass this intermediate translation to the unsupervised model to get the desired translation.

**Unsupervised, dissimilar languages** A multilingual model is trained to jointly perform supervised translation for reference language  $\leftrightarrow$  English and unsupervised translation for English  $\leftrightarrow$  Kannada as shown in Figure 3.2. For the former, we use the standard cross-entropy loss from Equation (1). For the latter, we use the unsupervised loss in Equation (4) between English and Kannada.

**Unsupervised, similar languages** The setup is closely related to the previous one. The main difference that the unsupervised translation is between Kannada and another Dravidian language, instead of English as shown in Figure 3.3. Similar to the motivation for the cascaded system, we hope to achieve strong unsupervised model by exploiting the similarity between Kannada and the reference language. During test time, we generate translations between English and Kannada directly unlike the cascaded system, which uses intermediate translations in the reference language.

**Unsupervised, dissimilar languages + cross translation** We add the cross translation loss from Equation (5) in order to generate higher-quality intermediate translations as shown in Figure 3.4. We extend the previous setup for unsupervised between dissimilar languages but also with cross translation loss using Kannada as the language where we generate intermediate translations.

**Unsupervised, similar languages + cross translation** This setup is based on the previous setup with unsupervised between similar languages as shown in Figure 3.5. We add the cross translation loss from Equation (5), with Kannada being the language where we generate the intermediate translations.

**Comparing the approaches** Among the approaches outlined above, the cascaded system consists of two separately-trained NMT models, whereas all the other approaches use a single model. Comparing to the cascaded system provides insights to the benefit of multilinguality within a single model. Furthermore, we will evaluate the impact of using different languages as reference language  $R$ . This comparison allows us to evaluate whether a certain reference language (e.g. one similar to Kannada) is consistently more beneficial.

### 4.3 Measuring Language Similarity

A main motivation of our approach is to utilize similar languages to improve performance on another language. While a multitude of factors contribute to language similarity, such as morphological, syntactical, and phonetical features, a quantitative indicator could provide us guidance when choosing auxiliary multilingual data.

One way to measure language similarity is to consider the lexical overlap between two languages. As the Dravidian languages we work with are agglutinative, we choose to calculate the overlap based on character  $n$ -grams rather than at word level. Since our models operate on sub-word level rather than word level, taking character  $n$ -grams is more appropriate. In addition, we take into account the frequency of the shared character  $n$ -grams, as overlap between rare character  $n$ -grams is not as important compared to frequent ones. Below we explain the procedure for calculating the language similarity metric.

Let us consider two languages  $A$  and  $B$ . We create their vocabularies by taking unique character  $n$ -grams on the monolingual data denoted as  $V_A$  and

$V_B$  respectively. The union between  $V_A$  and  $V_B$  is the list of all character  $n$ -grams denoted as  $V$ . Let  $\text{freq}(\cdot)$  be a function that calculates the frequency of a character  $n$ -grams in the given language. We calculate the overlap of vocabularies between two languages  $A$  and  $B$  using the following formula

$$\text{overlap} = \frac{\sum_{w \in V} \min(\text{freq}(A, w), \text{freq}(B, w))}{\sum_{w \in V} \max(\text{freq}(A, w), \text{freq}(B, w))}, \quad (6)$$

where  $V = V_A \cup V_B$ . Consider a scenario of two languages that do not share any common character  $n$ -grams. The frequency of every character  $n$ -grams in the union of two vocabularies is 0 in the language that it is not present. Therefore, calculating overlap between two languages without any common character  $n$ -grams gives a score of 0 as we take the minimum in the numerator. Having minimum frequency in the numerator and maximum frequency in the denominator provides higher score to character  $n$ -grams that occur frequently in both the languages compared to a common character  $n$ -grams that occurs frequently in only one language.

## 5 Experiments and Results

First, we analyze the influence of having a unified writing system based on the approach described in Section 4.1. Next, we evaluate multiple architectures with different reference languages described in Section 4.2. We then discuss our results in the context of language similarity.

### 5.1 Datasets

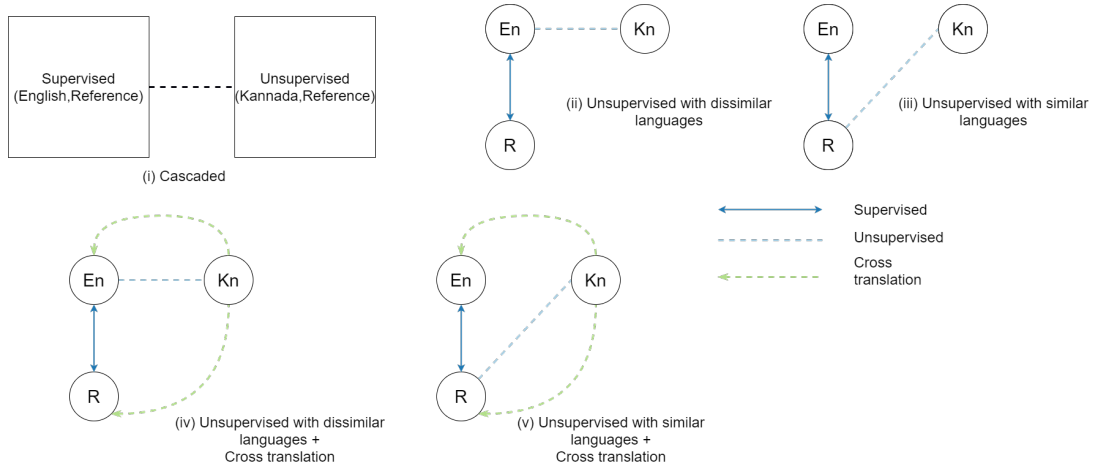
The monolingual datasets we use for Dravidian languages are from AI4Bharat-IndicNLP corpus (Kunchukuttan et al., 2020) while for English we use Wikipedia dumps. The parallel datasets we use from English to Dravidian languages are from the PMIndia dataset (Haddow and Kirefu, 2020). Almost 80% of English sentences are common in any two parallel datasets. We align these common English sentences and extract parallel data between Dravidian languages. The number of sentences in each dataset used for training is shown in Table 1.

### 5.2 Pre-Processing and Training Parameters

We use the *indic-transliteration*<sup>1</sup> toolkit for transliteration between Latin script and Indic script. We then use Byte Pair Encoding (Sennrich et al., 2016)

<sup>1</sup><https://pypi.org/project/indic-transliteration/>

Figure 3: Pictorial description of multiple architectures to include auxiliary parallel data in reference language  $R$ . Here  $R$  is chosen to be a language similar to Kannada.



to learn sub-words with 10k merge operations. We tokenize with *Indic NLP Library*<sup>2</sup> for Dravidian languages whereas for English we use Moses (Koehn et al., 2007). All the experiments have been performed using *XLM*<sup>3</sup> toolkit. For our multilingual models, we use the Transformer (Vaswani et al., 2017) architecture with 6 layers and 8 heads whereas for our bilingual supervised model baselines we have 5 layers and 2 heads following the same configuration as in Guzmán et al. (2019). We use a smaller model for the supervised model to avoid overfitting the parallel data amount is minimal. The other parameters are set to default values as they are in *XLM*. The scores are reported in BLEU (Papineni et al., 2002) and characTER (Wang et al., 2016). When translating into Dravidian languages with transliteration, we report the scores after restoring the original scripts. We report detokenized BLEU scores using sacreBLEU<sup>4</sup> (Post, 2018).

### 5.3 Effect of Unifying Writing System

In order to assess the impact of transliteration, we perform experiments in the supervised and unsupervised setup between the following language pairs:

- English and a Dravidian language (dissimilar).
- Two Dravidian languages (similar).

We consider English and Kannada for dissimilar

<sup>2</sup>[https://github.com/anoopkunchukuttan/indic\\_nlp\\_library](https://github.com/anoopkunchukuttan/indic_nlp_library)

<sup>3</sup><https://github.com/facebookresearch/XLM>

<sup>4</sup>BLEU+case.mixed+numrefs.1+smooth.exp+tok.spm+version.1.4.12. We use the tokenizer “spm” instead of the default “13a” for better tokenization of Indian languages.

Languages	Monolingual data	Parallel data to English
Kannada (Kn)	14M	31k*
Tamil (Ta)	21M	34k
Telugu (Te)	15M	34k
Malayalam (Ml)	11M	30k
English (En)	46M	-

Table 1: The number of sentences in the monolingual and parallel datasets for different languages. \*Note that the English-Kannada parallel data is only used to train the supervised model (as performance upper bound). The training of all other models do not involve this set.

language pair and Kannada and Telugu for similar pair. In Table 2, we report the performance before and after transliteration.

First, for the supervised models, romanizing the Dravidian languages has no conclusive effects on translation quality. While there is a drop in BLEU points after transliteration, the characTER scores remain unchanged. Upon inspection of the translation results, we also notice some of the outputs suffer from the information loss described in Section 4.1.

Second, we observe that the purely unsupervised systems perform poorly, especially between English and Kannada, the dissimilar language pair. The scores are nearly zero with or without romanization. This confirms the previous findings (Marchisio et al., 2020; Kim et al., 2020) that unsupervised translation is difficult between distant languages. Between Telugu and Kannada, the unsupervised performance largely improves after romanization. The BLEU scores increase by 3.6 and 2.8 points respectively when translating into and from

		Supervised				Unsupervised			
		Original Script		Romanized		Original Script		Romanized	
Source	Target	BLEU	characTER	BLEU	characTER	BLEU	characTER	BLEU	characTER
En	Kn	<b>16.4</b>	76	16	76	0.1	85	0.2	84
Kn	En	<b>12.3</b>	77	11.5	77	0.3	93	0.6	93
Te	Kn	<b>20.1</b>	69	18.9	69	3.5	79	<b>7.1</b>	<b>74</b>
Kn	Te	<b>11.9</b>	74	11.1	74	2.3	86	<b>5.1</b>	<b>82</b>

Table 2: Impact of transliteration with similar and distant language pairs. Scores with noticeable differences to the other writing system are shown in bold. The translation quality is measured in BLEU ( $\uparrow$ ) and characTER ( $\downarrow$ ).

Kannada. The characTER scores also decrease by 5 and 4 points absolute. This demonstrates that mapping the text into a common script helps in exploiting the similarity between related languages for UNMT. This finding also contrasts the observations under the supervised model, where transliteration shows no conclusive impact. The difference leads us to conclude that the role of transliteration is scenario-dependent, in this case on the amount of available supervision signals from parallel data.

Furthermore, we use the similarity metric proposed in Section 4.3 to validate this finding. We calculate this by taking character tri-grams on the monolingual data. Indeed, in Table 3, we see that the overlap between Kannada and Telugu increase from 1% to 34% after transliteration, exhibiting the highest similarity scores between Kannada and {Te, Ta, MI}. This finding is also in line with the close relation between Kannada and Telugu from the linguistic literature (Datta, 1988; Bright, 1996). This evidence shows that transliteration bridges across different writing systems and enables similar representations for related languages. After transliteration, the unsupervised model can more easily leverage the similarity between these related languages and therefore achieve better performance.

	En	Kn	Te	Ta	MI
En	100	0 $\rightarrow$ 6	0 $\rightarrow$ 6	0 $\rightarrow$ 3	0 $\rightarrow$ 6
Kn		100	1 $\rightarrow$ <b>34</b>	1 $\rightarrow$ 10	1 $\rightarrow$ 26
Te			100	1 $\rightarrow$ 9	1 $\rightarrow$ 27
Ta				100	1 $\rightarrow$ 11
MI					100

Table 3: Lexical overlap (%; calculated using Equation 6) before and after transliteration (before  $\rightarrow$  after).

#### 5.4 Impact of Model Architectures

We evaluate the setups described in Section 4.2 with different reference languages and report these scores in Table 4. Given the previously-found positive impact of unifying the writing systems on

unsupervised learning, we transliterate Dravidian languages in all experiment setups here.

First, the unsupervised models on similar languages consistently outperform those with dissimilar languages. This is in line with previous results in Table 2 and further validates our finding on the positive role of language similarity in UNMT.

Second, adding the cross translation loss improves the translation quality into English. The model using Telugu as reference language achieves 5.1 BLEU points when translating into English, the highest among the setups explored so far. However, there is not a similar gain when translating into Kannada. One reason is that cross translation generates more training signals when translating into English but does not explicitly encourage better translation into Kannada.

#### 5.5 Impact of Reference Languages

Table 4 provides a clear view of which reference language is more beneficial to a purely unsupervised system. In line with previous observations, when using Telugu as the reference language, the model performance exceeds when using other languages. The only exception is unsupervised dissimilar + cross translation scenario where it falls slightly behind Malayalam. The positive impact of adding Telugu is within expectation given the previous finding on its close relation to Kannada. After Telugu, the next best choices for the reference languages are Malayalam, followed up Tamil. This particular order agrees with the vocabulary overlap with Kannada shown in Table 3, with the highest value on Telugu (34%) followed by Malayalam (26%) and lastly Tamil (10%). Based on our similarity metric, we are able to show the impact of using languages that are more similar.

Given the positive impact of incorporating related reference languages, we extend our investigation to using multiple reference languages. We build upon the best architecture (unsupervised sim-

Reference Language	Cascaded		Unsupervised dissimilar		Unsupervised similar		Unsupervised dissimilar + Cross translation		Unsupervised similar + Cross translation	
	BLEU	characTER(%)	BLEU	characTER(%)	BLEU	characTER(%)	BLEU	characTER(%)	BLEU	characTER(%)
Te	3.7/ <b>1.6</b>	86/ <b>82</b>	1.3/2.1	89/87	3.8/1	86/85	2.2/0.4	87/90	<b>5.1</b> /1.4	<b>85</b> /93
MI	2.3/ <b>1.1</b>	86/ <b>87</b>	1.1/0.6	89/92	2.0/0.7	89/93	2.7/0.9	85/91	<b>3.3</b> /0.2	<b>84</b> /96
Ta	1.2/0.2	87/ <b>89</b>	0.3/ <b>0.4</b>	89/93	1.4/0.4	88/94	0.7/0.3	<b>86</b> /94	<b>1.8</b> /0.1	87/97

Table 4: Unsupervised translation performance between Kannada and English using different reference languages in multiple setups described in Section 4.2. Scores are shown as x/y where x indicates in Kn  $\rightarrow$  En direction and y indicates in En  $\rightarrow$  Kn respectively. The translation quality is measured in BLEU ( $\uparrow$ ) and characTER ( $\downarrow$ ). The best scores (BLEU & characTER) for each reference language in both directions are shown in bold.

	Kn $\rightarrow$ En		En $\rightarrow$ Kn	
	BLEU	characTER(%)	BLEU	characTER(%)
Supervised	11.5	77	16	76
Unsupervised	0.6	93	0.2	84
MUNMT, Te	5.1	85	1.4	93
MUNMT, Te+MI	<b>5.4</b>	<b>84</b>	1.6	95

Table 5: Scores for baselines and proposed model that includes reference languages. MUNMT corresponds to unsupervised similar + cross translation architecture. “MUNMT, Te” uses only Telugu as a reference language whereas “MUNMT, Te+MI” uses both Telugu and Malayalam as reference languages.

ilar + cross translation) and train it with using auxiliary data of two reference languages. We choose Telugu and Malayalam as they are more related to Kannada and yielded stronger performance than Tamil. The scores for this along with our baselines are in Table 5. By additionally incorporating Malayalam, we see an improvement from 5.1 to 5.4 BLEU points when translating to English. The overlap in the English sentences for the parallel data with the reference languages is more than 80%. Even though we are adding additional data, very few new in-domain English sentences are seen by the model. This might explain the minimal increase in score after adding multiple reference languages. While this performance is still 6.1 points short from supervised model, it is significantly better than the purely unsupervised baseline when translating into English.

## 5.6 Examples of Kn $\rightarrow$ En Translation

We highlight that our overall task for unsupervised translation remains very challenging, as the auxiliary parallel data amount with the reference languages is also minimal ( $\sim 30k$  sentences). Moreover, there is a domain discrepancy between the monolingual data and the test data from government announcements. In Figure 4, we show exam-

Kannada : ಮಹಾತ್ಮಾ ಗಾಂಧೀ ಅವರು ಕೂಡ ಸಮಾಜದ ಕಟ್ಟಕಡೆಯ ವ್ಯಕ್ತಿಗೆ ಏನಿದೆ ಎಂದು ಹೇಳುತ್ತಿದ್ದರು.  
English : And even Mahatma Gandhi used to say that what is there for the last man?  
Hypothesis : " Mahatma Gandhi used to say , " No Science has dropped from the skies in a perfect form ."

Kannada : ನೇಪಾಳದಲ್ಲಿನ ಬೆಳವಣಿಗೆಗಳ ಬಗ್ಗೆ ಶ್ರೀ ಬಿಮಲೇಂದ್ರ ನಿಧಿ ಅವರು ಪ್ರಧಾನಮಂತ್ರಿಯವರಿಗೆ ವಿವರಿಸಿದರು.  
English : Mr. Bimalendra Nidhi briefed the Prime Minister about developments in Nepal.  
Hypothesis : Mr. Bimalendra Nidhi was also briefed on the Prime Minister about the progress achieved in Nepal .

Figure 4: Example outputs when translating Kn  $\rightarrow$  En.

ple outputs when translating Kn  $\rightarrow$  En using the previous best “MUNMT, Te+MI” model. The first example shows the problem of hallucination (Koehn and Knowles, 2017), hinting towards negative impacts of domain shifts. Nonetheless, the model sometimes achieves translation with adequate quality such as the last example in Figure 4.

## 6 Conclusion

We focus on the challenging yet realistic scenario of translating between English and Kannada without any parallel data. After confirming the difficulty of purely unsupervised translation between distant language pairs, we show that increasing the lexical overlap by transliteration is necessary for achieving UNMT for our focal language. Moreover, we utilize monolingual and limited amount of auxiliary parallel data in languages related to Kannada. The best performance is achieved by UNMT between similar languages along with cross translation. We show the importance of incorporating languages with high relatedness to the focus language by evaluating different reference languages. Finally, we propose a metric for measuring lexical overlap and show that it serves as a good indicator for selecting the best reference languages.



**Acknowledgement** This work is supported by the Facebook Sponsored Research Agreement “Language Similarity in Machine Translation”.

## References

- Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. [Unsupervised neural machine translation](#). In *International Conference on Learning Representations*.
- Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.
- William Bright. 1996. Kannada and telugu writing. *The world’s writing systems*, pages 413–419.
- Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2018. [Improving wordnets for under-resourced languages using machine translation](#). In *Proceedings of the 9th Global Wordnet Conference*, pages 77–86, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.
- Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2019a. [Comparison of Different Orthographies for Machine Translation of Under-Resourced Dravidian Languages](#). In *2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 70 of *OpenAccess Series in Informatics (OA-SIcs)*, pages 6:1–6:14, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.
- Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2019b. [WordNet gloss translation for under-resourced languages using multilingual neural machine translation](#). In *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*, pages 1–7, Dublin, Ireland. European Association for Machine Translation.
- Bharathi Raja Chakravarthi, Ruba Priyadarshini, Bernardo Stearns, Arun Jayapal, Sridevy S, Mihael Arcan, Manel Zarrouk, and John P McCrae. 2019c. [Multilingual multimodal machine translation for Dravidian languages utilizing phonetic transcription](#). In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 56–63, Dublin, Ireland. European Association for Machine Translation.
- Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.
- Amaresh Datta. 1988. *Encyclopaedia of Indian Literature: devraj to jyoti*, volume 2. Sahitya Akademi.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 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. Association for Computational Linguistics.
- Alexander Fraser. 2020. [Findings of the WMT 2020 shared tasks in unsupervised MT and very low resource supervised MT](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 765–771, Online. Association for Computational Linguistics.
- Xavier Garcia, Pierre Foret, Thibault Sellam, and Ankur Parikh. 2020a. [A multilingual view of unsupervised machine translation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3160–3170, Online. Association for Computational Linguistics.
- Xavier Garcia, Aditya Siddhant, Orhan Firat, and Ankur P Parikh. 2020b. [Harnessing multilinguality in unsupervised machine translation for rare languages](#). *arXiv preprint arXiv:2009.11201*.
- Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc’Aurelio Ranzato. 2019. [The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English](#). 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)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.
- Barry Haddow and Faheem Kirefu. 2020. [PMIndia—a collection of parallel corpora of languages of india](#). *arXiv preprint arXiv:2001.09907*.
- Yunsu Kim, Miguel Graça, and Hermann Ney. 2020. [When and why is unsupervised neural machine translation useless?](#) In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 35–44, Lisboa, Portugal. European Association for Machine Translation.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. [Moses: Open source toolkit for statistical machine translation](#). In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.
- Philipp Koehn and Rebecca Knowles. 2017. [Six challenges for neural machine translation](#). In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

- Vishnupriya Kolipakam, Fiona M Jordan, Michael Dunn, Simon J Greenhill, Remco Bouckaert, Russell D Gray, and Annemarie Verkerk. 2018. A bayesian phylogenetic study of the dravidian language family. *Royal Society open science*, 5(3):171504.
- Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. AI4Bharat-IndicNLP corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.
- Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018. **Unsupervised machine translation using monolingual corpora only**. In *International Conference on Learning Representations*.
- Zuchao Li, Hai Zhao, Rui Wang, Kehai Chen, Masao Utiyama, and Eiichiro Sumita. 2020a. **SJTU-NICT’s supervised and unsupervised neural machine translation systems for the WMT20 news translation task**. In *Proceedings of the Fifth Conference on Machine Translation*, pages 218–229, Online. Association for Computational Linguistics.
- Zuchao Li, Hai Zhao, Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2020b. **Reference language based unsupervised neural machine translation**. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4151–4162, Online. Association for Computational Linguistics.
- Zuchao Li, Hai Zhao, Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2020c. Reference language based unsupervised neural machine translation. *arXiv preprint arXiv:2004.02127*.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. **Multilingual denoising pre-training for neural machine translation**. *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. **When does unsupervised machine translation work?** In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583, Online. Association for Computational Linguistics.
- Kishore Papineni, 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. Association for Computational Linguistics.
- Matt Post. 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. Association for Computational Linguistics.
- Siva Reddy and Serge Sharoff. 2011. **Cross language POS taggers (and other tools) for Indian languages: An experiment with Kannada using Telugu resources**. In *Proceedings of the Fifth International Workshop On Cross Lingual Information Access*, pages 11–19, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. **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. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008.
- Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. **CharacTer: Translation edit rate on character level**. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510, Berlin, Germany. Association for Computational Linguistics.