Bilingual Lexicon Induction across Orthographically-distinct Under-Resourced Dravidian Languages

Bharathi Raja Chakravarthi[♣], Navaneethan Rajasekaran[♥], Mihael Arcan[♣], Kevin McGuinness[♥], Noel E. O'Connor[♥], John P. McCrae[♣]
[♣,♥]Insight SFI Research Centre for Data Analytics,
[♣]Data Science Institute, National University of Ireland Galway, Galway, Ireland
[♥]Dublin City University, Dublin, Ireland
bharathi.raja@insight-centre.org

Abstract

Bilingual lexicons are a vital tool for under-resourced languages and recent state-of-the-art approaches to this leverage pretrained monolingual word embeddings using supervised or semi-supervised approaches. However, these approaches require cross-lingual information such as seed dictionaries to train the model and find a linear transformation between the word embedding spaces. Especially in the case of low-resourced languages, seed dictionaries are not readily available, and as such, these methods produce extremely weak results on these languages. In this work, we focus on the Dravidian languages, namely Tamil, Telugu, Kannada, and Malayalam, which are even more challenging as they are written in unique scripts. To take advantage of orthographic information and cognates in these languages, we bring the related languages into a single script. Previous approaches have used linguistically sub-optimal measures such as the Levenshtein edit distance to detect cognates, whereby we demonstrate that the longest common sub-sequence is linguistically more sound and improves the performance of bilingual lexicon induction. We show that our approach can increase the accuracy of bilingual lexicon induction methods on these languages many times, making bilingual lexicon induction approaches feasible for such under-resourced languages.

1 Introduction

Bilingual lexicon induction (BLI) is the process of creating lexicons for two or more languages from monolingual corpora (Irvine and Callison-Burch, 2017). It is a time-consuming process to do it manually so automatically inducing bilingual lexicons based on edit-distance (Haghighi et al., 2008), comparable corpora (Turcato, 1998), bilingual corpora (Rosner and Sultana, 2014) or pretrained embeddings from monolingual corpora (Vulić and Moens, 2015) is more suitable. However, sentence-aligned parallel data is not available for all languages. Methods based on unsupervised or semi-supervised learning can utilise readily available monolingual data to induce bilingual lexicons. Artetxe et al. (2018) showed that an iterative self-learning method could bootstrap this approach without the need of a seed dictionary by utilising numbers as seed dictionary through adversarial training. However, Patra et al. (2019) showed that with even a small seed dictionary, the results could be improved considerably. This task is further complicated by the fact that many languages use distinct scripts, and as such learning the similarities between cognates is a non-trivial task. As such, BLI is a challenging task for under-resourced languages due to lack of seed dictionaries and large monolingual corpora. For this work, we proposed to use the IndoWordNet as a seed dictionary for the closely related Dravidian languages, namely Tamil, Telugu, Kannada, and Malayalam, which use different scripts.

BLI between closely-related languages has shown to perform better than unrelated languages (Irvine and Callison-Burch, 2017), since closely related languages often share similar linguistics properties and cognates (Nasution et al., 2016). Cognates are words that have a similar meaning and similar orthography based on etymological relationships (Kondrak et al., 2003). Computational models of monolingual embeddings also exhibit isomorphism across closely related languages (Mikolov et al., 2013b; Ormazabal

Proceedings of the 7th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, pages 57–69
Barcelona, Spain (Online), December 13, 2020

et al., 2019) based on the assumption that word embeddings in different languages have approximately the same structure. This isomorphic property was exploited by Artetxe et al. (2018) and Lample et al. (2018) to map monolingual word embeddings in different languages to a shared space through a linear transformation. For closely-related languages, it follows that cognates can be used as a form of alignment as words that have a similar form are quite likely to be cognates and therefore could be used as a weak seed dictionary. Given previous work on the use of seed dictionaries (Patra et al., 2019), the usage of such alignments is likely to improve performance of BLI. Previous works used the Levenshtein distance (Riley and Gildea, 2018); however, this is not linguistically well-motivated as it allows for multiple changes that are not consistent with the kinds of changes seen etymologically.

The goal of this work is to exploit the orthographic information between languages that use a different script. For that purpose, we bring the languages into a single script, which allows us to take advantage of the cognate properties of closely related languages. This paper has two principal contributions: first, we study the use of transliteration, and we demonstrate that it is an effective and necessary step which yields more isomorphic embeddings and obtains more robust BLI. Second, we show that the use of the longest common subsequence (LCS) is a superior method of assessing the cognate similarity.

2 Dravidian Languages

Dravidian languages are the common terminology (Caldwell, 1856) used to represent the South Indian languages, which consist of around 26 languages divided into four branches: 11 in the Southern group, 7 in the South-Central group, 5 in the Central group and 3 in the Northern group (Krishnamurti, 2003; Chakravarthi et al., 2018). Out of the 26 Dravidian languages, many of them are non-literary languages except the four languages chosen for this paper. Indigenous minority populations primarily use the non-literary languages. The modern society widely uses the four literary languages in literature, public communications, government institutions, academic settings and many other places in the day-to-day life of an ordinary person (Chakravarthi et al., 2020b; Chakravarthi et al., 2020a). For many natural language processing tasks such as machine translation (MT) systems, it is essential to have a corpus of written documents, as well as well-defined lexicons and grammar for the selected languages (Chakravarthi et al., 2020c). Hence in this work, we will focus on the four chosen Dravidian languages Tamil, Malayalam, Kannada and Telugu which are spoken by approximately 210 million people (Steever, 2015) across the world either as their first or second language.

ISO 639-3	tam	mal	kan	tel	eng
Script	Tamil	Malayalam	Kannada	Telugu	Latin
	உப்பு(Uppu)	ഉപ്പ് (Uppu)	ಉಪ್ಪು (Uppu)	ఉప్పు (Upp)	Salt
	நாணயம் (Nanayam)	നാണയം (Nanayam)	ನಾಣ್ಯ (Nanya)	నాణెం (Nanem)	Coin
	வாகனம்(Vakanam)	വാഹനം(Vahanam)	ವಾಹನ (Vahan)	వాహనం (Vahanam)	Vehicle
	நடை(Nadai)	നട (Nada)	ನಡಿ (Nadi)	నడిచి (Nadici)	Walk

Figure 1: Example of cognate words for the Dravidian language

Even though these languages share a common root, they cannot be termed as regional dialects of a language of the same origin (Caldwell, 1856). Tamil and Malayalam are more closely related such that a regional speaker of one language can understand another without translation (Burrow and Emeneau, 1961). Figure 1 illustrates the words for 'salt', 'coin', 'vehicle', 'walk' are similar in both Tamil and Malayalam. The Tamil, Malayalam and Telugu languages have their own written script symbols whereas Telugu and Kannada have significant similarities in their script symbols. Though Telugu and Kannada have these similarities, they are not readily intelligible for speakers of the language. The study of languages suggests that these languages formed a single language around late 4000 BCE and then started evolving on their own (Steever, 2015). Since the languages evolved sharing geographical, etymological and political borders, the cognates may have evolved similar meanings or borrowed words from each other. Chakravarthi et al. (2019a) have compared the Latin script and the International Phonetic Alphabet (IPA) for multilingual

translation systems and shown that bringing the Dravidian languages into Latin script outperforms a multilingual neural machine translation system trained on native script and IPA. Inspired by this, we transform the Dravidian language monolingual corpora into a single script (Latin script).

3 Our Approach

3.1 Bilingual Lexicon Induction

State-of-the-art approaches to BLI use monolingual (Haghighi et al., 2008) or comparable corpora (Fung, 1995; Tamura et al., 2012) to identify pairs of translated words with or without a seed dictionary (Vulić and Korhonen, 2016). The induced translation can improve MT systems (Golan et al., 1988) to expand the coverage of translation models by translating Out-Of-Vocabulary (OOV) words. Nevertheless, prior work in BLI treated it as stand-alone task (Irvine and Callison-Burch, 2017).

Using monolingual word embeddings for BLI has attracted significant attention in recent years. State-of-the-art BLI results are based on bilingual word embedding models (Irvine and Callison-Burch, 2017). Given the source and target language word embeddings trained independently on monolingual data, unsupervised models (Vulić and Moens, 2015; Artetxe et al., 2016; Zhang et al., 2017; Artetxe et al., 2017; Artetxe et al., 2018; Riley and Gildea, 2018; Artetxe et al., 2019) learn a linear mapping W between the source and target space such that:

$$W^* = \arg\min_W \sum_i \sum_j D_{ij} \left\| X_{i*} W - Z_{j*} \right\|^2, \tag{1}$$

where X and Z are two aligned matrices of embedding size *d* containing the embeddings of the words in the parallel vocabulary. The vocabulary of each language are V_s and V_t, and $D \in \{0,1\}^{|V_s| \times |V_t|}$ is a binary matrix representing a dictionary such that $D_{ij} = 1$ if the i-th word in the source language is aligned with the j-th word in the target language. Equation (1) is equivelent to:

$$W^* = \arg\max_W \mathrm{Tr}\left(XWZ^T D^T\right), \tag{2}$$

where $\mathrm{Tr}(\cdot)$ is the trace operator (the sum of all diagonal entries). The optimal solution to this equation is $W^* = UV^T$, where $X^T DZ = U\Sigma V^T$ is the singular value decomposition of $X^T DZ$. To get this required seed dictionary D, Artetxe et al. (2018) introduced an iterative, self-learning framework that uses numerals as a seed dictionary for the first time to determine W and uses it to calculate D. From the next iteration on, it appends D to the seed dictionary to learn as shown in the Algorithm 1. However, it suffers from the non-availability of seed dictionary.

Algorithm 1: Self-learning framework

Input: X (source embeddings)
Input: Z (target embeddings)
Input: D (seed dictionary)
1: **repeat**
2: $W \leftarrow$ LEARN-MAPPING (X, Z, D)
3: $D \leftarrow$ LEARN-DICTIONARY (X, Z, W)
4: **until** convergence criterion
5: EVALUATE-DICTIONARY(D)

The goal of the function LEARN-MAPPING (X, Z, D) is to find the optimal mapping matrix W^* so that the sum of squared Euclidean distances between the mapped source embeddings $X_{i*}W$ and target embeddings Z_{j*} for the dictionary entries D_{ij} is minimised as per Equation (1) and (2).

LEARN-DICTIONARY (X, Z, W) uses the dot product between the mapped source language embeddings and the target language embeddings as the similarity measure, which is roughly equivalent to cosine similarity. Then D_{ij} is set to 1, if $j = \mathrm{argmax}_k \left(X_{i*}W\right) \cdot Z_{k*}$ and D_{ij} is set to 0 otherwise.

We evaluate our process by using D to create a translation for words from test set and then compare it with the true values. EVALUATE-DICTIONARY(D) calculates translation accuracy on the test set:

$$\text{Translation Accuracy} = \frac{\text{Correct Translations}}{\text{All Samples}}, \tag{3}$$

where the *'Correct Translation'* is the number of correct target words in the source translation from the test set. *'All Samples'* is the total number of samples in the test set.

Riley and Gildea (2018) proposed two methods to utilise the orthographic information to improve the BLI. The first method is an orthographic extension of word embeddings, where each word embedding in the monolingual embedding is appended with a vector of length equal to the size of the union of the two language alphabets.

Mathematically, let A be an ordered set of alphabet containing all characters appearing in both language's alphabets:

$$A = A_{\text{source}} \cup A_{\text{target}}. \tag{4}$$

Let O_{source} and O_{target} be the orthographic extension matrices for each language, containing counts of the characters appearing in each word w_i, scaled by a constant factor c_e:

$$O_{ij} = c_e \cdot \text{count}\left(A_j, w_i\right), \quad O \in \left\{O_{\text{source}}, O_{\text{target}}\right\}. \tag{5}$$

Then embedding matrices were concatenated with orthographic matrices as below:

$$X' = \left[X; O_{\text{source}}\right], \quad Z' = \left[Z; O_{\text{target}}\right]. \tag{6}$$

Finally, in the normalized embedding matrices X'' and Z'', each row has magnitude 1:

$$X''_{i*} = \frac{X'_{i*}}{\|X'_{i*}\|}, \quad Z''_{i*} = \frac{Z'_{i*}}{\|Z'_{i*}\|}. \tag{7}$$

X'' and Z'' are new matrices that are used in the place of X and Z to include orthographic information.

The second approach of Riley and Gildea (2018) modifies the similarity score to include orthographic information for each word pair during the dictionary induction phase of the self-learning phase. Instead of using the dot product of the words' embeddings to quantify similarity, the approach modifies the similarity score by adding a measure of orthographic similarity, which is a function of Levenshtein distance (Levenshtein, 1966) divided by the length of the longer word. The normalised Levenshtein distance, denoted NL, is:

$$\text{NL}\left(w_1, w_2\right) = \frac{\text{L}\left(w_1, w_2\right)}{\max\left(\left|w_1\right|, \left|w_2\right|\right)}. \tag{8}$$

The Orthographic similarity of two words w_1 and w_2 is $\log\left(2.0 - \text{NL}\left(w_1, w_2\right)\right)$. The edit distance for a subset of possible word pairs is just considered as by how far most of word sets are orthographically unique, resulting in a normalised edit distance close to 1 and an orthographic similarity close to 0.

3.2 Longest Common Subsequence

The Levenshtein distance is a standard measure of the distance between two sequences by a minimum number of single-character edits required to map one string from another based on deletions, additions and substitution. This approach makes a binary decision about whether a pair of characters match. LCS (Paterson and Dančík, 1994; Melamed, 1999) is a similarity measure of two or more strings to find the longest subsequence common to all sequence in two or more strings. LCS was previously used to extract the morphological variations and generate lexicons (Hulden et al., 2014; Sorokin, 2016). LCS was also used to identify cognate candidates during the construction of N-best translation lexicons from parallel texts (Melamed, 1995; Kondrak et al., 2003), and for the automatic evaluation of translation quality (Lin and Och, 2004). Recent work on the creation of a large-scale multilingual lexical database based on cognates was introduced by Batsuren et al. (2019), called CogNet, which uses LCS ratio to find cognates.

Karakanta et al. (2018) also used the LCS ratio to extract cognate pairs from Wikipedia titles between Russian and Belarusian. Inspired by this, we use LCS for our work.

A sequence $Z = [z_l, z_2, \ldots, z_n]$ is a subsequence of another sequence $X = [x_1, x_2, \ldots, x_m]$, if there exists a sequence $[i_l, i_2, \ldots, i_k]$ of indices of X such that for all $j = 1, 2, \ldots, k$, where $x_{ij} = z_j$. Given two sequences X and Y, the LCS of X and Y is a common subsequence with maximum length. More formally:

$$\mathrm{LCS}(X_i, Y_j) = \begin{cases} \emptyset & \text{if } i = 0 \text{ or } j = 0 \\ \mathrm{LCS}(X_{i-1}, Y_{j-1})' x_i & \text{if } i, j > 0 \text{ and } x_i = y_j \\ \max\{\mathrm{LCS}(X_i, Y_{j-1}), \mathrm{LCS}(X_{i-1}, Y_j)\} & \text{if } i, j > 0 \text{ and } x_i \neq y_j \end{cases} \tag{9}$$

In previous works, Artetxe et al. (2018) uses the dot product of two words embeddings to quantify similarity. Riley and Gildea (2018) uses normalised string edit distance based on Levenshtein distance during the dictionary induction phase of the self-learning framework. In our method, we used LCS during the dictionary induction phase of self-learning framework. LCS is used to measure the orthographic similarity of the languages (Melamed, 1995; Nakov and Ng, 2012). Dravidian languages have rich morphological features, which will be beneficial in comprehending their cognate information by LCS. To include LCS, we modify the Riley and Gildea (2018) similarity score for each word pair during the dictionary induction phase. The normalised edit distance of Equation (8) is modified as to become:

$$\mathrm{NLCS}(w_1, w_2) = 1 - \frac{\mathrm{LCS}(w_1, w_2)}{\max(|w_1|, |w_2|)}. \tag{10}$$

Now the orthographic similarity of two words w_1 and w_2 is $\log(2.0 - \mathrm{NLCS}(w_1, w_2))$.

3.3 Phonetic Transcription

Phonetic transcription is used for describing a speech by means of symbols. The most common type of phonetic transcription uses a phonetic alphabet, such as the International Phonetic Alphabet (IPA); however, transcribing into the Latin script or non-native script is prevalent due to the ubiquity of the US/UK keyboard. The IPA is an evolving standard initially developed by the International Phonetic Association in 1888 with the goal of transcribing the sounds of all human languages. Transliteration is used to help language learners to read words written in foreign scripts, by writing the sound of the word using the equivalent letters. Romanisation remains a popular technique for transliteration of various languages (Hermjakob et al., 2018). The use of the Latin script for text entry of South Asian languages is common, even though there is no standard orthography for these languages in the script (Wolf-Sonkin et al., 2019). The 107 symbols used for writing the IPA are taken primarily from the Latin and Greek scripts some are novel creations. Diacritics are used for subtle distinctions in sounds and to show nasalisation of vowels, length, stress, and tones. Using IPA symbols, one can represent the pronunciation of words. Nevertheless, the study by Chakravarthi et al. (2019a) and Chakravarthi et al. (2019c) shows that transliteration into the Latin Script is best suited to take advantage of cognate information from closely related languages.

For example, LCS of the input sequence "AABCDH" and "AABHEDE" is "AAB" of length 3. However, LCS might be zero even though the Dravidian languages share a common root. This is due to the difference in the orthography of these languages. They must be converted to a single script to take advantage of the closeness of these languages. The phonetic transcription of Dravidian languages by Chakravarthi et al. (2019b) showed improvement in the translation of WordNet entries and compared the results with IndoWordNet. Chakravarthi (2020) showed that the usage of Latin script outperforms the IPA for Multilingual NMT for Dravidian languages. This was proven with a cosine similarity of the corpus showing that transcribing the text into the Latin script retain more similarity. Inspired by this, we used the Indic-trans library by Bhat et al. (2015) to transliterate. We show the example of a comparison of NLCS and NL between languages for examples of cognate words in Table 1.

Previous methods based on edit-distance and orthographic similarity are proposed for using linguist features for word alignments by supervised and unsupervised methods (Dyer et al., 2011; Berg-Kirkpatrick et al., 2010; Hauer et al., 2017). Hauer et al. (2017) created a seed dictionary based on the cognates of

Language Pairs	Word Pair	English Translation	NLCS	NL
kan-mal	hajaradant-hajarulla	arriving	0.4545	0.6363
kan-mal	rahasyadaan-rahasyadan	secret	0.2307	0.3076
kan-tam	navratn-navmani	nine gems	0.3750	0.5000
kan-tam	tandeilladant-tacoppanillat	having no living father	0.4285	0.7857
kan-tel	poojaniyavadantah-poojyaniyulu	worthy of adoration	0.5555	0.6666
kan-tel	atyagatyavadant-atyavasaramin	primary	0.5000	0.5625
mal-tam	navaratnam-navmani	nine gems	0.5454	0.6363
mal-tam	tatanillat-tacoppanillat	having no living father	0.3571	0.4285
mal-tel	navaratnam-navratnalu	nine gems	0.2727	0.3636
mal-tel	tatanillat-tandriless	having no living father	0.5454	0.7272
tam-tel	tacoppanillat-tandriless	having no living father	0.6428	0.7857
tam-tel	sammadam-samardhinchada	approval	0.6000	0.6666

Table 1: Example training set comparison of NLCS and NL

the related languages using orthographic information. They have shown that approaches that include orthographic information outperform the previous approaches for closely related languages. Unlike our work, previous works did not study languages that have different scripts, and used Levenshtein distance without considering morphological properties.

3.4 Data

Lexicons such as WordNet (Miller, 1995; Miller, 1998) for English or EuroWordNet for European languages are lexical resources which were used to improve MT quality. EuroWordNet is a cross-lingual synonym resource that linked WordNet synsets across European languages (Vossen, 1997). Similarly, IndoWordNet (Bhattacharyya, 2010) links WordNet synsets across major Indian languages from the Indo-Aryan, Dravidian and Sino-Tibetan families. An online multilingual dictionary for Indian languages was developed by Redkar et al. (2015) from IndoWordNet. However, this dictionary is not publicly accessible. To train and evaluate the quality of BLI, a seed dictionary and a test set of the bilingual lexicon is required. For the Dravidian languages, there is no existing seed dictionary, so we used the IndoWordNet. To create a seed dictionary, we used the IndoWordNet ID to link the WordNet entries for Tamil, Telugu, Malayalam and Kannada. We map the one-to-many word mapping from IndoWordNet to one-to-one word mapping by replicating the source word. Table 2 shows the seed dictionary statistics and Table 3 shows the statistics for the test set. The test set was randomly chosen among the mapping from IndoWordNet for languages under study. Even though IndoWordNet is not perfect, it is one of the resources readily available for the under-resourced Dravidian language experiments.

Language Pair	Number of entries
Tamil-Telugu	11,666
Tamil-Kannada	4,353
Tamil-Malayalam	18,731
Telugu-Kannada	12,769
Telugu-Malayalam	3,791
Kannada-Malayalam	4,639

Table 2: Number of entries in the initial bilingual lexicons used as a seed dictionary for the experiment

Wikipedia is a free online encyclopedia that created by volunteers from a different regions of the world. It has contents in more than 200 languages. Wikipedia dumps (Wikidump)[1] for Tamil, Telugu, Malayalam

[1]https://dumps.wikimedia.org/

Language Pair	Number of entries
Tamil-Telugu	1,982
Tamil-Kannada	1,930
Tamil-Malayalam	1,918
Telugu-Kannada	2,000
Telugu-Malayalam	2,000
Kannada-Malayalam	1,999

Table 3: Number of entries in the test set for the experiment

and Kannada were downloaded to create a monolingual embedding for each language. Wikidumps were downloaded from July 2019. Wikiextractor[2] was used to extract the documents from the Wikidump. The total number of sentences and the number of tokens from Wikidump is given in Table 4. Even if Wikipedia is available for more than 200 languages, many dumps are relatively small in size compared to other high resourced language such as English. The Wikidumps for Dravidian languages are considerably smaller corpora than that of the pre-trained embeddings for other high resourced languages studied by Artetxe et al. (2017) and Riley and Gildea (2018). We used the indic-trans library[3] to transliterate the corpus into Latin script. We trained the embedding based on the skip-gram model with 300 dimensions and default parameters.

Language	Number of sentences	Number of tokens
Tamil	1,088,753	18,761,579
Telugu	1,423,448	27,229,563
Kannada	416,764	11,109,735
Malayalam	539,755	10,501,347

Table 4: Number of sentences and number of tokens extracted from wikimedia.org for Dravidian languages

4 Experimental Settings

Words that share a similar context are semantically related. Based on their word embeddings, methods represent words in a vector space by grouping semantically similar words near each other. Word embeddings are useful for several lexical-semantic tasks such as detecting synonyms and disambiguating word sense. Several pre-trained embedding models are publicly available such as word2vec[4] (Mikolov et al., 2013a; Mikolov et al., 2013b), global word representation-based models (GloVe)[5] (Pennington et al., 2014) and FastText[6] (Bojanowski et al., 2017; Grave et al., 2018). FastText was used to create monolingual embeddings from Wikipedia articles. FastText enhances traditional word-based vectors by representing each word as a bag of character n-grams. Incorporating this subword information from FastText embeddings as well as semantic relatedness allows the capturing of orthographic and morphological similarity. We did not use the pre-trained embeddings from FastText since we also created embedding for transliteration. Given that the main focus of our work is on bringing closely related languages into a single script, we transcribed the Wikidump corpus before creating word embeddings and training seed dictionary. We conducted an experiment on BLI on language pairs, Tamil-Telugu, Tamil-Kannada, Tamil-Malayalam, Telugu-Malayalam, Telugu-Kannada, and Kannada-Malayalam.

[2]https://github.com/attardi/wikiextractor
[3]https://github.com/libindic/indic-trans
[4]https://github.com/tmikolov/word2vec
[5]https://nlp.stanford.edu/projects/glove/
[6]https://fasttext.cc/

Similar to Riley and Gildea (2018)[7], we stop training when the improvement on the average cosine similarity for the induced dictionary is below 10^{-6} between successive iterations. We compare our methods with baselines with and without a seed dictionary. An automatically-generated dictionary consisting only of numeral identity translations such as 4-4, 8-8, as in Artetxe et al. (2017), was used as a training set as the input dictionary to the baseline system without a seed dictionary.

5 Results and Discussion

We show results of all eight cases studied for BLI in Table 5. First, we added the seed dictionary created from IndoWordNet to the work by Artetxe et al. (2017) while maintaining the corpora in native script. Adding the seed dictionary showed small improvement over the baseline. Further, we did an experiment with the methodology proposed by Riley and Gildea (2018) as another comparable baseline still maintaining the corpora in the native script, which showed improvement over baselines. In our approach, we first transliterate the corpora work with numerals only and a seed dictionary. In the final experiment, we used LCS to improve baseline methodology with transliteration numerals only and a seed dictionary.

Once the dictionary D is learned by the self-learning process, we use the dictionary D to create translations for source words from the test set, and compare it against the target words within the evaluation set to calculate translation accuracy. Translation accuracy is the proportion of correct predictions among the total number of cases examined in the test set given in Equation 3. The total number of cases examined for language pairs under study is given in Table 3. Translation accuracy is our evaluation measure since most of the state-of-the systems are evaluated using accuracy. For example, our method with LCS-transliterated+seed dictionary for the Tamil-Malayalam language pair yields 220 correct word translations of a total of 2,000, giving a translation accuracy of 11.00%. The experimental results indicate a seed dictionary and transliteration improve accuracy. We further investigate our result to explicate the effects of cognates from similar languages.

	Approach	tam-tel	kan-tam	tam-mal	tel-kan	mal-tel	kan-ml
(Artetxe et al., 2017)	N-Numerals	0.00	0.00	0.00	0.00	0.00	0.00
	N-Seed-Dict	2.61	1.22	1.33	2.45	2.11	1.45
(Riley and Gildea, 2018)	N-Numerals	4.06	3.08	4.01	4.80	3.20	3.66
	N-Seed-Dict	5.67	4.66	6.35	6.24	4.64	4.65
T-with NL	T-Numerals	8.93	6.03	10.34	9.24	4.98	5.02
	T-Seed-Dict	10.11	6.20	10.56	9.16	5.01	5.13
T-with NLCS	T-Numerals	9.01	6.10	10.38	9.28	5.05	5.02
	T-Seed-Dict	10.12	6.36	11.00	9.74	6.04	5.36

Table 5: Performance comparison of bilingual lexicon induction on test data for Dravidian languages. Translation accuracy is represented in percentage. N: native script, T: transliteration, seed-dict: seed dictionary.

As it can be seen from Table 5, our approach with LCS outperforms the baseline methods within their groups for all four languages. Moreover, the proposed transliteration approach gives the best accuracy compared against all baseline methods for all six language pairs (Table 1). Our method with a transliteration and LCS outperforms all baselines on Tamil-Telugu. Interestingly, the transliteration and LCS fails on Malayalam-Kannada with the numerals seed dictionary, since the monolingual corpus for these languages is very small compared with the high-resourced languages such as English, Spanish and German studied by Artetxe et al. (2018). For example, 1.3 billion + 65 billion tokens for German, 702 million + 36 billion tokens for Italian, and 127 million + 6 billion tokens for Finnish from Wikipedia

[7]https://github.com/Luminite2/vecmap

and Common Crawl respectively are used in training word vectors, a very high number compared to tokens in Dravidian languages as shown in Table 4. The results for the IndoWordNet seed dictionaries show that our method is comparable or even better than the baseline systems. As another reference, the best-published results using orthographic information used by Riley and Gildea (2018) for high-resource languages reported accuracy of 55.53% for English-German, 46.27% for English-Italian, and 41.78% for English-Finnish dictionary. In any case, the main focus of our work is on under-resourced languages, and it is under this setting that our method stands out.

As it can be seen, our method with LCS obtains the best results in all language pairs and directions, with a highest accuracy of 11.00% for the Tamil-Malayalam language pair and lowest of 5.36% for the Kannada-Malayalam language pair. These results are very consistent across all translation directions. This suggests that, while previous methods did not focus on languages with different scripts, there is a substantial margin of improvement when orthographic information is taken into consideration. We believe that, beyond the substantial gains in this particular task, our work has important implications for future research in MT and cross-lingual word embeddings mapping between languages that use different scripts.

All approaches do better with the transliteration corpora, indicating that this may be suitable for under-resourced closely related languages in different scripts. We observed that providing IndoWordNet as a seed dictionary helps with the training process when compared to purely a unsupervised approach using only numbers as the seed dictionary. When the word vectors are not rich enough, the baseline methods fail entirely to map the embeddings without a seed dictionary. Orthographic information added to the BLI does not face this problem. As can be observed, the model performs reasonably well even with numerals only as a seed dictionary.

6 Conclusion

In this paper, we have explored bringing closely related languages into a single script and their impact on the task of BLI from monolingual word embeddings. We created a seed dictionary for Tamil, Telugu, Malayalam and Kannada from IndoWordNet. Our initial experiments with 'off-the-shelf' BLI, based on the alignment of numbers, produced very poor results, showing that these methods were not possible to apply directly to these under-resourced languages. We found that mapping these languages to a single character system helps the system to discover cognates in these closely-related under-resourced languages. Further, we showed that LCS is more linguistically sound for cognate detection by quantitative evaluation and in application to BLI. This paper shows that importance of evaluating methodologies directly on under-resourced languages as the challenges related to these languages may require modification to existing methodologies so as to make them work effectively, as demonstrated in this paper.

Acknowledgements

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2 (Insight_2), co-funded by the European Regional Development Fund as well as by the EU H2020 programme under grant agreements 825182 (Prêt-à-LLOD), and Irish Research Council grant IRCLA/2017/129 (CARDAMOM-Comparative Deep Models of Language for Minority and Historical Languages).

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada, July. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 789–798, Melbourne, Australia, July. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. Bilingual lexicon induction through unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5002–5007, Florence, Italy, July. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Gabor Bella, and Fausto Giunchiglia. 2019. CogNet: A large-scale cognate database. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3136–3145, Florence, Italy, July. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, Alexandre Bouchard-Côté, John DeNero, and Dan Klein. 2010. Painless unsupervised learning with features. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 582–590, Los Angeles, California, June. Association for Computational Linguistics.

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. IIIT-H system submission for FIRE2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, pages 48–53, New York, NY, USA. ACM.

Pushpak Bhattacharyya. 2010. IndoWordNet. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Thomas Burrow and Murray Barnson Emeneau. 1961. *A Dravidian etymological dictionary: supplement*. Oxford University Press.

Robert Caldwell. 1856. A comparative grammar of the Dravidian or south-Indian family of languages.(Madras: University of Madras. 1961).

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 (GWC 2018)*, page 78.

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 (OASIcs)*, 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, August. European Association for Machine Translation.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, 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, August. European Association for Machine Translation.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020a. A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France, May. European Language Resources association.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020b. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France, May. European Language Resources association.

Bharathi Raja Chakravarthi, Priya Rani, Mihael Arcan, and John P McCrae. 2020c. A survey of orthographic information in machine translation. *arXiv preprint arXiv:2008.01391*.

Bharathi Raja Chakravarthi. 2020. *Leveraging orthographic information to improve machine translation of under-resourced languages*. Ph.D. thesis, NUI Galway.

Chris Dyer, Jonathan Clark, Alon Lavie, and Noah A. Smith. 2011. Unsupervised word alignment with arbitrary features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, page 409–419, USA. Association for Computational Linguistics.

Pascale Fung. 1995. Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In *Third Workshop on Very Large Corpora*.

Igal Golan, Shalom Lappin, and Mori Rimon. 1988. An active bilingual lexicon for machine translation. In *Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics*.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: HLT*, pages 771–779, Columbus, Ohio, June. Association for Computational Linguistics.

Bradley Hauer, Garrett Nicolai, and Grzegorz Kondrak. 2017. Bootstrapping unsupervised bilingual lexicon induction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 619–624, Valencia, Spain, April. Association for Computational Linguistics.

Ulf Hermjakob, Jonathan May, and Kevin Knight. 2018. Out-of-the-box universal romanization tool uroman. In *ACL*.

Mans Hulden, Markus Forsberg, and Malin Ahlberg. 2014. Semi-supervised learning of morphological paradigms and lexicons. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 569–578, Gothenburg, Sweden, April. Association for Computational Linguistics.

Ann Irvine and Chris Callison-Burch. 2017. A comprehensive analysis of bilingual lexicon induction. *Computational Linguistics*, 43(2):273–310, June.

Alina Karakanta, Jon Dehdari, and Josef Genabith. 2018. Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, 32(1–2):167–189, June.

Grzegorz Kondrak, Daniel Marcu, and Kevin Knight. 2003. Cognates can improve statistical translation models. In *Companion Volume of the Proceedings of HLT-NAACL 2003 - Short Papers*, pages 46–48.

Bhadriraju Krishnamurti. 2003. *The Dravidian languages*. Cambridge University Press.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710, Feb. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, page 605–es, USA. Association for Computational Linguistics.

I. Dan Melamed. 1995. Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In *Third Workshop on Very Large Corpora*.

I. Dan Melamed. 1999. Bitext maps and alignment via pattern recognition. *Comput. Linguist.*, 25(1):107–130, March.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomas Mikolov, Google Inc, Mountain View, Quoc V. Le, Google Inc, Ilya Sutskever, and Google Inc. 2013b. Exploiting similarities among languages for machine translation.

George A. Miller. 1995. Wordnet: A lexical database for English. *Commun. ACM*, 38(11):39–41, November.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Preslav Nakov and Hwee Tou Ng. 2012. Improving statistical machine translation for a resource-poor language using related resource-rich languages. *J. Artif. Int. Res.*, 44(1):179–222, May.

Arbi Haza Nasution, Yohei Murakami, and Toru Ishida. 2016. Constraint-based bilingual lexicon induction for closely related languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3291–3298, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019. Analyzing the limitations of cross-lingual word embedding mappings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4990–4995, Florence, Italy, July. Association for Computational Linguistics.

Mike Paterson and Vlado Dančík. 1994. Longest common subsequences. In Igor Prívara, Branislav Rovan, and Peter Ruzička, editors, *Mathematical Foundations of Computer Science 1994*, pages 127–142, Berlin, Heidelberg. Springer Berlin Heidelberg.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy, July. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Hanumant Redkar, Sandhya Singh, Nilesh Joshi, Anupam Ghosh, and Pushpak Bhattacharyya. 2015. IndoWordNet dictionary: An online multilingual dictionary using IndoWordNet. In *Proceedings of the 12th International Conference on Natural Language Processing*, pages 71–78, Trivandrum, India, December. NLP Association of India.

Parker Riley and Daniel Gildea. 2018. Orthographic features for bilingual lexicon induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 390–394, Melbourne, Australia, July. Association for Computational Linguistics.

Michael Rosner and Kurt Sultana. 2014. Automatic methods for the extension of a bilingual dictionary using comparable corpora. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3790–3797, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Alexey Sorokin. 2016. Using longest common subsequence and character models to predict word forms. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 54–61, Berlin, Germany, August. Association for Computational Linguistics.

Sanford B Steever. 2015. *The Dravidian Languages*. Routledge.

Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita. 2012. Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 24–36, Jeju Island, Korea, July. Association for Computational Linguistics.

Davide Turcato. 1998. Automatically creating bilingual lexicons for machine translation from bilingual text. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 1299–1306, Montreal, Quebec, Canada, August. Association for Computational Linguistics.

Piek Vossen. 1997. EuroWordNet: a multilingual database for information retrieval. In *Proceedings of the DELOS workshop on Cross-language Information Retrieval, March 5-7, 1997 Zurich*. Vrije Universiteit. Proceedings of the DELOS workshop on Cross-language Information Retrieval, March 5-7, 1997, Zurich.

Ivan Vulić and Anna Korhonen. 2016. On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 247–257, Berlin, Germany, August. Association for Computational Linguistics.

Ivan Vulić and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 719–725, Beijing, China, July. Association for Computational Linguistics.

Lawrence Wolf-Sonkin, Vlad Schogol, Brian Roark, and Michael Riley. 2019. Latin script keyboards for south Asian languages with finite-state normalization. In *Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing*, pages 108–117, Dresden, Germany, September. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada, July. Association for Computational Linguistics.